

24^ο ΠΑΝΕΛΛΗΝΙΟ ΣΥΝΕΔΡΙΟ **ΕΣΙ**
Στατιστική &
Δεοντολογία - Διαφάνεια

27 ΑΠΡΙΛΙΟΥ · 1 ΜΑΪΟΥ 2011, ΠΑΤΡΑ



ΠΕΡΙΕΧΟΜΕΝΑ



ΕΛΛΗΝΙΚΟ ΣΤΑΤΙΣΤΙΚΟ ΙΝΣΤΙΤΟΥΤΟ
Ε.Σ.Ι
GREEK STATISTICAL INSTITUTE
G.S.I

Π Ρ Α Κ Τ Ι Κ Α

24^ο Πανελληνίου
Συνεδρίου Στατιστικής

PROCEEDINGS

of the 24th Panhellenic
Statistics Conference

Στατιστική & Δεοντολογία-Διαφάνεια

Statistics & Deontology-Transparency



**ΕΛΛΗΝΙΚΟ ΣΤΑΤΙΣΤΙΚΟ
ΙΝΣΤΙΤΟΥΤΟ
Ε.Σ.Ι
GREEK STATISTICAL INSTITUTE
G.S.I**

Π Ρ Α Κ Τ Ι Κ Α

**24^ο Πανελληνίου
Συνεδρίου Στατιστικής**

Στατιστική & Δεοντολογία-Διαφάνεια

Οργάνωση

ΕΛΛΗΝΙΚΟ ΣΤΑΤΙΣΤΙΚΟ ΙΝΣΤΙΤΟΥΤΟ

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΑΣ

**Τμήμα Μαθηματικών, Γενικό Τμήμα,
Τμήμα Μηχανολόγων και Αεροναυπηγών Μηχανικών**

Πάτρα, 27 Απριλίου-1^η Μαΐου 2011



ΕΛΛΗΝΙΚΟ ΣΤΑΤΙΣΤΙΚΟ ΙΝΣΤΙΤΟΥΤΟ

Σολωμού 5 (Πλατεία Εξαρχείων)

Τηλ. & Fax 210 33.03.909

Email: esi-stat@hol.gr

<http://www.esi-stat.gr>

ΠΕΡΙΕΧΟΜΕΝΑ

ΠΡΟΛΟΓΟΣ	αρ.σελ. 7
ΧΟΡΗΓΟΙ.....	11
ΠΡΟΓΡΑΜΜΑ ΣΥΝΕΔΡΙΟΥ.....	12
ΕΠΙΤΡΟΠΕΣ ΣΥΝΕΔΡΙΟΥ.....	27

Ομιλίες σε ολομέλεια

ΔΟΝΑΤΟΣ Γ.: Διαφάνεια, δεοντολογία και αποτελεσματικότητα στο Δημόσιο Τομέα και η Στατιστική	29
ΦΙΛΙΠΠΟΥ Α. : Τα στατιστικά στοιχεία, η Ευρωζώνη και η κρίση χρέους της Ελλάδας.....	45
ΧΡΙΣΤΟΦΙΔΗΣ Τ.: Δεοντολογία και εμπιστευτικότητα της πληροφορίας στις δειγματοληπτικές έρευνες.....	53

Εργασίες στα Ελληνικά

ΒΑΣΙΛΕΙΑΔΗΣ Γ.: Ομογενές Μαρκοβιανό σύστημα διακριτού χρόνου και ουρά διακριτού χρόνου με πεπερασμένη πηγή.....	62
ΒΑΣΙΛΕΙΑΔΗΣ Β.Γ., ΣΠΑΝΟΥ Ε.Ν., ΡΗΓΑΣ Α.Γ.: Βελτίωση της εκτίμησης των παραμέτρων ενός γενικευμένου γραμμικού μοντέλου με τη χρήση της παλινδρόμησης Ridge.....	69
ΒΟΤΣΗ Ε., ΛΗΜΝΙΟΣ Ν., ΤΣΑΚΛΙΔΗΣ Γ., ΠΑΠΑΔΗΜΗΤΡΙΟΥ Ε.: Κρυφά Μαρκοβιανά και ημι-Μαρκοβιανά μοντέλα για τον εντοπισμό του υποκειμένου κατά τη γένεση σεισμών πεδίου τάσεων.....	78
ΓΕΩΡΓΙΑΔΗΣ Σ., ΛΗΜΝΙΟΣ Ν.: Πολυδιάστατο συναρτησιακό κεντρικό οριακό θεώρημα για τον εμπειρικό εκτιμητή ενός ημι-μαρκοβιανού πίνακα.....	87
ΓΚΙΟΥΛΟΣ Α.Ν., ΤΣΑΚΛΙΔΗΣ Γ., ΣΤΑΜΑΤΕΛΛΟΣ Γ.: Εκτίμηση της από κοινού κατανομής στηθιαίας διαμέτρου και ύψους δένδρων με τη χρήση συζεύξεων (copulas).....	95
ΔΗΜΗΤΡΙΟΥ Ι.: Ένα σύστημα με επαναλαμβανόμενους πελάτες, προτεραιότητες, εξαρτημένη ροή αφίξεων και βλάβες υπαλλήλου.....	104
ΔΡΙΤΣΑΚΗΣ Ν.: Ισοτιμία αγοραστικής δύναμης και έλεγχοι μοναδιαίας ρίζας σε πάνελ δεδομένα.....	113
ΘΕΟΔΩΡΟΥ Ι., ΑΛΕΒΙΖΟΣ Φ., ΚΕΧΡΙΝΙΩΤΗΣ Α.: Ανάλυση αντιστοιχιών με ασαφή δεδομένα: η πρακτική εφαρμογή της μεθόδου (αντιπροσωπευτικό παράδειγμα με αποσαφοποιημένη παραγοντική γεωμετρική παράσταση).....	125
ΚΑΡΑΒΑΣΙΛΗΣ Γ., ΡΗΓΑΣ Α.: αναγνώριση στοχαστικών συστημάτων στάσιμων σημειακών ανελιξεων.....	134

ΚΑΡΑΒΑΣΙΛΗΣ Γ., ΡΗΓΑΣ Α., ΚΥΔΡΟΣ Δ.: Βασικά χαρακτηριστικά και ιδιαιτερότητες της στατιστικής γλώσσας προγραμματισμού R κατά τον υπολογισμό των συναρτήσεων πυκνότητας γινομένου στάσιμων σημειακών ανελίξεων.....	143
ΚΑΤΡΗΣ Χ., ΔΑΣΚΑΛΑΚΗ Σ.: Προβλεπτικά μοντέλα χρονοσειρών με εξάρτηση μακράς διάρκειας (LRD): εφαρμογή στο γενικό δείκτη του Χρηματιστηρίου Αξιών Αθηνών.....	152
ΚΕΤΖΑΚΗ Ε., ΦΑΡΜΑΚΗΣ Ν., ΚΕΤΖΑΚΗ Δ.: Σύγκριση των δεικτών ποικιλότητας χρησιμοποιώντας τη μέθοδο Jackknife.....	160
ΚΙΤΙΚΙΔΟΥ Κ., ΜΑΤΗΣ Κ.: Εφαρμογή της ανάλυσης επιβίωσης για την αξιολόγηση της αύξησης των δέντρων ελάτης στο πανεπιστημιακό δάσος Πετρουλίου.....	169
ΚΟΥΡΟΥΚΛΗΣ Σ.: Η δειγματική διασπορά και το τίμημα της αμεροληψίας της	178
ΚΟΥΤΣΟΧΕΡΑΣ Γ., ΠΑΠΑΪΩΑΝΝΟΥ Τ.: Μέτρα συνάφειας και μέτρα συμμετρίας-ασυμμετρίας για πίνακες συνάφειας.....	185
ΜΑΚΡΗ Φ.Σ., ΨΥΛΛΑΚΗΣ Ζ.Μ.: Συμβολοσειρές φραγμένου μήκους σε δυαδικές ακολουθίες.....	195
ΜΑΚΡΗΣ Γ., ΔΑΛΑΜΗΤΡΑ Ε., ΑΝΤΩΝΙΟΥ Ι.: Κατασκευή μοντέλου για την εκτίμηση της επιλογής κατεύθυνσης σπουδών των μαθητών της β΄ λυκείου	203
ΜΑΝΑΤΑΚΗΣ Μ., ΚΑΤΣΑΚΙΩΡΗ Π.: Στατιστική ανάλυση παραγόντων πρόκλησης θανατηφόρων εργατικών ατυχημάτων	211
ΜΗΛΙΕΝΟΣ Φ.Σ., ΚΟΥΤΡΑΣ Μ.Β.: Ασυμπτωματικά αποτελέσματα για χρόνους αναμονής σχηματισμών.....	220
ΜΠΑΝΤΗΣ Λ., ΤΣΙΜΗΚΑΣ Γ., ΓΕΩΡΓΙΟΥ Σ.: Εκτίμηση της συνάρτησης επιβίωσης με χρήση μονότονων κυβικών Spline	229
ΜΠΟΜΠΟΤΑΣ Π., ΚΟΥΡΟΥΚΛΗΣ Σ.: Νέες κλάσεις εκτιμητών για τον πίνακα διασπορών- συνδιασπορών πολυμεταβλητής κανονικής κατανομής	239
ΝΙΚΗΦΟΡΙΔΟΥ Ζ., ΠΑΓΓΕ Τ. ΧΑΤΖΗΠΑΝΤΕΛΗΣ Θ.: Στρατηγικές ανάληψης ρίσκου από παιδιά και ενήλικες	247
ΠΑΠΑΝΑ Α., ΚΟΥΓΙΟΥΜΤΖΗΣ Δ.: Εντοπισμός άμεσων συζεύξεων σε πολυμεταβλητά συστήματα.....	254
ΠΑΠΑΠΕΤΡΟΥ Μ., ΚΟΥΓΙΟΥΜΤΖΗΣ Δ.: Εκτίμηση τάξης της μαρκοβιανής αλυσίδας μέσω της δεσμευμένης αμοιβαίας πληροφορίας....	262
ΠΑΠΠΑΣ Β.Α., ΑΔΑΜΙΔΗΣ Κ., ΛΟΥΚΑΣ Σ.: Μία οικογένεια κατανομών στην ανάλυση επιβίωσης.....	271
ΠΕΡΙΚΛΕΟΥΣ Κ., ΚΟΥΝΙΑΣ Σ.: Βέλτιστοι σχεδιασμοί γραμμής με τρεις αγωγές και εξαρτημένες παρατηρήσεις.....	280

ΠΕΤΡΟΠΟΥΛΟΣ Κ.: Κλάσεις βελτιωμένων διαστημάτων εμπιστοσύνης για την παράμετρο κλίμακας από την πολυδιάστατη κατανομή Lomax.....	288
ΠΟΥΠΑΚΗΣ Σ.: Σύγκριση υποκειμενικών απαντήσεων σε έρευνες κοινής γνώμης και τρόποι αντιμετώπισης.....	295
ΡΑΚΙΤΖΗΣ Α.Χ., ΑΝΤΖΟΥΛΑΚΟΣ Δ.Λ.: Τεστ εκκίνησης αποδοχής εξαρτημάτων με δύο επίπεδα αποτυχίας.....	303
ΡΑΚΙΤΖΗΣ Α.Χ., ΜΠΟΥΤΣΙΚΑΣ Μ.Β., ΑΝΤΖΟΥΛΑΚΟΣ Δ.Λ.: Περί της κατανομής του αθροίσματος διακοπής: εφαρμογές στον ποιοτικό έλεγχο	312
ΡΕΚΟΥΤΗ Α., ΑΛΕΒΙΖΟΣ Φ.: Εφαρμογή μεθόδων πολυμεταβλητής στατιστικής ανάλυσης για την ανίχνευση & την περιγραφή των προτιμήσεων του ελληνικού πληθυσμού ως προς την κατανάλωση αλκοολούχων ποτών.....	321
ΣΙΓΓΙΡΙΔΟΥ Ε., ΚΟΥΓΙΟΥΜΤΖΗΣ Δ., ΜΠΟΡΑ-ΣΕΝΤΑ Ε.: Διερεύνηση της στατιστικής σημαντικότητας της συσχέτισης μεταξύ χρονοσειρών για τη δημιουργία δικτύων.....	329
ΤΡΙΑΝΤΑΦΥΛΛΟΥ Ι.Σ., ΚΟΥΤΡΑΣ Μ.Β.: Μελέτη ιδιοτήτων μιας γενικής κλάσης τυχαίων μεταβλητών που σχετίζονται με κατανομές διατεταγμένων παρατηρήσεων.....	338
ΦΑΡΜΑΚΗΣ Ν., ΜΑΚΡΗΣ Γ.: Δειγματοληψία στο διαδίκτυο: δυνατότητες ανάκτησης συγκεκριμένης πληροφορίας.....	346
ΧΑΛΙΚΙΑΣ Μ., ΚΟΥΝΙΑΣ Σ.: Σχεδιασμοί επαναλαμβανόμενων μετρήσεων, ικανές και αναγκαίες συνθήκες βελτιστοποίησης.....	355
Εργασίες στα Αγγλικά	
BASSIAKOS Y.C.: Brain drain and brain gain within eu27: Current picture and possible reasons.....	366
HOMBAS V.C.: Do the Goals per Game in 2012 World Cup of Soccer follow the Poison Distribution.....	373
POULOPOULOU S., KARLIS D., DAFNI U.: Adaptive phase II design for evaluating survival probabilities.....	378
PRODROMIDIS P.: Treating the regression misspecification problem through space. Analyzing employment and unemployment evidence from Greece.....	387
TSOLAKI E., HARALAMPOUS H., PAPADOPOULOS H.: Forecasting total electron content over Cyprus.....	397
VIANA M.: Dihedral analysis of symmetry rankings.....	406

ΠΡΟΛΟΓΟΣ

Το 24^ο Πανελλήνιο Συνέδριο Στατιστικής διοργανώθηκε από το Ελληνικό Στατιστικό Ινστιτούτο (Ε.Σ.Ι.) στη Πάτρα την περίοδο 27 Απριλίου -1^η Μαΐου 2011 σε συνεργασία με το Τμήμα Μαθηματικών, το Γενικό Τμήμα και το Τμήμα Μηχανολόγων & Αεροναυπηγών Μηχανικών του Πανεπιστημίου Πατρών. Το θέμα του Συνεδρίου ήταν *Στατιστική και Δεοντολογία-Διαφάνεια*.

Έλαβαν μέρος 213 Σύεδροι, από τους οποίους 145 ήταν φοιτητές και 28 συνοδεύοντα μέλη από την Ελλάδα και την Κύπρο. Η συμμετοχή των προπτυχιακών και των μεταπτυχιακών φοιτητών υπήρξε ικανοποιητική, συγκρίσιμη με το παρελθόν.

Στο Συνέδριο προσκεκλημένοι ομιλητές ήταν η κ. Δώρα Κυριακίδου, Γενική Διευθύντρια της Στατιστικής Υπηρεσίας Κύπρου, ο κ. Τάσος Χριστοφίδης Καθηγητής του Πανεπιστημίου της Κύπρου, ο κ. Ανδρέας Φιλίππου Καθηγητής του Πανεπιστημίου Πατρών, ο κ. Γεώργιος Δονάτος Καθηγητής του Πανεπιστημίου Αθηνών και ο κ. Τάκης Μερκούρης Επίκουρος Καθηγητής του Οικονομικού Πανεπιστημίου Αθηνών.

Το Επιστημονικό Πρόγραμμα ήταν ιδιαίτερα πλούσιο και περιελάμβανε εβδομήντα τρεις (73) ανακοινώσεις και δεκαοχτώ (18) posters σε θέματα Πιθανοτήτων, Στοχαστικών Μοντέλων, Μαθηματικής και Εφαρμοσμένης Στατιστικής, Χρονοσειρών, Στατιστικής και Οικονομίας, Αξιοπιστίας και Ελέγχου Ποιότητας, Πειραματικών Σχεδιασμών, Βιοστατιστικής, Δημογραφίας και Διατροφής, Μεθόδων Ανάλυσης Δεδομένων κτλ.

Η εναρκτήρια τελετή του Συνεδρίου πραγματοποιήθηκε στο αμφιθέατρο του Συνεδριακού και Πολιτιστικού Κέντρου του Πανεπιστημίου Πατρών. Την έναρξη κήρυξε ο Πρύτανης του Πανεπιστημίου Πατρών Καθηγητής Γεώργιος Παναγιωτάκης, ενώ εκ μέρους του Δημάρχου Πατρέων απηύθυνε χαιρετισμό ο Αντιδήμαρχος Παιδείας, Διαφάνειας και Ηλεκτρονικής Διαβούλευσης Καθηγητής Ανδρέας Φίλιας. Επίσης, στην εναρκτήρια τελετή απηύθυναν χαιρετισμούς ο Πρόεδρος του Τμήματος Μαθηματικών Καθηγητής Παναγιώτης Πιντέλας, ο Πρόεδρος του Τμήματος Μηχανολόγων & Αεροναυπηγών Μηχανικών Καθηγητής Σπύρος Παντελάκης, ο Αναπληρωτής Πρόεδρος του Γενικού Τμήματος Καθηγητής Στάθης Περγίδης, ο Καθηγητής του Γενικού Τμήματος Ιωάννης Κουτρουβέλης εκ μέρους της Τοπικής Οργανωτικής Επιτροπής, και ο Πρόεδρος του ΕΣΙ Καθηγητής Χαράλαμπος Χαραλαμπίδης. Η τελετή έναρξης ολοκληρώθηκε με τις εναρκτήριες ομιλίες του Συνεδρίου που έδωσαν ο Καθηγητής του Τμήματος Μαθηματικών του Πανεπιστημίου Πατρών Ανδρέας Φιλίππου με τίτλο «Τα στατιστικά στοιχεία, η Ευρωζώνη και η κρίση χρέους της Ελλάδας»

και ο Επίκουρος Καθηγητής του Τμήματος Στατιστικής του Οικονομικού Πανεπιστημίου Αθηνών Παναγιώτης Μερκούρης με τίτλο «Διαφάνεια στην Στατιστική Πρακτική».

Μετά τους χαιρετισμούς και με την ευκαιρία της συμπλήρωσης τριακονταετίας από την ίδρυση του ΕΣΙ ο Καθηγητής Χαράλαμπος Χαραλαμπίδης απένειμε αναμνηστικά διπλώματα στους προηγούμενους Προέδρους του ΕΣΙ, κ.κ. *Κωνσταντίνο Δρακάτο, Θεόφιλο Κάκουλλο, Στρατή Κουνιά και Τάκη Παπαϊωάννου*, και εξήρε τη συνεισφορά τους στο ΕΣΙ.

Κατά την εναρκτήριο τελετή απονεμήθηκε και το *Ελέναιο Βραβείο Διδακτορικής Διατριβής στη Στατιστική* για τη διετία 2009-2010 στον κ. Ανδρέα Αρτεμίου, Assistant Professor, Michigan Technology University USA, ο οποίος την επόμενη μέρα παρουσίασε τη διατριβή του με τίτλο «Hyperplane alignment for sufficient dimension reduction: Implementation, application and advantages». Ο κ. Αρτεμίου εκπόνησε τη διατριβή του στο Pennsylvania State University με επιβλέποντα καθηγητή τον Bing Li. Για το βραβείο υποβλήθηκαν και οι ακόλουθες τέσσερις (4) διατριβές:

- Ονοματεπώνυμο: Ευάγγελος Βασιλείου
Τίτλος: Μέθοδοι προσαρμογής δεδομένων με διηρημένες διαφορές και εφαρμογές στην Οικονομετρία
Ίδρυμα: Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών, Τμήμα Οικονομικών Επιστημών
Επιβλέπων καθηγητής: Ιωάννης Δημητρίου
- Ονοματεπώνυμο: Σπύρος Δαφνής
Τίτλος: Κατανομές Σχηματισμών, Γενικεύσεις και Επεκτάσεις Κατανομών Ροών και Εφαρμογές
Ίδρυμα: Πανεπιστήμιο Πατρών, Τμήμα Μαθηματικών
Επιβλέπων καθηγητής: Ανδρέας Φιλίππου
- Ονοματεπώνυμο: Παναγιώτης Μπομποτάς
Τίτλος: Εκτιμητές τύπου Strawderman για παραμέτρους κλίμακας
Ίδρυμα: Πανεπιστήμιο Πατρών, Τμήμα Μαθηματικών
Επιβλέπων καθηγητής: Σταύρος Κουρούκλης
- Ονοματεπώνυμο: Μιλτώ Χατζηκυριάκου
Τίτλος: Probability and Moment Inequalities for Demimartingales and Associated Random Variables
Ίδρυμα: Πανεπιστήμιο Κύπρου, Τμήμα Μαθηματικών και Στατιστικής
Επιβλέπων καθηγητής: Τάσος Χριστοφίδης

Και οι πέντε διδακτορικές διατριβές ήταν πολύ αξιόλογες και καταδεικνύουν την, σε βάθος, γνώση της ευρύτερης περιοχής του θέματος της διατριβής από την πλευρά των υποψηφίων. Και οι πέντε υποστηρίζονται με δημοσιεύσεις σε διεθνή περιοδικά όπως *Statistica Sinica*, *Statistical Papers*, *JSPI*, *Statistics and Probability Letters* etc. Μέλη της Επιτροπής του Ελένιου Βραβείου Διδακτορικής Διατριβής ήταν ο Καθηγητής Ιωάννης Κουτρουβέλης, ο Αναπληρωτής Καθηγητής Δημήτρης Καρλής και ο Ομότιμος Καθηγητής Τάκης Παπαϊωάννου.

Για έκτη συνεχή χρονιά απονεμήθηκε το *Βραβείο Καλύτερης Εργασίας Νέου Στατιστικού*. Για το σκοπό αυτό υποβλήθηκαν τέσσερις (4) εργασίες, οι οποίες και παρουσιάστηκαν σε ξεχωριστή συνεδρία (βλ. Πρόγραμμα Συνεδρίου). Την τελευταία μέρα του Συνεδρίου η Τριμελής Επιτροπή, ύστερα από μελέτη των εργασιών και παρακολούθηση των παρουσιάσεων, κατέθεσε τη βαθμολογία της και το Βραβείο Καλύτερης Εργασίας Νέου Στατιστικού για το 2011 απονεμήθηκε στον Δημήτρη Λυμπερόπουλο, υποψήφιο διδάκτορα του Πανεπιστημίου Πειραιώς, για την από κοινού με τον Αναπληρωτή Καθηγητή Νικόλαο Μαχαιρά εργασία του με τίτλο: «Χαρακτηρισμοί μεμειγμένων διαδικασιών Poisson». Την Τετραμελή Επιτροπή αποτελούσαν οι συνάδελφοι *Τάσος Χριστοφίδης, Χαράλαμπος Χαραλαμπίδης, Χρόνης Μωυσιάδης και Ιωάννης Τζούφρας*.

Οι κοινωνικές εκδηλώσεις ήταν πλούσιες: Cocktail προσφέρθηκε την πρώτη μέρα του Συνεδρίου στο Συνεδριακό Κέντρο του Πανεπιστημίου Πατρών. Εκδρομή, για τα συνοδευόμενα μέλη, στην πόλη της Πάτρας, η οποία περιελάμβανε επισκέψεις στα αξιοθέατα της πόλης συμπεριλαμβανομένων του Αρχαιολογικού Μουσείου και του Οινοποιείου Achaia Clauss. Η εκδρομή για τους συνέδρους και τα συνοδευόμενα μέλη περιελάμβανε επίσκεψη στην Αρχαία Ολυμπία, και ξενάγηση στο μουσείο και στον αρχαιολογικό χώρο. Το βράδυ της ίδιας μέρας διοργανώθηκε το επίσημο δείπνο του συνεδρίου στο κέντρο «Ιφιγένεια», όπου έλαβαν μέρος αρκετοί σύνεδροι με έντονο κέφι μέχρι και τις πρώτες πρωινές ώρες.

Στον τόμο αυτό περιλαμβάνονται εργασίες που παρουσιάστηκαν στο Συνέδριο και υποβλήθηκαν για δημοσίευση. Όλες οι εργασίες κρίθηκαν από κριτές με την φροντίδα των υπευθύνων έκδοσης πρακτικών.

Οι παρατηρήσεις και τα σχόλια των κριτών, σύμφωνα με την πάγια πολιτική που ακολουθεί το Ε.Σ.Ι αφορούν κυρίως στον τρόπο παρουσίασης της εργασίας και στην παρουσία ή όχι τυπογραφικών και σοβαρών επιστημονικών λαθών, έτσι ώστε η δημοσίευση να γίνεται με ευθύνη των ίδιων των συγγραφέων. Για το σκοπό αυτό υπάρχουν κριτήρια δημοσίευσης εργασιών στα Πρακτικά του Ε.Σ.Ι. τα οποία είναι αναρτημένα στην ιστοσελίδα του Ε.Σ.Ι.: www.esi.stat.gr. Όλες οι εργασίες, για τις οποίες

ζητήθηκε αναθεώρηση, κρίθηκαν εκ νέου από τους κριτές ή από τους υπεύθυνους έκδοσης των πρακτικών.

Συνολικά υποβλήθηκαν σαράντα έξι (46) εργασίες, από τις οποίες μία (1) εργασία ανακλήθηκε από τους συγγραφείς μετά την πρώτη αξιολόγηση και μία (1) απορρίφθηκε, σύμφωνα με απόφαση του Δ.Σ. και ύστερα από σχετική πρόταση των κριτών και των υπευθύνων της έκδοσης των Πρακτικών. Ως κριτές των εργασιών συνεργάστηκαν οι: Ε. Αγγελής, Ι. Ανδρεάδης, Ι. Αντωνίου, Α. Αρτεμίου, Μ. Βαμβακάρη, Β. Βασδέκης, Ι. Βόντα, Χ. Δαμιανού, Τ. Δάρας, Γ. Δονάτος, Ν. Δριτσάκης, Κ. Ζωγράφος, Δ. Ιωαννίδης, Θ. Κάκουλλος, Δ. Καρλής, Μ. Κατέρη, Φ. Κολυβά-Μαχαίρα, Δ. Κουγιουμτζής, Χ. Κουκουβίνος, Σ. Κουνιάς, Σ. Κουρούκλης, Ι. Κουτροβέλης, Ε. Κυριακίδης, Κ. Μακρής, Σ. Μαλεφάκη, Π. Μαραβελάκης, Λ. Μελιγκοτσίδου, Π. Μερκούρης, Μοσχονά Θεανώ-Εριφύλη, Ι. Μπασιάκος, Α. Μπατσίδης, Μ. Μπούτσικας Α. Νικολουλόπουλος, Δ. Παναγιωτάκος, Ν. Παπαδάτος, Α. Παπαδοπούλου, Τ. Παπαϊωάννου, Δ. Παπαναστασίου, Χ. Παυλόπουλος, Δ. Πόταρη, Μ. Σφακιανάκης, Γ. Τσακλίδης, Π. Τσιαμυρτζής, Ν. Φαρμάκης, Κ. Φερεντίνος, Α. Φιλίππου, Χ. Χαραλαμπίδης. Προς όλους αυτούς η Επιτροπή Έκδοσης Πρακτικών του ΕΣΙ εκφράζει τις ευχαριστίες της για την επιμελημένη και προσεκτική αξιολόγηση των εργασιών.

Η σειρά παρουσίασης των εργασιών στον παρόντα τόμο είναι αλφαβητική με βάση το επώνυμο του πρώτου συγγραφέα.

Το Διοικητικό Συμβούλιο του Ε.Σ.Ι αισθάνεται την ανάγκη να ευχαριστήσει το Τμήμα Μαθηματικών, το Γενικό Τμήμα και το Τμήμα Μηχανολόγων & Αεροναυπηγών Μηχανικών του Πανεπιστημίου Πατρών, τους χορηγούς και την Τοπική Οργανωτική Επιτροπή για την πολύ καλή οργάνωση και προσφορά τους.

Τέλος το ΔΣ του Ε.Σ.Ι. εκφράζει τις ευχαριστίες του στη Γραμματεία του ΕΣΙ Μαρία Χιώλου για την τεχνική επιμέλεια της έκδοσης των Πρακτικών και των CD.

ΕΚ ΜΕΡΟΥΣ ΤΟΥ Δ.Σ. ΤΟΥ Ε.Σ.Ι

Οι υπεύθυνοι Έκδοσης Πρακτικών 24^ο Συνεδρίου

Αλέξανδρος Καραγρηγορίου
Χρόνης Μουσιάδης

Στρατής Κουνιάς
Τάκης Παπαϊωάννου

ΧΟΡΗΓΟΙ ΣΥΝΕΔΡΙΟΥ



ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ
ΤΜΗΜΑ ΜΑΘΗΜΑΤΙΚΩΝ,
ΓΕΝΙΚΟ ΤΜΗΜΑ,
ΤΜΗΜΑ ΜΗΧΑΝΟΛΟΓΩΝ
& ΑΕΡΟΝΑΥΠΗΓΩΝ ΜΗΧΑΝΙΚΩΝ



ΠΡΟΓΡΑΜΜΑ ΣΥΝΕΔΡΙΟΥ

Τετάρτη, 27 Απριλίου 2011

ΣΥΝΕΔΡΙΑΚΟ & ΠΟΛΙΤΙΣΤΙΚΟ ΚΕΝΤΡΟ ΠΑΝΕΠΙΣΤΗΜΙΟΥ ΠΑΤΡΩΝ Αμφιθέατρο Ι-4	
15:00-16:45	Εγγραφή Συνέδρων Διανομή συνεδριακού υλικού
17:00-17:30	Έναρξη του συνεδρίου Χαιρετισμοί
17:30	ΕΝΑΡΚΤΗΡΙΕΣ ΟΜΙΛΙΕΣ Προεδρεύων: Τ. Παπαϊωάννου
17:30-18:05	Α. Φιλίππου Τα στατιστικά στοιχεία, η Ευρωζώνη και η κρίση χρέους της Ελλάδας
18:05-18:40	Τ. Μερκούρης Διαφάνεια στην Στατιστική Πρακτική
Αίθουσα Α	
	ΣΤΑΤΙΣΤΙΚΗ Προεδρεύων: Τ. Μερκούρης
19:00-19:20	Σ. Κουρούκλης Η δειγματική διασπορά και το τίμημα της αμεροληψίας της
19:20-19:40	Π. Μπομποτάς, Σ. Κουρούκλης Μία νέα κλάση εκτιμητών για τον πίνακα διασπορών-συνδιασπορών πολυμεταβλητής κανονικής κατανομής
19:40-20:00	Α. Μπατσίδης, Κ. Ζωγράφος Ένας τρόπος ελέγχου αποκλίσεων από συγκεκριμένους ελλειπτικούς πληθυσμούς

20:00-20:20	Π. Οικονόμου, Γ. Τζαβελάς Διαδικασίες αναγνώρισης μεροληπτικών ως προς το μέγεθος δειγματοληπτικών σχεδιασμών
20:20-20:40	Ι.Κ. Δημητρίου, Σ.Σ. Παπακωνσταντίνου Ένα θεώρημα χαρακτηρισμού της L1 γραμμικής παλινδρόμησης με ανισοτικούς περιορισμούς
Αίθουσα Β	ΒΙΟΣΤΑΤΙΣΤΙΚΗ - ΒΙΟΜΕΤΡΙΑ Προεδρεύων: Δ. Παναγιωτάκος
19:00-19:20	Λ. Μπαντής, Γ. Τσιμήκας, Σ. Γεωργίου Εκτίμηση της συνάρτησης επιβίωσης με χρήση μονότονων κυβικών spline
19:20-19:40	Ε. Κετζάκη, Ν. Φαρμάκης, Δ. Κετζάκη Σύγκριση των δεικτών ποικιλότητας χρησιμοποιώντας την μέθοδο Jackknife
19:40-20:00	Ε. Ματθαίου, Ν. Φαρμάκης, Ε. Λοσιάδη, Φ. Αθανασιάδου-Πιπεροπούλου Ανάλυση τάσης στην εμφάνιση παιδικού καρκίνου στη βόρεια Ελλάδα Εμπειρία 21 ετών: 1990-2010
20:00-20:20	Γ. Πούνης, Δ.Β. Παναγιωτάκος Περιγραφική μελέτη για την αξιολόγηση της εφαρμογής των στατιστικών μεθόδων σε δείγμα δημοσιευμένων άρθρων σε έγκριτο διατροφικό περιοδικό
20:20-20:40	Β. Μπουντζιούκα, Δ.Β. Παναγιωτάκος Η επίδραση της περιστροφής στην επαναληψιμότητα των διατροφικών προτύπων που προκύπτουν από την εφαρμογή της ανάλυσης σε κύριες συνιστώσες

Αίθουσα Γ	ΣΤΟΧΑΣΤΙΚΕΣ ΔΙΑΔΙΚΑΣΙΕΣ - ΕΦΑΡΜΟΓΕΣ Προεδρεύων: Ν. Τσάντας
19:00-19:20	Σ. Γεωργιάδης, Ν. Αημιός Πολυδιάστατο συναρτησιακό Κεντρικό Οριακό Θεώρημα για τον εμπειρικό εκτιμητή ενός ημιμαρκοβιανού πυρήνα
19:20-19:40	Ε. Βότση, Ν. Αημιός, Γ. Τσακλίδης, Ε. Παπαδημητρίου Κρυφά μαρκοβιανά και ημιμαρκοβιανά μοντέλα για τον εντοπισμό του υποκείμενου κατά τη γένεση σεισμών πεδίου τάσεων
19:40-20:00	Γ. Ηλιόπουλος, Σ. Μαλεφάκη Εκτίμηση συναρτησιακών της οριακής κατανομής ημιμαρκοβιανών διαδικασιών βάσει μοντέλων για τους μέσους χρόνους παραμονής
20:00-20:20	V.P. Koutras, S. Malefaki, A.N. Platis Dependability analysis of a software rejuvenation model based on Monte Carlo simulation
20:20-20:40	Μ. Σωτηρόπουλος Εφαρμογή της Μαθηματικής Θεωρίας Εννοιών σε στοχαστικούς χώρους
21:00	<i>ΔΕΞΙΩΣΗ στο κεντρικό foyer του ΣΥΝΕΔΡΙΑΚΟΥ & ΠΟΛΙΤΙΣΤΙΚΟΥ ΚΕΝΤΡΟΥ</i>

Πέμπτη , 28 Απριλίου 2011

ΚΤΗΡΙΟ ΠΡΥΤΑΝΕΙΑΣ ΠΑΝΕΠΙΣΤΗΜΙΟΥ ΠΑΤΡΩΝ

Αίθουσα Τελετών	ΠΡΟΣΚΕΚΛΗΜΕΝΟΙ ΟΜΙΛΗΤΕΣ Προεδρεύων: Α. Φιλίππου
9:00-9:35	Γ. Δονάτος Διαφάνεια, δεοντολογία και αποτελεσματικότητα στη Δημόσια Διοίκηση
9:35-10:10	Τ. Χριστοφίδης Δεοντολογία-διαφάνεια και εμπιστευτικότητα της πληροφορίας στις δειγματοληπτικές έρευνες
10:10-10:45	Δ. Κυριακίδου Δεοντολογία-διαφάνεια στις επίσημες στατιστικές της Κύπρου
Αίθουσα Τελετών	ΕΛΕΝΕΙΟ ΒΡΑΒΕΙΟ ΔΙΔΑΚΤΟΡΙΚΗΣ ΔΙΑΤΡΙΒΗΣ Προεδρεύων: Ι. Κουτρουβέλης
	Απονομή Ελένειου Βραβείου Διδακτορικής Διατριβής στη Στατιστική της διετίας 2009-2010
10:45-11:15	A. Artemiou Hyperplane alignment for sufficient dimension reduction: Implementation, application, and advantages
11:15-12:00	ΔΙΑΛΕΙΜΜΑ - ΠΕΡΙΠΑΤΗΜΑ ΜΕΧΡΙ ΤΟ ΣΥΝΕΔΡΙΑΚΟ ΚΕΝΤΡΟ - ΚΑΦΕΣ

ΣΥΝΕΔΡΙΑΚΟ & ΠΟΛΙΤΙΣΤΙΚΟ ΚΕΝΤΡΟ ΠΑΝΕΠΙΣΤΗΜΙΟΥ ΠΑΤΡΩΝ

Αίθουσα Α	ΣΤΑΤΙΣΤΙΚΗ Προεδρεύων: Σ. Κουρούκλης
12:00-12:20	Κ. Ζωγράφος Μέτρα και έννοιες στοχαστικής εξάρτησης
12:20-12:40	D. Karlis Model based clustering via copulas
12:40-13:00	P.N. Patil and D. Bagkavos A measure of asymmetry
13:00-13:20	Γ. Κουτσογέρας, Τ. Παπαϊωάννου Μέτρα συνάφειας και μέτρα συμμετρίας-ασυμμετρίας για πίνακες συνάφειας
13:20-13:40	Γ. Ηλιόπουλος, N. Balakrishnan Ακριβής συμπερασματολογία για τις παραμέτρους της κατανομής Laplace
Αίθουσα Β	ΣΤΑΤΙΣΤΙΚΗ ΚΑΙ ΔΙΑΔΙΚΤΥΟ - ΥΠΟΛΟΓΙΣΤΙΚΗ ΣΤΑΤΙΣΤΙΚΗ Προεδρεύων: Φ. Αλεβίζος
12:00-12:20	Θ. Χατζηπαντελής, Ι. Ανδρεάδης Ηλεκτρονικός σύμβουλος ψήφου: Μια ηλεκτρονική δημοσκοπήση ή ένα εργαλείο βοήθειας επιλογής και ανάλυσης;
12:20-12:40	Σ. Πουπάκης Σύγκριση υποκειμενικών απαντήσεων σε έρευνες κοινής γνώμης και τρόποι αντιμετώπισης
12:40-13:00	Ν. Φαρμάκης, Γ. Μακρής Δειγματοληψία στο Διαδίκτυο: Δυνατότητες ανάκτησης συγκεκριμένης πληροφορίας

13:00-13:20	V. Georgiou, B. Quost, T. Denœux A novel approach on improving the reliability of classifiers using contextual discounting
13:20-13:40	Γ. Μακρής, Ε. Δαλαμήτρα, Ι. Αντωνίου Κατασκευή μοντέλου για την εκτίμηση της επιλογής κατεύθυνσης σπουδών των μαθητών της Β Λυκείου
Αίθουσα Γ	ΣΤΟΧΑΣΤΙΚΕΣ ΔΙΑΔΙΚΑΣΙΕΣ - ΕΦΑΡΜΟΓΕΣ Προεδρεύων: Γ. Τσακλίδης
12:00-12:20	Μ. Παπαπέτρον, Δ. Κουγιουμτζής Εκτίμηση τάξης της μαρκοβιανής αλυσίδας μέσω της δεσμευμένης αμοιβαίας πληροφορίας
12:20-12:40	Γ. Βασιλειάδης Ομογενές μαρκοβιανό σύστημα διακριτού χρόνου και ουρά διακριτού χρόνου με πεπερασμένη πηγή
12:40-13:00	Β. Δημητρίου, Ν. Τσάντας Ελέγχοντας μαθηματικά μοντέλα παρακαταθήκης για συστήματα ανθρώπινου δυναμικού
13:00-13:20	Ι.Χ. Δημητρίου Ένα σύστημα εξυπηρέτησης με επαναλαμβανόμενους πελάτες, προτεραιότητες, εξαρτημένη ροή αφίξεων και βλάβες υπαλλήλου
13:20-13:40	Γ. Καραβασίλης, Α. Ρήγας Αναγνώριση στοχαστικών συστημάτων στάσιμων σημειακών ανελιξέων

13:40-16:00 **ΜΕΣΗΜΒΡΙΝΗ ΔΙΑΚΟΠΗ**

Αίθουσα Α	ΒΡΑΒΕΙΟ ΝΕΟΥ ΣΤΑΤΙΣΤΙΚΟΥ Προεδρεύων: Τ. Χριστοφίδης
16:00-16:20	G. Mytalas, M. Zazanis A central limit theorem for the number of successes and failures in ruins
16:20-16:40	T. Xifara, C. Sherlock Adaptive MCMC using a hidden Markov model for disease interactions
16:40-17:00	Δ. Λυμπερόπουλος, Ν. Μαχαιράς Χαρακτηρισμοί μειγμένων διαδικασιών Poisson
17:00-17:20	Ε. Σιγγιρίδου, Δ. Κουγιουμτζής, Ε. Μπόρα-Σέντα Διερεύνηση της στατιστικής σημαντικότητας της συσχέτισης μεταξύ χρονοσειρών για τη δημιουργία δικτύων

Αίθουσα Β	ΠΕΙΡΑΜΑΤΙΚΟΙ ΣΧΕΔΙΑΣΜΟΙ Προεδρεύων: Ν. Φαρμάκης
16:00-16:20	Κ. Περικλέους, Σ. Κουνιάς Βέλτιστοι σχεδιασμοί γραμμής με τρεις αγωγές και εξαρτημένες παρατηρήσεις
16:20-16:40	Φ. Κολυβά-Μαχαίρα, Σ.Α. Χατζόπουλος Τύπου 1 βέλτιστοι 3 ^m παραγοντικοί πειραματικοί σχεδιασμοί με $N \equiv -1 \pmod{9}$ παρατηρήσεις
16:40-17:00	Α. Σκούντζου, Χ. Κουκουβίνος Μέθοδος επιλογής μεταβλητών σε υπερκορεσμένου σχεδιασμούς
17:00-17:20	Μ. Χαλικιάς, Σ. Κουνιάς Σχεδιασμοί επαναλαμβανόμενων μετρήσεων, αναγκαίες και ικανές συνθήκες βελτιστοποίησης

Αίθουσα Γ	ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ Προεδρεύων: Μ. Μανατάκης
16:00-16:20	Α. Καράκος, Θ. Βουγιουκλής, Π. Καμπάκη-Βουγιουκλή Αντικατάσταση της κλίμακας Likert από τη ράβδο
16:20-16:40	Γ. Ματαλλιωτάκης, Χ. Σκιαδάς Μοντελοποίηση δεδομένων επιβίωσης και συγκριτική ανάλυση μεταξύ δυναμικών και αιτιοκρατικών μοντέλων
16:40-17:00	Υ.С. Bassiakos Brain drain and brain gain within EU27: current picture and possible reasons
17:00-17:20	P. Prodromidis Treating the regression misspecification problem through space. Analyzing employment and unemployment evidence from Greece

17:20-17:40 **ΔΙΑΔΕΙΜΜΑ - ΚΑΦΕΣ**

Αίθουσα Α	ΠΙΘΑΝΟΤΗΤΕΣ Προεδρεύων: Σ. Κουνιάς
17:40-18:00	Α. Kyriakoussis, Μ. Vamvakari On q-hypergeometric series distributions and the associated q-Racah orthogonal polynomials
18:00-18:20	Charalambides Ch.A. A q-Polya urn model and the q-Polya and inverse q-Polya distributions
18:20-18:40	Μ. Hadjikyriakou, T.C. Christofides Probability and moment inequalities for demimartingales and related results
18:40-19:00	Σ. Καποδίστρια, Γ. Ψαρράκος Μελέτη της αθροιστικής υπολειπόμενης εντροπίας σε μια οικογένεια σταθμισμένων κατανομών

19:00-19:20	Ι.Σ. Τριανταφύλλου, Μ.Β. Κούτρας Μελέτη ιδιοτήτων μιας γενικής κλάσης τυχαίων μεταβλητών που σχετίζονται με κατανομές διατεταγμένων παρατηρήσεων
Αίθουσα Β	ΠΡΟΒΛΕΨΗ - ΧΡΟΝΟΣΕΙΡΕΣ Προεδρεύων: Ι. Δημητρίου
17:40-18:00	Α. Παπάνα, Δ. Κουγιουμτζής Εντοπισμός άμεσων συζεύξεων σε πολυμεταβλητά συστήματα
18:00-18:20	Α. Σκουρκέας, Φ. Κολυβά-Μαχαίρα, Π. Μαχαίρας Διερεύνηση του ρόλου των τάσεων στην ανάπτυξη μοντέλων και σεναρίων κλιματικών παραμέτρων
18:20-18:40	H. Haralambous, H. Papadopoulos, E. Tsolaki Forecasting Total Electron Content (TEC) over Cyprus
18:40-19:00	N. Δριτσάκης Ισοτιμία αγοραστικής δύναμης και έλεγχοι μοναδιαίας ρίζας με πάνελ δεδομένα
19:00-19:20	X. Κάτρης, Σ. Δασκαλάκη Προβλεπτικά μοντέλα χρονοσειρών με εξάρτηση μακράς διάρκειας: Εφαρμογή στο Γενικό Δείκτη του Χρηματιστηρίου Αξιών Αθηνών
Αίθουσα Γ	ΣΤΑΤΙΣΤΙΚΗ Προεδρεύων: Φ. Μακρή
17:40-18:00	Α. Ρεκούτη, Φ. Αλεβίζος Εφαρμογή μεθόδων πολυμεταβλητής στατιστικής ανάλυσης για την ανίχνευση & περιγραφή των προτιμήσεων του Ελληνικού πληθυσμού ως προς την κατανάλωση αλκοολούχων ποτών
18:00-18:20	Ι. Θεοδώρου, Φ. Αλεβίζος, Α. Κεχρινιώτης Ανάλυση Αντιστοιχιών με Ασαφή Δεδομένα: Η πρακτική εφαρμογή της μεθόδου (αντιπροσωπευτικό παράδειγμα με αποασαφοποιημένη παραγοντική γεωμετρική παράσταση)

18:20-18:40	Β.Γ. Βασιλειάδης, Ε.Ν. Σπανού, Α.Γ. Ρήγας Βελτίωση της εκτίμησης των παραμέτρων ενός γενικευμένου γραμμικού μοντέλου με τη χρήση της παλινδρόμησης Ridge
18:40-19:00	Α.Χ. Ρακιτζής, Δ.Α. Αντζουλάκος Τεστ εκκίνησης αποδοχής εξαρτημάτων με δύο επίπεδα αποτυχίας
19:00-19:20	Θ. Κάκουλλος Η γεωμετρία των ελαχίστων τετραγώνων

Παρασκευή, 29 Απριλίου 2011

08:15-17:30	ΕΚΔΡΟΜΗ (Excursion) <i>Αρχαία Ολυμπία</i>
20:30	ΕΠΙΣΗΜΟ ΔΕΙΠΝΟ (Official Dinner)

Σάββατο, 30 Απριλίου 2011

ΣΥΝΕΔΡΙΑΚΟ & ΠΟΛΙΤΙΣΤΙΚΟ ΚΕΝΤΡΟ ΠΑΝΕΠΙΣΤΗΜΙΟΥ ΠΑΤΡΩΝ	
Αίθουσα Α	ΠΙΘΑΝΟΤΗΤΕΣ Προεδρεύων: Χ. Χαραλαμπίδης
10:00-10:20	Γ. Αφένδρας, Ν. Παπαδάτος Ισχυροποιημένη ανισότητα τύπου Chernoff και εκτεταμένα τύπου Poisson φράγματα διασποράς σε υποοικογένεια των κατανομών Pearson
10:20-10:40	Χ.Α. Νικολαΐδου, Β.Ε. Πιπερίγκου Μελέτη της διδιάστατης Geometrico-Poisson κατανομής, συναφών κατανομών και εφαρμογές αυτών

10:40-11:00	Φ.Σ. Μακρή, Ζ.Μ. Ψυλλάκης Συμβολοσειρές φραγμένου μήκους σε δυαδικές ακολουθίες
11:00-11:20	Φ.Σ. Μηλιένος, Μ.Β. Κούτρας Ασυμπτωτικά αποτελέσματα για χρόνους αναμονής σχηματισμών
11:20-11:40	A.N. Philippou, S.D. Dafnis, F.S. Makri On (k_1, k_2) events in Markov dependent trials

Αίθουσα Β	ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ Προεδρεύων: Γ. Δονάτος
------------------	---

10:00-10:20	Κ. Κιτικίδου, Κ. Μάτης Εφαρμογή της ανάλυσης επιβίωσης για την αξιολόγηση της αύξησης των δέντρων ελάτη στο Πανεπιστημιακό δάσος Περτουλίου
10:20-10:40	Ζ. Νικηφορίδου, Τ. Παγγέ, Θ. Χατζηπαντελής Στρατηγικές ανάληψης ρίσκου από παιδιά και ενήλικες
10:40-11:00	Μ. Χαλικιάς, Θ.Ε. Μοσχονά Ομαδοποίηση ελληνικών τουριστικών προορισμών σε σχέση με οικονομικοκοινωνικά χαρακτηριστικά ταξιδιωτών
11:00-11:20	Β. Χομπάς Τα γκολ ανά παιχνίδι στο Παγκόσμιο Κύπελλο ποδοσφαίρου του 2010 ακολουθούν την κατανομή Poisson;

11:40-12:00	ΔΙΑΛΕΙΜΜΑ - ΚΑΦΕΣ
--------------------	--------------------------

12:00-13:00

POSTERS (Αναρτημένες εργασίες) *

1

D. Ioannides, E. Matzner-Lober

Conditional quantiles with measurement errors

2

S. Meintanis

Monitoring changes in autoregressive models

3

S. Pouloupoulou, D. Karlis, U. Dafni

Adaptive phase II design for evaluating survival probabilities

4

Α. Γκιούλος, Γ. Τσακλίδης, Γ. Σταματέλλος

Εκτίμηση της κοινής κατανομής στηθιαίας διαμέτρου και ύψους δασικών δέντρων με τη χρήση συζεύξεων (copulas)

5

Γ. Καραβασίλης, Α. Ρήγας, Α. Κύδρος

Βασικά χαρακτηριστικά και ιδιαιτερότητες της στατιστικής γλώσσας προγραμματισμού R κατά τον υπολογισμό συναρτήσεων γινομένου και συσσώρευσης στάσιμων σημειακών ανελίξεων

6

Μ. Μανατάκης, Π. Κατσακιώρη

Στατιστική ανάλυση παραγόντων πρόκλησης θανατηφόρων εργατικών ατυχημάτων

7

Γ. Ματαλλιωτάκης, Χ. Σκιαδάς, Κ. Ζοπουνίδης

Δημοσκόπηση για τις ανανεώσιμες πηγές ενέργειας και την πράσινη ανάπτυξη

8

Μ.Β. Μπούτσικας, Α.Χ. Ρακιτζής, Δ.Α. Αντζουλάκος

Περί της κατανομής του αθροίσματος διακοπής: Εφαρμογές στον Ποιοτικό Έλεγχο

9

Δ. Στογιάννης, Χ. Καρώνη

Διερεύνηση προσαρμογής μοντέλου παλινδρόμησης χρόνου πρώτης μετάβασης

10

Θ. Θεοδοσίου, Σ. Βαλσαμίδης, Γ. Φλώρου, Α. Καραύκος

Ανάλυση δεδομένων σε σύστημα ηλεκτρονικής ψηφοφορίας

13:00-15:30	ΜΕΣΗΜΒΡΙΝΗ ΔΙΑΚΟΠΗ
Αίθουσα Α	ΣΤΑΤΙΣΤΙΚΗ Προεδρεύων: Θ. Κάκουλλος
15:30-15:50	M. Viana Fourier Analysis of symmetry preference data
15:50-16:10	Κ. Πετρόπουλος Κλάσεις βελτιωμένων διαστημάτων εμπιστοσύνης για την παράμετρο κλίμακας από την πολυδιάστατη κατανομή Lomax
16:10-16:30	Π. Οικονόμου, Ι.Α. Κουτρουβέλης Γραφικοί έλεγχοι καλής προσαρμογής για κατανομές θέσης-κλίμακας με βάση συναρτήσεις ημιαναλλοιώτων
16:30-16:50	Μ. Δουλή, Γ. Τζαβελάς Μελέτη της ασυμπτωτικής αποδοτικότητας του εκτιμητή ημιπιθανοφάνειας
16:50-17:10	Β.Α. Παππάς, Κ. Αδαμίδης, Σ. Λουκάς Μια οικογένεια κατανομών στην ανάλυση επιβίωσης
Αίθουσα Β	ΑΝΑΛΟΓΙΣΜΟΣ - ΑΣΦΑΛΙΣΗ Προεδρεύων: Κ. Κουτσόπουλος
15:30-15:50	Β. Δερμιτζάκης, Κ. Πολίτης Μονοτονία και ασυμπτωτική συμπεριφορά για τη λύση μιας ελλειμματικής ανανεωτικής εξίσωσης
15:50-16:10	A.G. Bardoutsos, D.G. Konstantinides Structure of subclasses in the frame of heavy-tailed distributions
16:10-16:30	C.H. Skiadas A methodology for estimating life expectancy and related applications

16:30-16:50

D.G. Konstantinides, C.E. Kountzakis

The restricted convex risk measures in premium calculation

16:50-17:10

Γ. Μαυριδόγλου

Ένα αναλογιστικό υπόδειγμα για την ασφάλιση εισοδήματος από ασθένεια

17:30

ΛΗΞΗ ΣΥΝΕΔΡΙΟΥ

στην Αίθουσα Α

- 1) Απονομή Βραβείου Καλύτερης Εργασίας Νέου Στατιστικού
- 2) Προτάσεις για το Συνέδριο Στατιστικής του 2012
- 3) Συζήτηση

ΚΟΙΝΩΝΙΚΕΣ ΕΚΔΗΛΩΣΕΙΣ & Πρόγραμμα Μετακινήσεων

Τετάρτη 27 Απριλίου 2011

Δεξίωση: Συνεδριακό & Πολιτιστικό Κέντρο Πανεπιστημίου Πατρών
Ωρα: 21:00

Παρασκευή 29 Απριλίου 2011

Επίσημο Δείπνο Συνεδρίου
Προορισμός: Κέντρο «Ιφιγένεια»
Ωρα: 20:30
Περιλαμβάνεται στο δικαίωμα συμμετοχής.

ΕΚΔΡΟΜΕΣ

Πέμπτη 28 Απριλίου 2011
Εκδρομή για συνοδεύοντα μέλη

Πρόγραμμα: Αναχώρηση από το κτήριο της Πρυτανείας, στις 09:30πμ
Επίσκεψη στο Αρχαιολογικό Μουσείο (είσοδος ελεύθερη)
Περιήγηση στην πόλη της Πάτρας
Επίσκεψη στο οινοποιείο της “Achaia Claus”

Επιστροφή στο Πανεπιστήμιο περίπου στις 13:30.
Περιλαμβάνεται στο δικαίωμα συμμετοχής. Η εκδρομή δεν περιλαμβάνει φαγητό.

Παρασκευή 29 Απριλίου 2011
Ολοήμερη εκδρομή για συνέδρους και συνοδεύοντα μέλη

Πρόγραμμα: Αναχώρηση από το Συνεδριακό Κέντρο, στις 08.15 πμ.
Ξενάγηση στον αρχαιολογικό χώρο και το μουσείο της Αρχαίας Ολυμπίας.

Επιστροφή στο Πανεπιστήμιο, περίπου στις 17:30.
Περιλαμβάνεται στο δικαίωμα συμμετοχής. Η εκδρομή δεν περιλαμβάνει φαγητό, ούτε την είσοδο στον αρχαιολογικό χώρο και το μουσείο.

ΕΠΙΣΤΗΜΟΝΙΚΗ ΕΠΙΤΡΟΠΗ

- Δ. Κωνσταντινίδης**, Πανεπιστήμιο Αιγαίου
Α. Φιλίππου, Πανεπιστήμιο Πατρών
Χ. Χαραλαμπίδης, Πανεπιστήμιο Αθηνών
Τ. Μερκούρης, Οικονομικό Πανεπιστήμιο Αθηνών
Ν. Φαρμάκης, Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης
Χ. Μωυσιάδης, Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης
Τ. Χριστοφίδης, Πανεπιστήμιο Κύπρου
Μ. Κατέρη, Πανεπιστήμιο Ιωαννίνων
Σ. Κουνιάς, Πανεπιστήμιο Αθηνών
Α. Κυριακούσης, Χαροκόπειο Πανεπιστήμιο
Τ. Παπαϊωάννου, Πανεπιστήμια Πειραιώς και Ιωαννίνων (Πρόεδρος)
Δ. Ιωαννίδης, Πανεπιστήμιο Μακεδονίας
Α. Ρήγας, Δημοκρίτειο Πανεπιστήμιο Θράκης
Γ. Δονάτος, Πανεπιστήμιο Αθηνών
Ι. Κουτρουβέλης, Πανεπιστήμιο Πατρών

ΟΡΓΑΝΩΤΙΚΗ ΕΠΙΤΡΟΠΗ

- Γ. Ηλιόπουλος**, Πανεπιστήμιο Πειραιώς
Σ. Κουρούκλης, Πανεπιστήμιο Πατρών
Ι. Κουτρουβέλης, Πανεπιστήμιο Πατρών
Σ. Μαλεφάκη, Πανεπιστήμιο Πατρών
Μ. Μανατάκης, Πανεπιστήμιο Πατρών
Π. Οικονόμου, Πανεπιστήμιο Πατρών
Κ. Πετρόπουλος, Πανεπιστήμιο Πατρών
Β. Πιπερίγκου, Πανεπιστήμιο Πατρών

ΕΠΙΤΡΟΠΗ ΠΡΟΓΡΑΜΜΑΤΟΣ

- Γ. Ηλιόπουλος**, Πανεπιστήμιο Πειραιώς
Α. Καραγρηγορίου, Πανεπιστήμιο Κύπρου
Τ. Παπαϊωάννου, Πανεπιστήμια Πειραιώς και Ιωαννίνων
Χ. Χαραλαμπίδης, Πανεπιστήμιο Αθηνών

ΔΙΟΙΚΗΤΙΚΟ ΣΥΜΒΟΥΛΙΟ Ε.Σ.Ι.

- Πρόεδρος:** Χαράλαμπος Χαραλαμπίδης
Αντιπρόεδρος: Χρόνης Μωυσιάδης
Γενικός Γραμματέας: Αλέξανδρος Καραγρηγορίου
Ταμίας: Γεώργιος Ηλιόπουλος
Ειδικός Γραμματέας: Θεανώ Μοσχονά
Έφορος Βιβλιοθήκης: Τάκης Παπαϊωάννου
Μέλος: Στρατής Κουνιάς

εργασίες

Ύστερα από πρόσκληση

Ολομέλεια



ΔΙΑΦΑΝΕΙΑ, ΔΕΟΝΤΟΛΟΓΙΑ ΚΑΙ ΑΠΟΤΕΛΕΣΜΑΤΙΚΟΤΗΤΑ ΣΤΟΝ ΔΗΜΟΣΙΟ ΤΟΜΕΑ ΚΑΙ Η ΣΤΑΤΙΣΤΙΚΗ

Γεώργιος Δονάτος
Πανεπιστήμιο Αθηνών
gdonat@econ.uoa.gr

ΠΕΡΙΛΗΨΗ

Η αξιολόγηση της αποτελεσματικότητας του Δημοσίου Τομέα είναι αναγκαία διαδικασία πριν από κάθε μεταρρύθμιση. Η πληρότητα και η ποιότητα των στατιστικών στοιχείων και η τήρηση της διαφάνειας και των άλλων αρχών δεοντολογίας, κατά τη συγκέντρωσή τους και σε όλα τα στάδια της στατιστικής ανάλυσής τους, αποτελούν προϋπόθεση για την μη αμφισβήτηση των αποτελεσμάτων της αποτίμησης της αποτελεσματικότητας. Για το λόγο αυτό στο άρθρο προτείνεται η δημιουργία μιας συνεχώς ενημερωμένης αυτοματοποιημένης Τράπεζας Δεδομένων που θα αποτελεί τη βάση ενός επισήμου και ανεξάρτητου Πληροφοριακού Συστήματος Διοίκησης. Ακόμη, στην εργασία αυτή, επιχειρείται να αναδειχθούν ορισμένα ζητήματα, ερωτήματα και διαπιστώσεις που προκύπτουν από την δειγματοληπτική έρευνα και τις εμπειρικές μελέτες, που πραγματοποιήθηκαν για τον ελληνικό Δημόσιο Τομέα, και ναδειχθεί ότι η στατιστική επιστήμη διασφαλίζει: (α) Τη διαφάνεια και την τήρηση των βασικών αρχών δεοντολογίας στη συγκέντρωση και ανάλυση των στοιχείων και στην ερμηνεία και αξιοποίηση των αποτελεσμάτων των εμπειρικών μελετών. (β) Τη διάθεση των κατάλληλων ποσοτικών και ποιοτικών στοιχείων και συμβάλλει ουσιαστικά: (i) Στον έλεγχο και στην ερμηνεία της συμπεριφοράς και της παραγωγικότητας του ανθρώπινου δυναμικού και της Διοίκησης. (ii) Στην τεκμηριωμένη και αξιόπιστη αξιολόγηση της αποτελεσματικότητας του Δημοσίου Τομέα, με βάση τους κανόνες της επιστήμης και της δεοντολογίας. (iii) Στη λήψη των βέλτιστων αποφάσεων για τον εκσυγχρονισμό του Δημοσίου τομέα και την εφαρμογή αναπτυξιακών πολιτικών.

Λέξεις Κλειδιά: Διαφάνεια, Δεοντολογία, Αποτελεσματικότητα, Δημόσιος Τομέας, Πληροφοριακό Σύστημα Διοίκησης, Επιστημονική Γνώση, Ανθρώπινο Κεφάλαιο, Ηγετικά Χαρακτηριστικά.

1. ΕΙΣΑΓΩΓΗ

Σε ένα συνεχώς μεταβαλλόμενο κοινωνικοοικονομικό περιβάλλον, που εξελίσσεται με ταχύτατους ρυθμούς, σημαντικά λειτουργικά προβλήματα απασχολούν τις Δημόσιες Υπηρεσίες και Επιχειρήσεις και τους Δημόσιους

Οργανισμούς που αφορούν τις δαπάνες, την παραγωγικότητα, την ποιότητα των υπηρεσιών κλπ. Οι υπεύθυνοι Διοίκησης αυτών των Δημοσίων Υπηρεσιών, Επιχειρήσεων και Οργανισμών θεωρούν ότι τα προβλήματα αυτά οφείλονται κυρίως στην ανεπάρκεια των πόρων που διαθέτουν. Εν τούτοις, το ερώτημα που ανακύπτει είναι: Αν η αιτία της κατάστασης αυτής είναι μόνο η έλλειψη πόρων ή και η μη αποτελεσματική διαχείριση των πόρων καθώς και άλλοι παράγοντες;

Πράγματι, ζητούμενο στη σχέση Διοίκησης και κοινωνίας των πολιτών είναι αν, εκτός των αναγκαίων πόρων για την αντιμετώπιση του, υπό συνθηκών αγοράς, λογικού κόστους των παραγομένων προς προσφορά υπηρεσιών, η ποιότητα και η αποτελεσματικότητα της διοικητικής δράσης αναφέρονται στη νομιμότητα του περιεχομένου πράξεων και ενεργειών, εν γένει, στην εντιμότητα (fairness) των ακολουθούμενων διαδικασιών και χρησιμοποιούμενων μεθόδων στην επιστημονική διοίκηση, αλλά και σε άλλους σημαντικούς ποιοτικούς παράγοντες που συνδέονται με την ορθή διαχείριση των πόρων. Επίσης, ιδιαίτερα απασχολεί τους ερευνητές το ότι η εφαρμογή των στατιστικών και των άλλων ποσοτικών μεθόδων και η αξιοπιστία των εμπειρικών αναλύσεων εξαρτώνται, σε μεγάλο βαθμό από την ποιότητα των διαθέσιμων στοιχείων, επίσημων ή προσφερομένων από ιδιωτικές τράπεζες δεδομένων κλπ., που συνήθως παρουσιάζουν σημαντικές ελλείψεις και άλλα προβλήματα. Τέλος, για τις έρευνες αξιολόγησης και μέτρησης της αποτελεσματικότητας δεν έχει συνταχθεί κώδικας διαφάνειας, αμεροληψίας και εν γένει δεοντολογίας, με αποτέλεσμα να ανακύπτουν ζητήματα διασφάλισης των αρχών αυτών στη συλλογή των στοιχείων, στην επεξεργασία τους, στη γενικότερη στατιστική ανάλυσή τους και στην ερμηνεία και γνωστοποίηση των αποτελεσμάτων – πορισμάτων και, όπως είναι φυσικό, να εγείρονται πολλά ηθικά, νομικά και πολιτικά ζητήματα στη διαδικασία των ερευνών και κυρίως κατά τη δημοσιοποίηση των αποτελεσμάτων που προκύπτουν.

2. ΔΙΟΙΚΗΤΙΚΗ ΑΠΟΤΕΛΕΣΜΑΤΙΚΟΤΗΤΑ

Διεθνώς υποστηρίζεται ότι: Στη σύγχρονη παγκοσμιοποιημένη και ανταγωνιστική κοινωνία της γνώσης και της πληροφορίας, για να επιτύχουν οι Δημόσιες Υπηρεσίες, οι Δημόσιοι Οργανισμοί και οι Δημόσιες Επιχειρήσεις να αποκτήσουν μια ευέλικτη, ανοικτή, συμμετοχική και αποτελεσματική μορφή οργάνωσης, πρέπει να αλλάξουν τον τρόπο που «σκέπτονται» και «ενεργούν». [Fox and Miller (1995), Kettl and Dijulio (1995), Kettl (2000)]. Ειδικότερα: Ορισμένοι διοικητικοί επιστήμονες προβάλλουν ως αιτία της «διοικητικής αναποτελεσματικότητας» το ότι η Διοίκηση διατηρεί σε μεγάλη έκταση τις «παραδοσιακές μεθόδους» και το ότι δεν χρησιμοποιεί επαρκώς την «επιστημονική γνώση», η οποία, ως γνωστό, έχει αναπτυχθεί διεθνώς στο πλαίσιο της κίνησης «της επιστημονικής διοίκησης», που ξεκίνησε τις πρώτες δεκαετίες του 20^{ου} αιώνα με τα «ταιηλοριανά υποδείγματα» και κατέληξε τα τελευταία 30 χρόνια στο «αγοραίο υπόδειγμα». Με άλλα λόγια, διατείνονται ότι για να αρθούν η ανομοιογένεια, η ασάφεια, η αβεβαιότητα, η πολυπλοκότητα και οι συγκρούσεις που χαρακτηρίζουν το αντικείμενο «Διοίκηση», και για να υπάρξει ουσιαστική βελτίωση, θα πρέπει η διοικητική πρακτική να βασίζεται μόνο σε συγκεκριμένη επιστημονική γνώση, με

κανόνες που να ισχύουν διεθνώς και μάλιστα τόσο στον ιδιωτικό όσο και στον Δημόσιο Τομέα. Στόχος δε της διαδικασίας αυτής είναι η αναζήτηση της καλύτερης μεθόδου για τη μεγιστοποίηση της παραγωγικότητας και την προσέγγιση του άριστου αποτελέσματος.

Όμως, η σχετική «με τις ανθρώπινες σχέσεις» κίνηση, μετά από σειρά ερευνών, θεμελίωσε την άποψη ότι μόνο το οικονομικό όφελος του ανθρώπου δεν μπορεί να ερμηνεύσει τη συνολική εργασιακή συμπεριφορά του και, άρα, την απόδοσή του. [Nichols (1997), Box (1999), De Leon and Denhardt (2000), Denhardt and Denhardt (2000)]. Επομένως, μια τεχνοκρατική αντίληψη ότι η εφαρμογή τυποποιημένων υποδειγμάτων, που προέρχονται, συνήθως, από την ξένη εμπειρία, οδηγεί στο άριστο για τη Διοίκηση αποτέλεσμα, δίχως να λαμβάνονται υπόψη οι βασικές αρχές δεοντολογίας και «ο ανθρώπινος παράγοντας», είναι εσφαλμένη, όχι μόνο για λόγους ηθικής, δεοντολογίας και δημοκρατικής συμπεριφοράς, αλλά και για λόγους σκοπιμότητας, αφού ένα υπόδειγμα, για να απεικονίζει την πραγματική κατάσταση, πρέπει να περιέχει τουλάχιστο τους σημαντικότερους ποιοτικούς και ποσοτικούς παράγοντες, που ερμηνεύουν το υπό εξέταση φαινόμενο.

3. ΕΥΡΥΤΕΡΗ ΕΠΙΣΤΗΜΟΝΙΚΗ ΓΝΩΣΗ

Οι διοικητικές αποφάσεις χαρακτηρίζονται από πολυπλοκότητα, αντιθέσεις ή συγκρούσεις, αβεβαιότητα, επειδή οι μελλοντικές συνθήκες καθορίζουν την αξία μιας απόφασης, και ασάφεια, λόγω του πλημμελούς καθορισμού στοιχείων καθώς και οριοθέτησης προβλημάτων. Άρα, η συστηματική εφαρμογή εμπειρικών αναλύσεων τυγχάνει όλο και μεγαλύτερης αποδοχής στη διεθνή πρακτική [Ackoff (1981), Jackson and Keys (1984), Checkland (1985), Richard et. al. (1986)], προκειμένου να μελετηθεί με διαφάνεια η πολυπλοκότητα, η ασάφεια και η αβεβαιότητα διοικητικών θεμάτων καθώς και η αποτελεσματικότητα της Διοίκησης, μέσα από κατάλληλα υποδείγματα. Τα υποδείγματα που χρησιμοποιούνται αναπαριστούν πραγματικά φαινόμενα και καταστάσεις, λαμβάνοντας υπόψη τα αναγκαία λειτουργικά δεδομένα που αφορούν τους παραγωγικούς συντελεστές και προϊόντα, τη διοικητική και γενικότερη συμπεριφορά του ανθρώπινου δυναμικού, την ποιότητα της Διοίκησης, τις θεσμικές, διοικητικές και κοινωνικοοικονομικές συνθήκες της χώρας και τις εξελίξεις που αφορούν το ευρωπαϊκό και διεθνές περιβάλλον κλπ.

Συνεπώς, η επίτευξη της βελτιστοποίησης της διαχείρισης των πόρων απαιτεί «ευρύτερη» επιστημονική γνώση, που αφορά: α. Την επιστημονική διοίκηση. β. Τις αρχές δεοντολογίας. γ. Τους θεσμούς. δ. Το ευρύτερο κοινωνικοοικονομικό, φυσικό και τεχνικό περιβάλλον. ε. Τις ποσοτικές μεθόδους και την τεχνολογία.

α. Επιστημονική Διοίκηση

Η επιστημονική διοίκηση βασίζεται: (1) Στις αρχές και μεθόδους οργάνωσης. (2) Στις αρχές και μεθόδους διοίκησης. (3) Στον προγραμματισμό. (4) Στη στοχοθεσία. (5) Στην αξιολόγηση και λήψη αποφάσεων. (6) Στις μεθόδους βελτίωσης της παραγωγικότητας.

β. Αρχές Δεοντολογίας

Τα τελευταία χρόνια αυξάνονται συνεχώς οι απαιτήσεις της κοινωνίας των πολιτών για διαφάνεια, αξιοκρατία, ίση μεταχείριση, κοινωνική δικαιοσύνη. Στον Ιδιωτικό Τομέα όροι, όπως εταιρική κοινωνική ευθύνη, εταιρική διακυβέρνηση, οργανωσιακή ηθική και εταιρική συνείδηση, χρησιμοποιούνται στην προσπάθεια να γίνουν οι επιχειρήσεις καλοί «εταιρικοί πολίτες». Στον Δημόσιο Τομέα, που έχει και την ευθύνη της παρακολούθησης των ιδιωτικών επιχειρήσεων, προβλήματα που αφορούν όχι μόνο την εφαρμογή των νόμων και τη διαύγεια της διαχείρισης, αλλά και ευρύτερα τη δημόσια ηθική και δεοντολογία, απασχολούν όλο και συχνότερα την πολιτεία και τη δικαιοσύνη. Το ζητούμενο, λοιπόν, στη σχέση «Διοίκηση και κοινωνία των πολιτών», πέραν της επιβαλλομένης από την συνταγματική αρχή του κράτους δικαίου νομιμότητας της διοικητικής δράσης, είναι και η τήρηση των αρχών που συνδέονται με την «Καλή Διοίκηση» [Μακρυδημήτρης (2003, 2010)]. Η έμφαση στη διοικητική ηθική και η προσήλωση στις βασικές αξίες και αρχές, που πρέπει να διέπουν τη Διοίκηση, προσθέτουν ποιότητα στη λειτουργία της, κυρίως, όσον αυτή αφορά την καθημερινότητα του πολίτη, αλλά και τη μεθόδευση επίτευξης των στόχων που έχει θέσει.

γ. Οι θεσμοί

Η Δημόσια Διοίκηση και κάθε διοικητικό όργανο, που δρα στο πλαίσιο της οφείλει να ενεργεί σύμφωνα με το νόμο ή να εναρμονίζεται με αυτόν. Τούτο σημαίνει ότι η Δημόσια Διοίκηση και τα όργανα που τη συγκροτούν μπορεί να αναλάβουν δράση και να αναπτύξουν πρωτοβουλίες μόνο εφόσον τούτο προβλέπεται ή επιτρέπεται από τις ισχύουσες σχετικές διατάξεις και τους κανόνες και με τον τρόπο που αυτό προβλέπεται. [Μακρυδημήτρης (2003)].

δ. Πληροφόρηση για το Ευρύτερο Περιβάλλον

Ιδιαίτερα σημαντικό και απαραίτητο είναι να υπάρχει γνώση για την αποτίμηση του οικονομικού, κοινωνικού, φυσικού και τεχνολογικού περιβάλλοντος. Στην ουσία αναφερόμαστε σε όλους τους παράγοντες που επηρεάζουν το εσωτερικό και εξωτερικό περιβάλλον και δημιουργούν έναν κύκλο αλληλεπιδράσεων. Το ίδιο συμβαίνει και με το διεθνές περιβάλλον, ως προς τις αλλαγές που πραγματοποιούνται σε αυτό, τα μηνύματα που εκπέμπει και τις επιδράσεις που ασκεί. [Παπούλιας (2002)].

ε. Ποσοτικές Μέθοδοι και Τεχνολογία

Οι στατιστικές και λοιπές ποσοτικές μέθοδοι προσφέρουν στις Διοικήσεις των Δημοσίων Υπηρεσιών, Οργανισμών και Επιχειρήσεων τη δυνατότητα να συγκεντρώσουν σημαντικές πληροφορίες για τη λήψη αποφάσεων, σχετικά με τα θέματα που αντιμετωπίζουν. Η συστηματική χρήση της ποσοτικής ανάλυσης, προκειμένου ν'αντιμετωπισθεί η πολυπλοκότητα, η αβεβαιότητα και η αστάθεια των διοικητικών θεμάτων τυγχάνει όλο και μεγαλύτερης αποδοχής στη διεθνή βιβλιογραφία [Ackoff (1981), Jackson and Keys (1984), Checkland (1985), Richard et. al. (1986), Papoulias and Darzentas (1990)]. Οι, υπό ευρεία έννοια, ποσοτικές

μέθοδοι, που εφαρμόζονται συνήθως στην πράξη, είναι οι εξής: Εφαρμοσμένα Μαθηματικά, Γραμμικός Προγραμματισμός, Προγραμματισμός Στόχων, Δυναμικός Προγραμματισμός, Θεωρία των Παιγνίων, Θεωρία Αποθεμάτων, Προσομοίωση, Θεωρία Αναμονής, Ευρετικός Προγραμματισμός, Στατιστική, Οικονομετρία, Δικτυωτή Ανάλυση κλπ. [Δονάτος (2002)]. Η σύγχρονη τεχνολογία της πληροφορικής και οι λειτουργικές δυνατότητες που παρέχουν τα «ολοκληρωμένα» Πληροφοριακά Συστήματα Διοίκησης επιτρέπουν την καλύτερη αξιοποίηση των ποσοτικών μεθόδων και τη λήψη ικανοποιητικών αποτελεσμάτων από τις σχετικές εμπειρικές αναλύσεις [Δονάτος (2002)].

4. ΟΛΟΚΛΗΡΩΜΕΝΟ ΠΛΗΡΟΦΟΡΙΑΚΟ ΣΥΣΤΗΜΑ ΔΙΟΙΚΗΣΗΣ

Η εφαρμογή των στατιστικών και λοιπών ποσοτικών μεθόδων και η αξιοπιστία της όλης εμπειρικής ανάλυσης εξαρτώνται σε μεγάλο βαθμό από τα διαθέσιμα στοιχεία (ποσοτικά και ποιοτικά), τα οποία παρέχονται συνήθως: (α) Από δημοσιεύσεις της Ελληνικής Στατιστικής Αρχής ή άλλων Δημοσίων ή Τοπικής Αυτοδιοίκησης Υπηρεσιών - Οργανισμών (β) Από Βάσεις Δεδομένων Ερευνητικών Κέντρων και Εταιρειών (γ) Από δειγματοληπτικές έρευνες ή μελέτες, που διεξήχθησαν για ορισμένους σκοπούς. Τα δημοσιευμένα επίσημα θεσμικά στοιχεία παρατηρείται ότι παρουσιάζουν σημαντικές ελλείψεις, όσον αφορά την ποιότητα, το χρόνο, τον τόπο και το είδος της πληροφορίας. Τα προβλήματα προσauxάζονται στην περίπτωση που τα λειτουργικά δεδομένα προέρχονται από ιδιωτικές Τράπεζες Δεδομένων ή άλλες πηγές. Όμως, η διεξαγωγή εμπειρικών αναλύσεων, προκειμένου να ληφθούν ικανοποιητικές πληροφορίες για την λήψη σημαντικών αποφάσεων, απαιτεί ακριβή, σαφή, πλήρη, επαρκή και αξιόπιστα στοιχεία, για τα οποία πρέπει να έχει εξασφαλισθεί συστηματικός-ομοιόμορφος τρόπος καταχώρησης, τακτική ενημέρωση, ποιοτική επεξεργασία και εύκολη προσπέλαση. Στην πράξη η έλλειψη ολοκληρωμένης πληροφόρησης αποτελεί σημαντικό εμπόδιο για την ανάπτυξη των αναγκαίων διαδικασιών για τη στήριξη μιας απόφασης, ενώ συνήθως, για το διατιθέμενο χρονικό διάστημα, η πραγματοποίηση της αναγκαίας δειγματοληπτικής έρευνας, που θα μπορούσε να συγκεντρώσει τα αναγκαία συμπληρωματικά λειτουργικά στοιχεία τα οποία θα ικανοποιούν τις στατιστικές απαιτήσεις, δεν είναι πάντοτε δυνατή. Είναι λοιπόν φανερό ότι η ύπαρξη μιας αυτοματοποιημένης Τράπεζας Δεδομένων, στο πλαίσιο ενός επίσημου, ανεξάρτητου, «ολοκληρωμένου» Πληροφοριακού Συστήματος Διοίκησης, που θα συνδέει τις Δημόσιες Υπηρεσίες και Επιχειρήσεις, και τους Δημόσιους Οργανισμούς της Χώρας, θα διασφαλίσει θέματα διαφάνειας και δεοντολογίας και θα επιτρέψει την τροφοδότηση των υποδειγμάτων με στατιστικά στοιχεία: ακριβή, διαθέσιμα, αξιόπιστα, επαρκή, συγκρίσιμα, μη επικαλυπτόμενα και συνεπή σε σχέση με το αντικείμενο, το χρόνο και τον τόπο. Παράλληλα, η σύγχρονη τεχνολογία της πληροφορικής και οι λειτουργικές δυνατότητες, που παρέχει ένα «ολοκληρωμένο» Πληροφοριακό Σύστημα Διοίκησης, μεγιστοποιούν τις δυνατότητες των σχετικών εμπειρικών υποδειγμάτων. [Δονάτος (2002)].

Έτσι, θα είναι δυνατό να λειτουργήσει ένα ευρύτερο Σύστημα Διοίκησης και Αποτίμησης της Αποτελεσματικότητας, για το οποίο θα μπορεί να εξασφαλισθεί η

αναγκαία πληροφόρηση για το σχεδιασμό και την εφαρμογή των στόχων, αλλά και για την αξιολόγηση της αποτελεσματικότητας του Δημοσίου Τομέα. Μάλιστα, η μέτρηση της αποτελεσματικότητας θα μπορεί να αποτελέσει μια συνεχή διαδικασία, η οποία θα οδηγεί στον επαναπροσδιορισμό των στόχων και στη βελτίωση της αποτελεσματικότητας του Δημοσίου Τομέα.

5. ΑΞΙΟΛΟΓΗΣΗ ΤΗΣ ΑΠΟΤΕΛΕΣΜΑΤΙΚΟΤΗΤΑΣ

5.1. Μελέτες Αξιολόγησης

Ο ελληνικός Δημόσιος Τομέας είναι ένα σύνθετο σύστημα, η δε αξιολόγηση της αποτελεσματικότητάς του εξαρτάται από την αποτελεσματικότητα των υποσυστημάτων που τον απαρτίζουν. Ειδικότερα, η αξιολόγηση της αποτελεσματικότητας του Δημοσίου Τομέα, επιτυγχάνεται με την πραγματοποίηση εμπειρικών μελετών που αποτιμούν τη σχετική αποτελεσματικότητα των μονάδων των επιμέρους φορέων, που τον συγκροτούν, καθώς και με δειγματοληπτικές έρευνες, που αποσκοπούν, κυρίως, στον έλεγχο και στην ερμηνεία της συμπεριφοράς και της παραγωγικότητας του ανθρώπινου δυναμικού του.

Για την εκτίμηση της σχετικής αποτελεσματικότητας των μονάδων των Δημοσίων Υπηρεσιών ή των Δημοσίων Οργανισμών κλπ. εφαρμόστηκαν, κυρίως, η Data Development Analysis (DEA) και Οικονομετρικές Μέθοδοι (π.χ. η Μέθοδος Ελαχίστων Τετραγώνων). Οι μέθοδοι αυτές υπερτερούν της απλής εφαρμογής των δεικτών, επειδή δεν περιορίζονται σε μία εισροή ή εκροή, αλλά λαμβάνουν υπόψη το σύνολο των παραγωγικών μέσων και μάλιστα την αλληλεπίδραση των εισροών – εκροών των μονάδων. Ακόμη, σενάρια προσομοιώσεων παρέχουν τη δυνατότητα διερεύνησης διαφόρων καταστάσεων, για τις εμφανιζόμενες ιδιαιτερότητες των μονάδων. Στην Ελλάδα εμπειρικές μελέτες εκτίμησης της σχετικής αποτελεσματικότητας έχουν εφαρμοσθεί σε Νοσοκομεία, Τηλεπικοινωνίες, Λιμάνια, Περιφέρειες – Νομαρχίες, Αεροδρόμια, Ξενοδοχεία, Μεταφορές, Αγορά Ακινήτων, Ασφαλιστική Αγορά, Δημόσιους Οργανισμούς, Πανεπιστήμια, Τράπεζες κλπ. [Miliotis (1922), Athanassopoulos (1995), Athanassopoulos and Giokas (1998), Giokas (2001), Δονάτος και Γκιώκας (2002), Δονάτος κ.ά. (2002), Donatos and Giokas (2008), Giokas (2008), Halkos and Tzeremes (2010), Karagianis and Valentzas (2010), Kritikos et. al. (2010), Michaelidis et. al. (2010), Tsekeris (2011)].

5.2. Εκτίμηση της Σχετικής Αποτελεσματικότητας των Ελληνικών Νοσοκομείων

Επειδή δεν είναι δυνατό να περιγραφεί το σύνολο των εμπειρικών μελετών, αναφερόμαστε, ενδεικτικά και συνοπτικά, στη μελέτη: «Εκτίμηση της Σχετικής Παραγωγικότητας των Ελληνικών Νοσοκομείων» [Δονάτος και Γκιώκας (1998)]. Στη μελέτη εφαρμόστηκαν η DEA και η Μέθοδος Ελαχίστων Τετραγώνων και χρησιμοποιήθηκαν εναλλακτικά σενάρια στο συνολικό κόστος, βρέθηκε ο βαθμός της σχετικής αποτελεσματικότητας των μονάδων και υπολογίστηκε η παραγωγή που θα μπορούσε να πραγματοποιηθεί από κάθε μονάδα, αν χρησιμοποιούσε τους πόρους της με τον σχετικά αποτελεσματικότερο τρόπο. Ειδικά, με την DEA υπολογίστηκαν, μεταξύ των άλλων, και οι βελτιώσεις που πρέπει να πραγματοποιήσουν οι μη

αποτελεσματικές μονάδες για να γίνουν αποτελεσματικές, καθώς και οι οικονομίες κλίμακας των μονάδων.

Τα κύρια συμπεράσματα τα οποία εξάγονται από τη μελέτη είναι τα εξής: (1) Από τα 96 νοσοκομεία, που εξετάστηκαν, 25 παρουσίασαν το μέγιστο βαθμό σχετικής αποτελεσματικότητας (100%), 14 μικρότερο από 70%, 35 μεταξύ 70% και 90% και 22 μεταξύ 90 και 99%. (2) Τα μη αποτελεσματικά νοσοκομεία, για να χαρακτηριστούν αποτελεσματικά, θα πρέπει να μειώσουν το συνολικό κόστος, τουλάχιστο κατά μέσο όρο 17%. Το 61%, περίπου, της μείωσης αφορά τη μισθοδοσία του προσωπικού, το 4% τα λειτουργικά έξοδα, το 14% τις προμήθειες φαρμάκων και το υπόλοιπο 21% τις λοιπές προμήθειες. (3) Τα μεσαίου μεγέθους νοσοκομεία (ανάλογα με τον αριθμό των κλινών) παρουσιάζουν υψηλότερα επίπεδα αποτελεσματικότητας. (4) Τα μεγάλα μεγέθους νοσοκομεία λειτουργούν κάτω από φθίνουσες οικονομίες κλίμακας. Αντίθετα, τα μικρότερα νοσοκομεία τείνουν να λειτουργούν στη ζώνη των αυξουσών αποδόσεων κλίμακας. (5) Η υπερβολική κατανάλωση πόρων, από τα μη αποτελεσματικά νοσοκομεία, οδηγεί στο συμπέρασμα ότι τουλάχιστο το 4.1% των δαπανών υγείας του ΑΕΠ οφείλεται στην αναποτελεσματική χρησιμοποίηση των πόρων των νοσοκομείων.

5.3. Δειγματοληπτική Έρευνα

i. Ικανότητες και χαρακτηριστικά στελεχών

Μεγάλη σημασία για την ερμηνεία και τη συμπεριφορά της Δημόσιας Διοίκησης έχουν το δίκαιο, η δημόσια ηθική και δεοντολογία, το πολιτικό και πολιτιστικό περιβάλλον, εντός του οποίου δρα, το κοινωνικό, οικονομικό και αξιακό πλαίσιο αναφοράς της, καθώς οι άνθρωποι που συμμετέχουν στη λήψη και υλοποίηση των αποφάσεων. Το ανθρώπινο κεφάλαιο αποτελεί σπάνιο στρατηγικό πόρο, η δε αποτελεσματικότητα της Δημόσιας Διοίκησης εξαρτάται σε μεγάλο βαθμό από τις ικανότητες και τα χαρακτηριστικά των ηγετικών στελεχών και των άλλων ανθρώπων, που προσπαθούν ως άτομα, αλλά και ως ομάδες, να επιτύχουν τις μέγιστες δυνατές επιδόσεις. [Μπουραντάς (2005)]. Ειδικότερα, τα στοιχεία εκείνα, δηλαδή οι ικανότητες και τα χαρακτηριστικά προσωπικότητας, που πρέπει να εξετασθεί αν διαθέτει το ηγετικό στέλεχος, είναι τα εξής: Συστημική σκέψη, γνώσεις και εμπειρία για την αποτελεσματική αντιμετώπιση των στόχων που επιλέγονται, γενικότερη παιδεία, προσήλωση στις αληθινές αξίες και αρχές, εντιμότητα, υπευθυνότητα, θέληση, επιμονή, εργατικότητα, αυτοπεποίθηση, αυτοεπίγνωση, αυτορρύθμιση, διορατικότητα, έμπνευση οράματος, δημιουργικότητα, ενδιαφέρον για καινοτομίες και πρωτοβουλίες, αποδοχή νέων ιδεών και κριτικής, πνεύμα συνεργασίας και ενεργοποίηση συνεργατών. [Μπουραντάς (2005)].

ii. Ερωτηματολόγιο

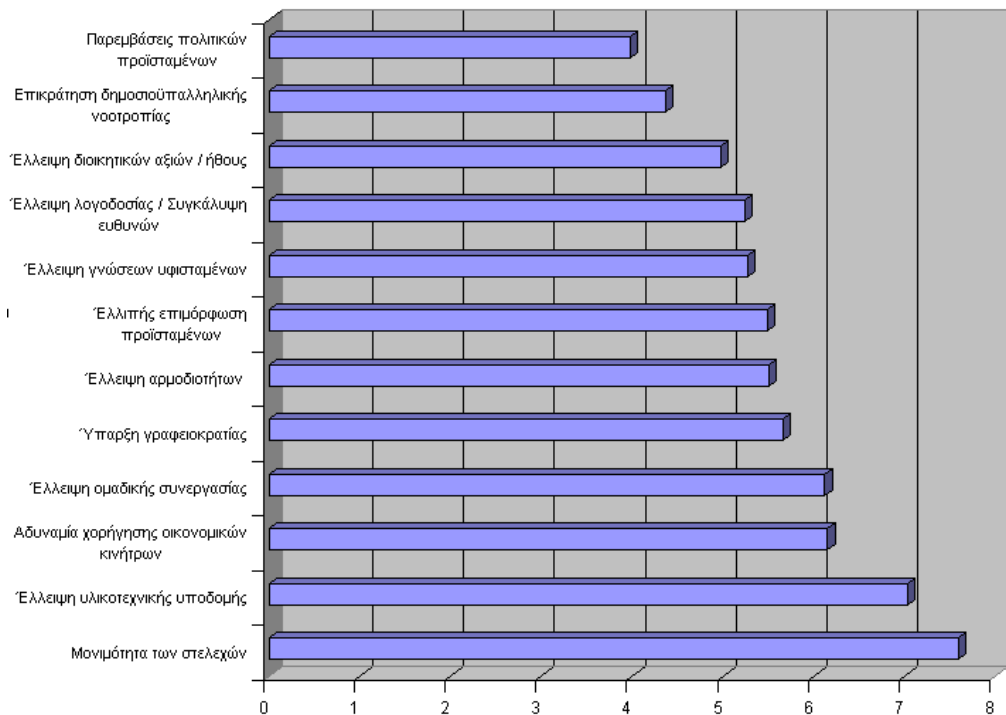
Επειδή η αποτίμηση των προαναφερομένων ποιοτικών στοιχείων θεωρήθηκε αναγκαία για να εξαχθούν χρήσιμα συμπεράσματα για την οργάνωση και τον τρόπο λειτουργίας του Δημοσίου Τομέα, πραγματοποιήσαμε το χρονικό διάστημα 2008 – 2011 δειγματοληπτική έρευνα χρηματοδοτούμενη από την Επιτροπή Ερευνών του Πανεπιστημίου Αθηνών. Συντάξαμε κατάλληλο Ερωτηματολόγιο, το οποίο

συμπληρώθηκε από 665 στελέχη και υπαλλήλους, που υπηρετούν στον Δημόσιο Τομέα και επελέγησαν τυχαία, για αξιοπιστία που επιτυγχάνεται με πιθανότητα 95%. Το ερωτηματολόγιο περιλαμβάνει σημαντικό αριθμό ερωτήσεων που αφορούν: Οργανωσιακά χαρακτηριστικά, τήρηση των αρχών της δεοντολογίας, συμπεριφορά των διοικητικών – ηγετικών στελεχών και λοιπών εργαζομένων καθώς και αξιολόγηση διαχείρισης πόρων, ποιότητας διαδικασιών, επίτευξης στόχων κλπ.

Τα στατιστικά στοιχεία, που έδωσε η δειγματοληπτική έρευνα, ταξινομήθηκαν στις εξής κατηγορίες: (1) Ποιότητα και αποτελεσματικότητα της Διοίκησης. (2) Διαφάνεια, ευελιξία και απλούστευση των διαδικασιών στη δράση της Διοίκησης. (3) Ορθή χρησιμοποίηση νέας τεχνολογίας / καινοτομίας. (4) Αποδοχή νέων ιδεών και κριτικής. (5) Συνεργασία – ομαδική συμπεριφορά στην εργασία. (6) Οργάνωση και συντονισμός στη δράση της Διοίκησης. (7) Παράγοντες που επηρεάζουν το έργο του προϊσταμένου. (8) Αξιοκρατία. (9) Κίνητρα βελτίωσης της αποτελεσματικότητας. (10) Στάση και συμπεριφορά της Διοίκησης έναντι του πολίτη. (11) Αρνητικοί παράγοντες που επηρεάζουν την απόδοση του διοικητικού στελέχους.

Ενδεικτικά, παραθέτουμε το Σχήμα 1, στο οποίο απεικονίζονται, σύμφωνα με τις απαντήσεις που δόθηκαν και σε κλίμακα βαθμολογίας, οι αρνητικοί παράγοντες που επηρεάζουν την απόδοση του διοικητικού στελέχους [κατηγορία (11)].

Σχήμα 1: Αρνητικοί Παράγοντες στην Απόδοση του Διοικητικού Στελέχους



Σημείωση: Κλίμακα 1 – 12, όπου: (1) ο λιγότερο αρνητικός παράγοντας και (12) ο πλέον αρνητικός παράγοντας.

iii. Δείγμα

Για να εκτιμηθεί η αναλογία $P = A/N$ των A μελών ενός πεπερασμένου πληθυσμού μεγέθους N , που χαρακτηρίζεται από μια συγκεκριμένη ιδιότητα (δίνουν συγκεκριμένη απάντηση), επιλέγεται ένα απλό (με ίσες πιθανότητες) τυχαίο δείγμα n χωρίς επανάθεση, στο οποίο εντοπίζονται a μέλη με αυτή την ιδιότητα.

Η εκτιμήτρια:

$$p = \frac{a}{n} \quad (1)$$

είναι αμερόληπτη εκτιμήτρια της αναλογίας P , με διακύμανση:

$$V(p) = \frac{NPQ}{n(N-1)} \left(1 - \frac{n}{N}\right) = \frac{PQ}{n} \left(\frac{N-n}{N-1}\right), \quad (2)$$

όπου για μεγάλα N χρησιμοποιείται στην πράξη η μορφή:

$$V(p) \approx \frac{PQ}{n} \left(1 - \frac{n}{N}\right), \quad (3)$$

ενώ αμερόληπτη εκτιμήτρια της $V(p)$ είναι η εκτιμήτρια:

$$\hat{V}(p) = \frac{pq}{n-1} \left(1 - \frac{n}{N}\right), \text{ όπου } q = 1 - p. \text{ [Cochran (1977)]} \quad (4)$$

Σημειώνεται ότι αν το μέγεθος N του πληθυσμού είναι άγνωστο ή αν $n < 0.05N$, όπως συμβαίνει στην έρευνα που πραγματοποιήθηκε, η «διόρθωση πεπερασμένου πληθυσμού» $1 - n/N$ παραλείπεται [Cochran (1977)].

Η ακρίβεια των εκτιμήσεων της έρευνας προσδιορίστηκε με βάση την υπόθεση ότι η εκτιμήτρια p κάθε εκτιμώμενης αναλογίας P του πληθυσμού είναι κανονική. Η υπόθεση της κανονικότητας πληρούται για εκτιμήτριες των αναλογιών P που έχουν τιμές από 0.1 μέχρι 0.9, αν το δείγμα είναι μεγαλύτερο του 600 [Yamane (1967)]. Συνεπώς, η υπόθεση αυτή πληρούται στη συγκεκριμένη έρευνα, αφού έχουμε δείγμα μεγέθους 665 ατόμων.

Με την υπόθεση της κανονικότητας, το αναγκαίο μέγεθος n_0 του δείγματος που αναμένεται, με πιθανότητα $1 - \alpha$, να δώσει εκτίμηση \tilde{p} της αναλογίας P , με προκαθορισμένη μέγιστη απόκλιση $d_0 = |P - \tilde{p}|$, προσδιορίζεται για μεγάλο N από τη σχέση [Cochran (1977)]:

$$n_0 = \frac{z_{\alpha/2}^2 PQ}{d_0^2}, \quad (5)$$

όπου η P είναι κατά προσέγγιση γνωστή από προηγούμενη έρευνα ή από πιλοτική έρευνα. Αν, όπως στη συγκεκριμένη έρευνα, η (κάθε) P είναι άγνωστη, θεωρούμε ως τιμή της την 0.5 για την οποία μεγιστοποιείται το n_0 , που θα δώσει προκαθορισμένη απόκλιση d_0 .

Στον Πίνακα 1 δίνονται τα αναγκαία μεγέθη n_0 (για $P=0.5$), ώστε να επιτευχθεί εκτίμηση \tilde{p} με προκαθορισμένες αποκλίσεις $d_0 = |P - \tilde{p}|$, για πιθανότητες $1 - \alpha$ ίσες προς 0.99, 0.95 και 0.90.

Πίνακας 1

1-α	d_0									$z_{\alpha/2}$
	0,020	0,025	0,030	0,035	0,037	0,038	0,040	0,047	0,050	
0,99	4147	2654	1843	1354	1212	1149	1037	751	663	2,58
0,95	2401	1537	1067	784	702	665	600	435	384	1,96
0,90	1691	1082	752	552	494	468	423	306	271	1,64

Με βάση τα στοιχεία του Πίνακα 1, θεωρήσαμε ως ικανοποιητική, από άποψη ακριβείας, την απόκλιση $d_0 = 0.038$, η οποία επιτυγχάνεται με πιθανότητα 0.95, όταν το δείγμα έχει μέγεθος 665, δηλαδή ακριβώς ίσο με το δείγμα των 665 ατόμων, που συμπλήρωσαν ορθά το ερωτηματολόγιο.

Προφανώς, οι υπό εκτίμηση αναλογίες P δεν είναι όλες ίσες με 0.5. Παρουσιάζει, λοιπόν, ενδιαφέρον να εκτιμηθούν με πιθανότητα 0.95 οι αποκλίσεις $d_i = |P_i - \tilde{p}_i|$, σε περιπτώσεις που οι εκτιμήσεις \tilde{p}_i είναι διαφορετικές από 0.5. Οι τιμών των d_i δίνονται στον Πίνακα 2, για επιλεγμένες τιμές των \tilde{p}_i , σύμφωνα με τη σχέση:

$$d_i = 1.96 \sqrt{\hat{V}(\tilde{p}_i)} = 1.96 \sqrt{\frac{\tilde{p}_i \tilde{q}_i}{n_i - 1}}, \text{ όπου } \tilde{q}_i = 1 - \tilde{p}_i :$$

Πίνακας 2

\tilde{p}_i	0,5	0,4 ή 0,6	0,3 ή 0,7	0,2 ή 0,8	0,1 ή 0,9
d_i	0,0380	0,0372	0,0348	0,0304	0,0228

Σημειώνεται, ακόμη, ότι μια εκτίμηση \tilde{p} θεωρείται στην πράξη ικανοποιητική, από άποψη ακριβείας, όταν ο συντελεστής μεταβλητότητας της $CV(\tilde{p}) = \sqrt{\hat{V}(\tilde{p})}/\tilde{p}$ είναι μικρότερος ή ίσος της τιμής 0.05. Στον Πίνακα 3 δίνονται τα ελάχιστα αναγκαία n_0 των δειγμάτων για να επιτευχθεί συγκεκριμένη εκτίμηση \tilde{p} , με $CV(\tilde{p}) \leq 0.05$, σύμφωνα με τη σχέση:

$$n_0 \geq 1 + \frac{\tilde{q}}{0.0025 \tilde{p}},$$

που συνάγεται από τη σχέση:

$$CV^2(\tilde{p}) = \frac{\tilde{p}\tilde{q}}{\tilde{p}^2} = \frac{n_0 - 1}{\tilde{p}^2} \leq 0.05^2 \text{ [Kish (1978)].}$$

Πίνακας 3

\tilde{p}	0,1	0,2	0,3	0,35	0,38	0,4	0,5	0,6	0,7	0,8	0,9
n_0	3601	1601	934	744	654	601	401	268	172	101	45

Συνεπώς, οι εκτιμήσεις \tilde{p} της αναλογίας P που έχουν τιμή μεγαλύτερη της τιμής 0.38, για κάθε δείγμα μεγέθους $n_0 = 654$ (<665), θεωρούνται ικανοποιητικές από άποψη ακριβείας, αφού έχουν συντελεστή μεταβλητότητας μικρότερο ή ίσο του 0.05, εφ' όσον αναφέρονται σε απαντήσεις από ολόκληρο το δείγμα και όχι από εξειδικευμένες κατηγορίες του δείγματος.

iv. Συμπερασματικές Παρατηρήσεις

Από την ανάλυση των στοιχείων, που προέκυψαν από τις απαντήσεις στο ερωτηματολόγιο, συνάγεται ότι για μια ποιοτική και αποτελεσματική Δημόσια Διοίκηση, που θα σέβεται τις αξίες και θα τηρεί τις αρχές της δεοντολογίας, χρειάζεται: (1) Βελτίωση της οργάνωσης και της ποιότητας. (2) Αξιοκρατία και διαφάνεια. (3) Στοχοθεσία και έλεγχο πραγματοποίησής της. (4) Συνεργασία και ομαδική συμμετοχή για την επίτευξη των στόχων. (5) Ενσωμάτωση της νέας τεχνολογίας και στήριξη της καινοτομίας. (6) Αποδοχή νέων ιδεών και κριτικής. (7) Υιοθέτηση ηθικών και οικονομικών κινήτρων για τη βελτίωση της αποτελεσματικότητας. (8) Ανεξαρτησία της διοίκησης από πολιτικούς παράγοντες. (9) Έρευνα για τις ανάγκες των πολιτών.

5.4. Ζητήματα και Ερωτήματα

Από τις μελέτες και τη δειγματοληπτική έρευνα, που πραγματοποιήσαμε ανακλύπτουν τα παρακάτω ζητήματα και ερωτήματα, σχετικά με τον ελληνικό Δημόσιο Τομέα :

α. Ζητήματα νομικά, πολιτικά και διασφάλισης των αρχών δεοντολογίας και αξιοπιστίας: i. Στην επιλογή, συλλογή, προμήθεια, κωδικοποίηση, πληρότητα, συνέπεια, επαλήθευση, βελτίωση των στοιχείων. ii. Στην επεξεργασία των στοιχείων. iii. Στη στατιστική μεθοδολογία που ακολουθείται. iv. Στην ανάλυση, ερμηνεία, προσπέλαση, δημοσιοποίηση και αξιοποίηση των αποτελεσμάτων.

β. Διασφαλίζεται από πολιτικές ή άλλες σκοπιμότητες, η Ελληνική Στατιστική Αρχή, οι Δημόσιες Υπηρεσίες και Επιχειρήσεις και οι Δημόσιοι Οργανισμοί και τα Κέντρα Ερευνών και Μελετών, ως προς τα θέματα: i. Αντικειμενικότητα, αξιοπιστία και ανεξαρτησία; ii. Επιστημονική και διαφανής λειτουργία; iii. Αμεροληψία και επάρκεια των δεδομένων; iv. Ολοκληρωμένη πληροφόρηση των αποτελεσμάτων, έναντι μερικής, ελλειπούς, μονομερούς και μεροληπτικής γνωστοποίησης – δημοσιοποίησης των; v. Διαθεσιμότητα χρησίων στοιχείων και ορθή διαχείριση απορρήτων πληροφοριών – αποτελεσμάτων;

γ. Υπάρχει επίσημο, σύγχρονο, αυτοματοποιημένο, ολοκληρωμένο, μη αμφισβητήσιμο Πληροφοριακό Σύστημα Διοίκησης που σε συνδυασμό με την πάταξη της γραφειοκρατίας και την κατάλληλη, αξιόπιστη δημοσιονομική και γενικότερα οικονομική πολιτική θα διευκολύνει την επενδυτική πολιτική;

δ. Η διαχείριση των πόρων βασίζεται:

i. Στη γνώση: (1) Των στοιχείων που εκφράζουν την πραγματική κατάσταση (ολοκληρωμένο Πληροφοριακό Σύστημα Διοίκησης); (2) Του γενικού περιβάλλοντος (θεσμοί, οικονομία, κοινωνία, τεχνολογία); (3) Του διεθνούς περιβάλλοντος; ii. Στις αρχές της συστημικής και συστηματικής Διοίκησης; iii. Στη διαφάνεια και στις αρχές της δεοντολογίας;

ε. Υπάρχει συνολική και υπεύθυνη αντιμετώπιση των αναγκών και στόχων του Δημοσίου Τομέα που να προβλέπει, αντί να αντιμετωπίζει πυροσβεστικά και ευκαιριακά τα διάφορα θέματα και τα παρουσιαζόμενα εκάστοτε προβλήματα;

Ιδιαίτερα: Γίνεται συντονισμένη Έρευνα – Μελέτη για όλους τους κλάδους του Δημοσίου Τομέα και όχι αποσπασματικά, που να οδηγεί: i. Σε ιεράρχηση των αναγκών; ii. Σε αξιολόγηση των αποτελεσμάτων που επιτυγχάνονται; iii. Σε στοχοθεσία, με συγκεκριμένους αποδέκτες εκτέλεσης; iv. Σε καθοριστικές συνολικές προτάσεις – αποφάσεις; v. Σε ανάδειξη τεκμηριωμένης πρότασης επενδυτικών ευκαιριών;

στ. Μήπως ξοδεύονται τεράστια ποσά για πλήθος μελετών, που πολλές φορές: i. Δεν αντιμετωπίζουν ιεραρχικά τις κύριες και ουσιαστικές ανάγκες; ii. Αντιμετωπίζουν αποσπασματικές καταστάσεις; iii. Επικαλύπτονται μεταξύ τους; iv. Δεν οδηγούν στο ζητούμενο; v. Δεν αναδεικνύουν τεκμηριωμένες προτάσεις επενδυτικών ευκαιριών;

ζ. Αξιοποιούνται, εφαρμόζονται τα αποτελέσματα και οι προτάσεις σημαντικών μελετών και δειγματοληπτικών ερευνών που έχουν πραγματοποιηθεί για τους Δημόσιους Φορείς, πολλές φορές μάλιστα κατά ανάθεση από αυτούς;

η. Όταν ζητείται η συναίνεση των κοινωνικών εταίρων στη διαβούλευση: i. Υπήρξε πριν την επιλογή των στρατηγικών, που θα εφαρμοστούν, η κατάλληλη τεκμηρίωση

και αξιολόγηση των καταστάσεων; ii. Υπήρξε διαβούλευση, σε τακτά διαστήματα κατά την αναπτυξιακή διαδικασία, πριν από το στάδιο των τελικών αποφάσεων, για νέες ρυθμίσεις, που πιθανώς να οδηγούν: (1) Σε κρίσιμες καταστάσεις; (2) Σε αναπόφευκτες συγκρούσεις;

6. ΕΠΙΛΟΓΟΣ

Στον διεθνή χώρο ο σχεδιασμός και τα προγράμματα διοικητικού εκσυγχρονισμού στοχεύουν στην οικοδόμηση διοικητικών συστημάτων που θα ικανοποιούν τον πολίτη, παρέχοντας υπηρεσίες υψηλής ποιότητας και αποτελεσματική διαχείριση [Kettl and Dijulio (1995), Kamensky (1966), Nichols (1997), Box (1999), De Leon and Denhardt (2000), Berg (2001)]. Αλλά, ενώ, κρίνεται αναγκαίος ο εκσυγχρονισμός του Δημοσίου Τομέα, η θέσπιση συγκεκριμένων κοινών κανόνων για την αποτίμηση της αποτελεσματικότητάς του, διαδικασία που πρέπει να προηγηθεί, δεν είναι εύκολη υπόθεση. Τούτο συμβαίνει, κυρίως, διότι οι διοικητικές δομές και η θεσμική και διοικητική οργάνωση καθώς και ο τρόπος λειτουργίας των διοικητικών μονάδων, διαφέρουν από χώρα σε χώρα, αλλά και στην ίδια χώρα. Επίσης, η έννοια της αποτελεσματικότητας του Δημοσίου Τομέα δεν ορίζεται μονότροπα και αναφέρεται σε ένα σύνθετο σύστημα, η δε μέτρησή της εξαρτάται από τη δυναμική και τη συνέργεια των υποσυστημάτων που τον απαρτίζουν.

Εξάλλου, τα στοιχεία που θα χρησιμοποιηθούν, στην αξιολόγηση των μονάδων μιας Δημοσίας Υπηρεσίας ή Επιχείρησης ή ενός Δημοσίου Οργανισμού κλπ., θα πρέπει να χαρακτηρίζονται από ακρίβεια, πληρότητα, επάρκεια, ομοιογένεια, συνέπεια και αξιοπιστία. Επίσης, η πληροφόρηση που θα προκύψει πρέπει να είναι ικανοποιητική ποιοτικά, μη αμφισβητήσιμη, ολοκληρωμένη, σαφής, αντικειμενική, προσπελάσιμη και αξιοποιήσιμη στατιστικά. Για να ισχύουν οι συνθήκες αυτές θα πρέπει να ικανοποιούνται οι αναγκαίες προϋποθέσεις που επιβάλλει η στατιστική επιστήμη και να τηρούνται η διαφάνεια και οι άλλες βασικές αρχές δεοντολογίας στις επιμέρους διαδικασίες, που συνδέονται με τη συγκέντρωση, κατάταξη, επεξεργασία και ανάλυση των στοιχείων, καθώς και σε όλα τα στάδια αποτίμησης της σχετικής αποτελεσματικότητας των μονάδων. Μάλιστα, η δημιουργία και λειτουργία μιας συνεχώς ενημερωμένης αυτοματοποιημένης Τράπεζας Δεδομένων, που θα αποτελεί τη βάση ενός επίσημου και ανεξάρτητου Πληροφοριακού Συστήματος Διοίκησης, θα εξασφαλίσει τη διάθεση ικανοποιητικών στατιστικών στοιχείων, θα βοηθήσει στην ανάπτυξη και εφαρμογή κατάλληλων υποδειγμάτων και θα συντελέσει στην τεκμηριωμένη και συνεχή διαδικασία αξιολόγησης της αποτελεσματικότητας του Δημοσίου Τομέα, ώστε να επιτευχθεί η αναγκαία πληροφόρηση για τη λήψη των σχετικών αποφάσεων για τον εκσυγχρονισμό και την ανάπτυξη.

Ειδικότερα, η στατιστική επιστήμη συμβάλλει στην ανάπτυξη της «Καλής και Επιστημονικής Διοίκησης» και στον επιτυχή «εκσυγχρονισμό» του Δημοσίου Τομέα, αλλά και στο σχεδιασμό, στην επιλογή, στην υλοποίηση, στην επανεξέταση και στη βελτίωση των αναπτυξιακών και λοιπών πολιτικών, διασφαλίζοντας:

α. Την ακρίβεια, αντικειμενικότητα, πληρότητα, συνέπεια, ομοιογένεια και αξιοπιστία των δεδομένων.

β. Τη διαφάνεια και την τήρηση βασικών αρχών δεοντολογίας στην επιλογή, συλλογή, επεξεργασία και ανάλυση των στοιχείων και στην ερμηνεία, διάθεση, δημοσιοποίηση και αξιοποίηση των αποτελεσμάτων των εμπειρικών ερευνών.

γ. Τη δημιουργία σύγχρονου, αυτοματοποιημένου, ολοκληρωμένου, ανεξάρτητου Πληροφοριακού Συστήματος Διοίκησης.

δ. Τις απαραίτητες συνθήκες για την κατασκευή, ανάπτυξη και λειτουργία των υποδειγμάτων.

ε. Τη διάθεση των αναγκαίων και αξιόπιστων ποσοτικών και ποιοτικών στοιχείων για τον έλεγχο και την ερμηνεία της συμπεριφοράς του ανθρώπινου δυναμικού και της Διοίκησης καθώς και για την τεκμηριωμένη και αξιόπιστη αξιολόγηση της αποτελεσματικότητας των επί μέρους μονάδων, με βάση τους κανόνες της επιστήμης και της δεοντολογίας.

στ. Την επαναξιολόγηση της απόδοσης των επιμέρους μονάδων σε προκαθορισμένα χρονικά διαστήματα και κυρίως μετά την εφαρμογή σημαντικών μεταβολών στη δομή και στον τρόπο της Διοίκησης.

ζ. Ευνοϊκό κλίμα για συναίνεση των κοινωνικών εταίρων στη διαβούλευση για αναγκαίες μεταρρυθμίσεις, λόγω της εμπειριστατωμένης και μη αμφισβητήσιμης πληροφόρησης που παρέχεται.

ABSTRACT

The evaluation of effectiveness of the Greek Public Sector is a compulsory process before each reform. The plenitude as well as the quality of statistical data, and the surveillance of transparency and ethics in general, at all stages of their statistical analysis, are considered as compulsory requirements for the without doubt assessment of effectiveness. The present article proposes the construction of a Data Bank, which should be continuously and automatically updated. This Data Bank should constitute the basis for an official and independent Information System Management. In addition, the study attempts to raise issues, questions and ascertainments that come from sampling research and empirical studies conducted on behalf of Greek Public Sector. The manuscript provides evidence that Statistical Science ensures: (a) the surveillance of transparency and ethics at all stages of data analysis, including inference, interpretation and utilization of empirical results (b) the availability of suitable quantitative and qualitative data. Also, Statistical Science contributes substantially to: (i) the control and interpretation of behavior and productivity of Human Resources and Administration, (ii) the documented and reliable evaluation of the effectiveness of the Public Sector, according to the rules of science and deontology (iii) the optimal decision-making for the modernization of Public Sector and the application of developmental policies.

ΑΝΑΦΟΡΕΣ

Ackoff, R.L. (1981). *Creating the Corporate Future*, New York: John Wiley.

- Athanassopoulos, A. (1995). Goal programming and data envelopment analysis (GoDEA) for target-based multi-level planning: Allocating central grants to the Greek local authorities. *European Journal of Operational Research*, **87**, 535-550.
- Athanassopoulos, A. and Giokas, D. (1998). Technical Efficiency and Economies of Scale in State Owned Enterprises: The Hellenic Telecommunications Organisation. *European Journal of Operational Research*, **107**, 62-75.
- Beer, M. (2001). How to develop an organization capable of sustained high performance. *Organisational Dynamics*, **29**, 233-247.
- Box, R. (1999). Running government like a business. *American Review of Public Administration*, **29**, 19-43.
- Checkland, P. (1985). From optimizing to learning: A development of systems thinking for the 1990's. *Journal of the Operational Research Society*, **36**, 757-768.
- Cochran, W.G. (1977). *Sampling Techniques*, New York: John Wiley.
- DeLeon, L. and Denhardt, R. (2000). The political theory of reinvention. *Public Administration Review*, **60**, 89-97.
- Denhardt, R. and Denhardt, J. (2000). The new public service: serving rather than steering. *Public Administration Review*, **60**, 549-559.
- Δονάτος, Γ. και Γκιώκας Δ. (1998). Εκτίμηση της σχετικής παραγωγικότητας των ελληνικών νοσοκομείων. *Πρακτικά 11ου Πανελληνίου Συνεδρίου Στατιστικής*, Χανιά, 77-86.
- Δονάτος, Γ. (2002). Οι ποσοτικές μέθοδοι στη διοικητική εκπαίδευση. Τμητικός Τόμος για την Ομότιμη Καθηγήτρια Λίτσα Νικολάου – Σμοκοβίτη. Πανεπιστήμιο Πειραιώς, 327-337.
- Δονάτος, Γ. και Γκιώκας, Δ. (2002). Η εκτίμηση της αποτελεσματικότητας σε δημόσιους οργανισμούς: Η περίπτωση του ΟΤΕ. *Πρακτικά 15ου Πανελληνίου Συνεδρίου Στατιστικής*, Ιωάννινα, 170-178.
- Δονάτος, Γ., Γκιώκας, Δ. και Αθανασόπουλος, Α. (2002). Εναλλακτικά υποδείγματα εισροών – εκροών για την εκτίμηση της σχετικής παραγωγικότητας του δικτύου των καταστημάτων ελληνικής τράπεζας. *Σπουδαί*, **52**, 136-159.
- Donatos, G. and Giokas, D. (2008). Relative efficiency in the branch network of a Greek bank: A quantitative analysis. *European Research Studies*, **XI**, 53-72.
- Fox, C. and Miller, H. (1995). *Postmodern Public Administration*, CA: Sage Publications.
- Giokas, D. (2001). Greek Hospitals: How well their resources are used. *Omega the International Journal of Management Science*, **29**, 73-83.
- Giokas, D. (2008). Assessing the efficiency in operations of a large Greek bank branch network adopting different economic behaviours. *Economic Modeling*, **25**, 559-574.
- Halkos, G. and Tzeremes, N. (2010). Measuring regional economic efficiency: The case of Greek prefectures. *Annals of Regional Studies*, **45**, 603-632.
- Jackson, N.K. and Keys, P. (1984). Towards a system of systems Methodologies. *The Journal of the Operational Research Society*, **35**, 473-486.
- Kamensky, J. (1996). The role of reinventing government movement in federal management reform. *Public Administration Review*, **56**, 247-255.

- Karagiannis, B. and Velentzas, K. (2010). Productivity and quality changes in Greek public hospitals. *Journal of Operational Research*, 1-10, in press.
- Kettl, D. (2000). The transformation of governance: globalisation, devolution and the role of government. *Public Administration Review*, **60**, 488-497.
- Kettl, D. and Dijulio, J. (1995). *Inside the Reinvention Machine: Appraising Governmental Reform*, Washington DC: Brooking Institution.
- Kish, L. (1978). *Survey Sampling*, New York: John Wiley.
- Kritikos, M., Markellos, R. and Prastakos, G. (2010). Corporate real estate analysis: evaluating telecom branch efficiency in Greece. *Applied Economics*, **42**, 1133-1143.
- Μακρυδημήτρης, Α. (2003). *Κράτος και Κοινωνία των Πολιτών*, Αθήνα: Μεταμεσονύκτιες Εκδόσεις.
- Μακρυδημήτρης, Α. (2010). *Δημόσια Διοίκηση: Στοιχεία Διοικητικής Οργάνωσης*, Αθήνα, Εκδόσεις: Σάκκουλας.
- Miliotis, P. (1992). Data envelopment analysis applied to evaluation distribution districts. *Journal of the Operational Research Society*, **43**, 549-555.
- Mixhaelides, P., Belegri-Roboli, A. and Marinos, T. (2010). Evaluating the technical efficiency of trolley buses in Athens, Greece. *Journal of Public Transportation*. **13**, 93-109.
- Μπουραντάς, Δ. (2005). *Ηγεσία*, Αθήνα, Εκδόσεις: Κριτική.
- Nichols, K. (1997). The crucial edge of reinvention: a primer on scoping and measuring for organizational change. *Public Administration Quarterly*, **21**, 405-418.
- Papoulias, D. and Darzentas, J. (1990). OR in Greece myth and reality. *European Journal of Operational Research*, **49**, 289-294.
- Παπούλιας, Δ. (2002). *Η Στρατηγική Διοίκηση Επιχειρήσεων και Αλλαγών*, Αθήνα, Εκδόσεις: Καστανιώτη.
- Richard, L. I., David, R.S. and Stinson, J.P. (1986). *Quantitative Approaches to Management*, New York: Mc Graw – Hill Book. Co.
- Tsekeris, T. (2011). Greek airports: Efficiency measurement and analysis of determinants. *Journal of Air Transport Management*, **17**, 139-141.
- Yamane, T. (1967). *Elementary Sampling Theory*, Prentice Hall Inc., Englewood Cliffs.



ΤΑ ΣΤΑΤΙΣΤΙΚΑ ΣΤΟΙΧΕΙΑ, Η ΕΥΡΩΖΩΝΗ ΚΑΙ Η ΚΡΙΣΗ ΧΡΕΟΥΣ ΤΗΣ ΕΛΛΑΔΑΣ

Ανδρέας Ν. Φιλίππου

Καθηγητής Πιθανοτήτων και Στατιστικής του Πανεπιστημίου Πατρών
Μέλος της Ελληνικής Στατιστικής Αρχής
www.math.upatras.gr/~anphilip

ΠΕΡΙΛΗΨΗ

Παρουσιάζουμε συνοπτικά εισηγήσεις και στοιχεία για τα στατιστικά, τα ελλείμματα και το χρέος, που οδήγησαν την Ελλάδα στην ίδρυση της Ελληνικής Στατιστικής Αρχής και την προσφυγή της χώρας στην Ευρωζώνη και το Διεθνές Νομισματικό Ταμείο για οικονομική βοήθεια. Αναφερόμαστε, επίσης, στο θέμα της αναδιάρθρωσης του χρέους και στην κυβερνητική δημοσιονομική στρατηγική για το 2012-15.

Λέξεις Κλειδιά: Ευρωπαϊκή Επιτροπή, ΕΛΣΤΑΤ, ελλείμματα, αναδιάρθρωση, ΔΝΤ

1. ΕΙΣΑΓΩΓΗ

Είναι γνωστή η ρήση “there are lies, damned lies and statistics”, που αποδίδεται στον Disraeli. Είναι ακόμη πιο γνωστή στους επιστήμονες στατιστικούς -και όχι μόνο- η πρόταση “how to lie with statistics”, που αποτελεί τον τίτλο του δημοφιλέστερου βιβλίου στατιστικής [1] του δημοσιογράφου και συγγραφέα Darrel Huff.

Θυμήθηκα αυτές τις προτάσεις, διότι πρόσφατα η φράση “Greek statistics” είχε καταντήσει ανέκδοτο και χαρακτηρισμός ανυποληψίας εναντίον της Ελλάδας. Πως έγινε αυτό;

2. ΣΤΑΤΙΣΤΙΚΑ ΣΤΟΙΧΕΙΑ ΔΗΜΟΣΙΟΥ ΕΛΛΕΙΜΜΑΤΟΣ ΚΑΙ ΧΡΕΟΥΣ

Στις 2 και 21 Οκτωβρίου 2009 οι ελληνικές αρχές διαβίβασαν στη Eurostat δύο διαφορετικές δέσμες στατιστικών στοιχείων για το έλλειμμα και το χρέος της ελληνικής κυβέρνησης για τα έτη 2005-2008 και μία εκτίμηση για το 2009. Η πρώτη εστάλη από τη κυβέρνηση του Κώστα Καραμανλή και η δεύτερη από τη νέα κυβέρνηση του Γιώργου Παπανδρέου. Στη δεύτερη, το έλλειμμα για το 2008 αναθεωρήθηκε από το 5,0% του ΑΕΠ (το ποσοστό που έστειλε η Ελλάδα και πιστοποίησε και δημοσίευσε η Eurostat τον Απρίλιο του 2009) στο 7,7% του ΑΕΠ.

Το προγραμματιζόμενο έλλειμμα για το 2009 αναθεωρήθηκε από το 3,7% (το ποσοστό που δηλώθηκε την άνοιξη) στο 12,5% του ΑΕΠ.

Η αξιωματική αντιπολίτευση διαμαρτυρήθηκε έντονα και κατηγόρησε την κυβέρνηση Παπανδρέου ότι διόγκωσε τα στοιχεία του ελλείμματος και του χρέους για πολιτικά οφέλη. Τα διόγκωσε;

Στις 10 Νοεμβρίου 2009 το Συμβούλιο ECOFIN κάλεσε την Ευρωπαϊκή Επιτροπή να ετοιμάσει έκθεση για τα «ανανεωμένα προβλήματα στα ελληνικά δημοσιονομικά στατιστικά στοιχεία» ... «και να προτείνει κατάλληλα μέτρα». Στην έκθεσή της [2], με ημερομηνία 8 Ιανουαρίου 2010 η Ευρωπαϊκή Επιτροπή αναφέρει:

Revisions of this magnitude in the estimated past government deficit ratios have been extremely rare in other EU Member States, but have taken place for Greece on several occasions. These most recent revisions are an illustration of the lack of quality of the Greek fiscal statistics (and of macroeconomic statistics in general) and show that the progress in the compilation of fiscal statistics in Greece, and the intense scrutiny of the Greek fiscal data by Eurostat since 2004 (including 10 EDP (excessive deficit procedure) visits and 5 reservations on the notified data), have not sufficed to bring the quality of Greek fiscal data to the level reached by other EU Member States.

Στη συνέχεια, τονίζει. In particular, the report has found evidence of:

- Severe irregularities in the EDP notifications of April and October 2009, including submission of incorrect data, and non-respect of accounting rules ... ;
- Poor cooperation between the national services involved in the compilation of EDP figures, **as well as lack of independence of the NSSG and the GAO from the Ministry of Finance;**
- Lack of accountability in the individual provision of figures used in EDP notifications, (e.g. absence of written documentation or certification in some cases, exchange of data by phone);
- Unclear responsibility and/or lack of responsibility of the national services

These findings indicate that ..., the current set-up does not guarantee the independence, integrity and accountability of the national statistical authorities. In particular the professional independence of the NSSG from the Ministry of Finance is not assured, which has allowed the reporting of EDP data to be influenced by factors other than the regulatory and legally binding principles for the production of high quality European statistics.

Η Επιτροπή συνεχίζει ...και εισηγείται:

The Greek authorities need to tackle resolutely not only the outstanding methodological issues, but also and crucially they need to put in place transparent and reliable working practices between the national services concerned, and to revise the institutional setting in order to guarantee the professional independence and full accountability of the NSSG and of the other

services involved in the domain of EDP data. Unless the institutional weaknesses identified in this report are addressed and proper checks and balances introduced, the reliability of Greek deficit and debt data will remain in question (η έμφαση είναι δική μου).

Η Ελληνική κυβέρνηση ανταποκρίθηκε στην κριτική της Ευρωπαϊκής Επιτροπής. Πρότεινε σε δημόσια διαβούλευση πρώτα, και κατέθεσε μετά στη Βουλή των Ελλήνων και ψήφισε τον νόμο 3832 *Ελληνικό Στατιστικό Σύστημα (ΕΛ.Σ.Σ.) Σύσταση της Ελληνικής Στατιστικής Αρχής (ΕΛ.ΣΤΑΤ.) ως Ανεξάρτητης Αρχής*. Ο νόμος δημοσιεύθηκε στην Εφημερίδα της Κυβερνήσεως με ημερομηνία 9 Μαρτίου 2010.

Στη συνέχεια, μετά από διεθνή προκήρυξη και πρόταση του υπουργού Οικονομικών Γιώργου Παπακωνσταντίνου, η Διάσκεψη των Προέδρων της Βουλής των Ελλήνων ενέκρινε με πλειοψηφία μεγαλύτερη ή ίση των 4/5 τα 4 από τα 7 μέλη της Αρχής (Γεωργίου, Λογοθέτης, Γεωργαντά και Φιλίππου). Τον κ. Γεωργίου καταψήφισαν 2, ενώ οι υπόλοιποι ψηφίστηκαν ομόφωνα, με 2 λευκά. Απαντώντας σε κριτική ερώτηση βουλευτή, κατά τη διάρκεια της ακρόασης στη Βουλή, είπα το αυτονόητο: Η ανεξαρτησία και η εύρυθμη λειτουργία της Αρχής δεν κατοχυρώνονται μόνο από τον νόμο. Θα εξαρτηθούν από τα μέλη της, από τον υπουργό Οικονομικών που προτείνει την πλειοψηφία των μελών, και από τη Βουλή που τα εγκρίνει και εποπτεύει τη λειτουργία της Αρχής. Για παράδειγμα, στις Ηνωμένες Πολιτείες της Αμερικής, όπως είναι γνωστό, η ισότητα κατοχυρώνεται από το Σύνταγμα. Όμως, χρειάστηκε ένας εμφύλιος για να καταργηθεί η δουλεία και περισσότερο από 100 χρόνια μετά τον εμφύλιο για να επιτευχθεί η ισότητα στην πράξη.

Στις 2 Αυγούστου 2010 η νέα ανεξάρτητη Αρχή ανέλαβε τα καθήκοντά της και άρχισε να παράγει έργο. Στις 15 Νοεμβρίου 2010, η Eurostat ανακοίνωσε στοιχεία [3, 4] πολύ χειρότερα για την Ελλάδα από εκείνα που έστειλε στις 21 Οκτωβρίου 2009 η κυβέρνηση Παπανδρέου, μόλις ανέλαβε την εξουσία, και κατηγορήθηκε ότι διόγκωσε το έλλειμμα και το χρέος για πολιτικά οφέλη:

«Τα μεγαλύτερα κυβερνητικά **ελλείμματα** (στην Ευρωζώνη των 16 και την ΕΕ των 27), ως ποσοστά του ΑΕΠ, παρατηρήθηκαν στην Ελλάδα (15,4%), την Ιρλανδία (14,4%), το Ηνωμένο Βασίλειο (11,4), την Ισπανία (11,1%), τη Λετονία (10,2%) και την Πορτογαλία (9,3%)». ... Η Ελλάδα ήταν πρωταθλήτρια και στο «**δημόσιο χρέος**, που ανερχόταν σε 126,8% του ΑΕΠ το 2009», ενώ «τα προηγούμενα δύο χρόνια ήταν 110,3% και 105,0% του ΑΕΠ», με αντίστοιχα «ελλείμματα 9,4% και 6,4% του ΑΕΠ».

Συνεπώς, η κυβέρνηση Παπανδρέου όχι μόνο δεν διόγκωσε τα στοιχεία, όπως κατηγορήθηκε, αλλά τα υποεκτίμησε. Αντίθετα, η προηγούμενη κυβέρνηση δημιούργησε και απέκρυψε χρέη επιμελώς, κυρίως μεταξύ 2007-2009. Μετά από ένα χρόνο συνεχών προσπαθειών και αποτελεσματικού έργου, η Ελλάδα ξανακέρδισε την στατιστική αξιοπιστία της.

3. Η ΕΥΡΩΖΩΝΗ ΚΑΙ Η ΚΡΙΣΗ ΧΡΕΟΥΣ ΤΗΣ ΕΛΛΑΔΑΣ

Όπως είναι γνωστό, η Ευρωζώνη, με κύρια συμβολή της Γερμανίας και της Γαλλίας, είναι η δεύτερη μεγαλύτερη οικονομία στον κόσμο μετά τις Ηνωμένες Πολιτείες. Ακολουθούν η Κίνα, η Ιαπωνία και η Βραζιλία. Η παγκόσμια οικονομική κρίση και τα προβλήματα χρέους της Ελλάδας, της Ιρλανδίας και της Πορτογαλίας, παρά το μικρό μέγεθος των οικονομιών τους, προκάλεσαν στις αρχές του περασμένου χρόνου μεγάλα προβλήματα στο ευρώ. Μερικοί μάλιστα καταστροφολόγοι προέβλεπαν την κατάργηση του ευρωπαϊκού νομίσματος και τη διάλυση της Ευρωζώνης.

Η Ελλάδα, στις 19 Μαΐου 2010, αντιμετώπιζε στάση πληρωμών. Μετά από μαραθώνιες διαβουλεύσεις, ο πρωθυπουργός Γιώργος Παπανδρέου προσέφυγε στην Ευρωπαϊκή Ένωση και το Διεθνές Νομισματικό Ταμείο για δανειακή βοήθεια.

Οι αρχηγοί των κρατών και κυβερνήσεων της ζώνης του ευρώ, που συνήλθαν στις Βρυξέλες στις 7 Μαΐου 2010, αποφάσισαν να δώσουν στην Ελλάδα το μεγαλύτερο δάνειο που πήρε ποτέ Ευρωπαϊκή χώρα. Η Ελλάδα είχε ήδη ψηφίσει ένα φιλόδοξο πρόγραμμα περιορισμού των δαπανών και αύξησης των εσόδων με στόχο την μείωση του ελλείμματος κατά 5,5% του ΑΕΠ το 2010 (από 13,6% σε 8,1%). Ανάμεσα σε άλλα, οι αρχηγοί δήλωσαν [5]:

«Υστερα από το αίτημα της ελληνικής κυβέρνησης, την 23^η Απριλίου, και από τη συμφωνία στην οποία κατέληξε η Ευρωομάδα στις 2 Μαΐου, θα διαθέσουμε στην Ελλάδα 80 δισεκατομμύρια ευρώ σε κοινό πακέτο με το ΔΝΤ, ύψους 110 δισεκατομμυρίων ευρώ. Η Ελλάδα θα λάβει μια πρώτη δόση τις προσεχείς ημέρες, πριν από τις 19 Μαΐου. Το πρόγραμμα που υιοθέτησε η ελληνική κυβέρνηση είναι φιλόδοξο και ρεαλιστικό. Αντιμετωπίζει τις σοβαρές δημοσιονομικές ανισορροπίες, θα καταστήσει την οικονομία ανταγωνιστικότερη και θα θέσει τις βάσεις για ισχυρότερη και διαρκέστερη οικονομική μεγέθυνση και δημιουργία θέσεων απασχόλησης. ... Οι αποφάσεις που λαμβάνουμε αντικατοπτρίζουν τις αρχές της υπευθυνότητας και της αλληλεγγύης, που κατοχυρώνονται από τη Συνθήκη της Λισσαβόνας και βρίσκονται στο επίκεντρο της νομισματικής ένωσης».

Δuo μέρες αργότερα ιδρύθηκε από τις χώρες της Ευρωζώνης ο προσωρινός ευρωπαϊκός μηχανισμός δημοσιονομικής σταθερότητας (European Financial Stability Facility (EFSF)). Θα αντικατασταθεί το 2013 από τον μόνιμο ευρωπαϊκό μηχανισμό σταθερότητας (European Stability Mechanism (ESM)).

Είχε προηγηθεί στις 6 Μαΐου 2010 η υπερψήφιση σε νόμο [6] του νομοσχεδίου «μέτρα για την εφαρμογή του μηχανισμού στήριξης της Ελληνικής οικονομίας από τα κράτη μέλη της ζώνης του ευρώ και το ΔΝΤ.» Υπέρ ψήφισαν οι βουλευτές του ΠΑΣΟΚ, πλην Δημαρά, Σακοράφας και Οικονόμου που δήλωσαν «παρών», οι βουλευτές του ΛΑΟΣ και η Ντόρα Μπακογιάννη. Το ΠΑΣΟΚ διέγραψε τους τρεις πρώτους από την Κοινοβουλευτική του Ομάδα και η Νέα Δημοκρατία διέγραψε την κα Μπακογιάννη από το κόμμα. Τα μέτρα καταψήφισαν η Νέα Δημοκρατία, το ΚΚΕ και ο ΣΥΡΙΖΑ.

Ήταν η προσφυγή και το «μνημόνιο» αναγκαία; Είτε προσφεύγαμε, είτε πτωχεύαμε. Αποφύγαμε την πτώχευση; Ναι, αν ελαττώσουμε όσο πρέπει τις δαπάνες και αν αυξήσουμε την ανάπτυξη και τα έσοδα, παράγοντας πρωτογενή πλεονάσματα.

4. ΑΝΑΔΙΑΡΘΡΩΣΗ;

Πρέπει να διεκδικήσει η Ελλάδα αναδιάρθρωση και περικοπή μέρους του χρέους της; Την έχουν εισηγηθεί δημοσίως διάφοροι οικονομολόγοι, τα περιοδικά Economist και Spiegel, η Wall Street Journal και ο πρώην πρωθυπουργός Κώστας Σημίτης [7]. Η κυβέρνηση όμως και η αξιωματική αντιπολίτευση την απορρίπτουν και απαντούν ότι η δημόσια συζήτηση για αναδιάρθρωση βλάπτει την χώρα. Συζητά η κυβέρνηση αναδιάρθρωση; Όχι δημόσια. Όμως, όπως ορισμένα πράγματα λέγονται και δεν γίνονται, ορισμένα γίνονται και δεν λέγονται πρόωρα. Για παράδειγμα, παρότι δεν έχει ονομαστεί έτσι, η επιμήκυνση αποπληρωμής των 110 δισεκατομμυρίων ευρώ και η μείωση του επιτοκίου των 80 δις κατά μία ποσοστιαία μονάδα, που πέτυχε η Ελλάδα στο Ευρωπαϊκό Συμβούλιο της 25^{ης} Μαρτίου 2011 [8, 9], αποτελεί επίτευγμα και είναι αναδιάρθρωση. Μικρή βέβαια ποσοτικά (6 δισεκατομμύρια ευρώ περίπου), σε σύγκριση με το χρέος μας (περίπου 340 δις ευρώ), αλλά απαραίτητη ως προς το σκέλος της επιμήκυνσης. Δεν θα μπορούσε να συμφωνηθεί κάτι ανάλογο και για τα άλλα δάνεια, ή / και επιπλέον δανειοδότηση και περικοπή;

5. ΜΕΣΟΠΡΟΘΕΣΜΟ ΠΛΑΙΣΙΟ ΔΗΜΟΣΙΟΝΟΜΙΚΗΣ ΣΤΡΑΤΗΓΙΚΗΣ 2012-2015

Πρόσφατα, μιλώντας στο Υπουργικό του Συμβούλιο και στην Κοινοβουλευτική Ομάδα και το Εθνικό Συμβούλιο του ΠΑΣΟΚ, ο έλληνας Πρωθυπουργός προανήγγειλε το «μεσοπρόθεσμο πλαίσιο δημοσιονομικής στρατηγικής 2012-2015» [10], που αναμένεται να κατατεθεί στη Βουλή εντός του Μαΐου 2011.

Στο πλαίσιο αυτό, ανάμεσα σε άλλα, αναφέρεται:

«Έχουμε καταφέρει την μεγαλύτερη μείωση ελλείμματος που έχει γίνει ποτέ στην Ελλάδα ή σε χώρα της Ευρωζώνης, μέσα σε ένα χρόνο. Θέσαμε τις βάσεις για διαφάνεια, αξιοκρατία και αξιοπιστία στη Δημόσια Διοίκηση. Βάλαμε τέλος στην αδιαφάνεια και αναξιοπιστία των στατιστικών».

Επίσης, με την αξιοποίηση της δημόσιας περιουσίας, η κυβέρνηση στοχεύει:

Να στηρίξει την αναπτυξιακή διαδικασία φέρνοντας επενδύσεις και θέσεις εργασίας.

Να αντλήσει 15 δισεκατομμύρια ευρώ την περίοδο 2011-2013 και 50 δισεκατομμύρια ευρώ συνολικά έως το 2015.

Αυτά τα έσοδα θα μειώσουν το χρέος κατά 20% του ΑΕΠ έως το 2015, καθιστώντας το πιο βιώσιμο με σημαντικά χαμηλότερη επιβάρυνση σε τόκους

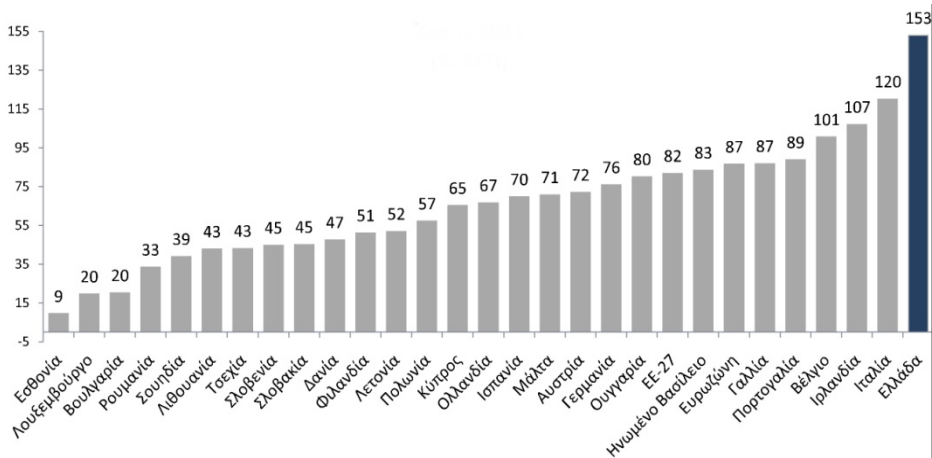
Σύμφωνα με το πλαίσιο:

Η ύφεση στην Ελλάδα το 2011 θα περιοριστεί στο 3% από το 4, 5% το 2010.

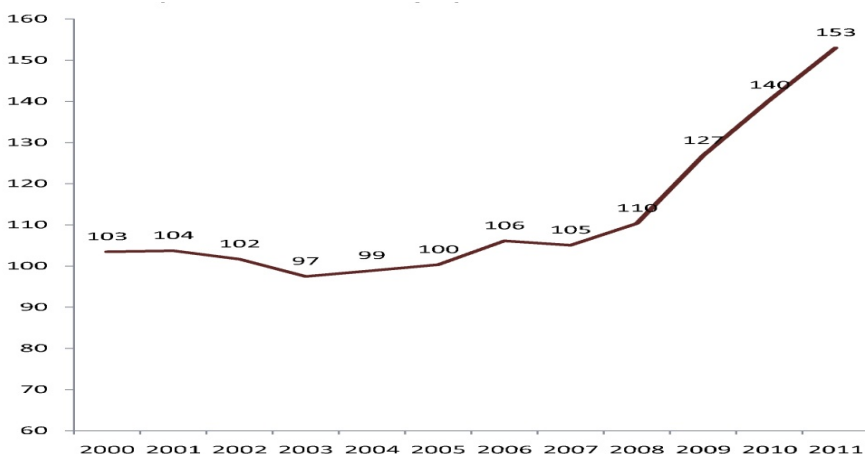
Η Ελλάδα αναμένεται να επιστρέψει σε θετικούς ρυθμούς ανάπτυξης το 2012. Ο δημοσιονομικός μας στόχος είναι να επιτύχουμε επιδόσεις που είχαμε και στο παρελθόν. Δηλαδή, έως το 2015, οι δαπάνες του κράτους να μειωθούν από 53% του ΑΕΠ το 2009 σε 44% (επίπεδο δαπανών του 2003) και τα έσοδα να φτάσουν στο 43% του ΑΕΠ (επίπεδο εσόδων του 2000), από 38% του ΑΕΠ.

Είναι ένας φιλόδοξος και δύσκολος, αλλά εφικτός στόχος (Διαγράμματα 1-4 [10]).

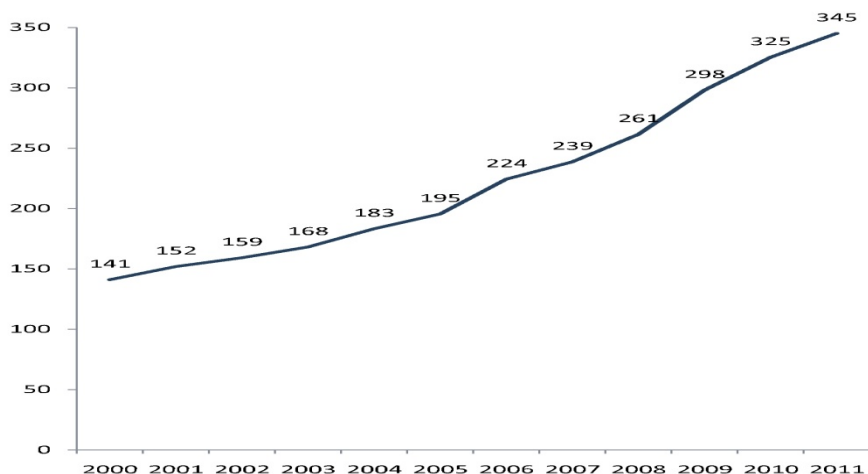
Διάγραμμα 1. ΧΡΕΟΣ ΕΥΡΩΠΑΙΚΩΝ ΧΩΡΩΝ ΩΣ % ΑΕΠ ΤΟ 2011



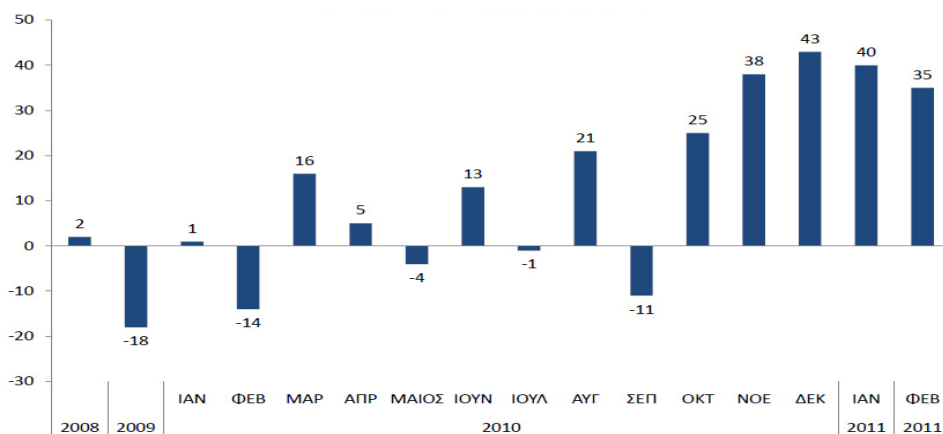
Διάγραμμα 2. ΧΡΕΟΣ ΕΛΛΑΔΑΣ (ΓΚ) ΩΣ % ΑΕΠ 2000-2011



Διάγραμμα 3. ΧΡΕΟΣ ΕΛΛΑΔΑΣ (ΓΚ) ΣΕ ΔΙΣ. ΕΥΡΩ 2000-2011



Διάγραμμα 4. ΡΥΘΜΟΣ ΜΕΤΑΒΟΛΗΣ ΕΛΛΗΝΙΚΩΝ ΕΞΑΓΩΓΩΝ (2008-11)



6. ΤΕΛΕΥΤΑΙΑ ΣΤΟΙΧΕΙΑ ΤΗΣ EUROSTAT

Σύμφωνα με τα τελευταία στοιχεία της EUROSTAT [11], που ανακοινώθηκαν στις 26 Απριλίου 2011, «το δημόσιο έλλειμμα της Ελλάδας μειώθηκε το 2010 στο 10,5% του ΑΕΠ από 15, 4% του ΑΕΠ που ήταν το 2009». Είναι η μεγαλύτερη μείωση ελλείμματος που έχει γίνει ποτέ στην Ελλάδα ή σε χώρα της Ευρωζώνης, μέσα σε ένα χρόνο. Ο αρχικός στόχος βέβαια για το έλλειμμα το 2010 ήταν 8,1% του ΑΕΠ, αλλά τότε η εκτίμηση για το έλλειμμα το 2009 ήταν 13,6% του ΑΕΠ.

Η πορεία της Ελλάδας επαφίεται στον πατριωτισμό και την εργατικότητα των Ελλήνων, στην αλληλεγγύη των εταίρων της στην Ευρωζώνη, και στην κρίση της ηγεσίας και των πολιτών της.

ABSTRACT

We quote parts of a study of the European Commission on Greek deficit and debt statistics and discuss how it led to the establishment of the Hellenic Statistical Authority. We also argue that large fiscal deficits and the excessive debt forced Greece to seek help from the Eurozone and the International Monetary Fund. Finally, we touch upon the question of debt restructuring and fiscal strategy for 2012-15.

Ευχαριστίες: Εκφράζουμε ευχαριστίες στην Επιτροπή Έκδοσης των Πρακτικών και στον ανώνυμο κριτή για τις χρήσιμες παρατηρήσεις του.

ΑΝΑΦΟΡΕΣ

Huff D. (1954). *How to Lie with Statistics*. Norton, New York.

Report on Greek government deficit and debt statistics. European Commission Brussels, 8.1.2010 COM(2010) 1 final.

Euro area and EU27 government deficit at 6.3% and 6.8% of GDP respectively (15 November 2010).

Information note on Greek fiscal data (15 November 2010)

Δήλωση των αρχηγών κρατών και κυβερνήσεων της ζώνης του ευρώ, Βρυξέλλες, 7 Μαΐου 2010 .

N. 3845/10 Μέτρα για την εφαρμογή του μηχανισμού στήριξης της ελληνικής οικονομίας από τα κράτη-μέλη της ζώνης του ευρώ και το ΔΝΤ (ΦΕΚ Α 65/06.05.2010) Παράρτημα I Παράρτημα II Παράρτημα III - Μνημόνιο Παράρτημα IV - Μνημόνιο.

Πρόταση-σοκ από Σημίτη. Βήμα, 17 Απριλίου 2011.

Συμπεράσματα των αρχηγών κρατών και κυβερνήσεων της ζώνης του ευρώ της 11^{ης} Μαρτίου 2011.

European Council - 24 and 25 March 2011 - Conclusions.

Μεσοπρόθεσμο Πλαίσιο Δημοσιονομικής Στρατηγικής 2012-2015.

Euro area and EU27 government deficit at 6.0% and 6.4% of GDP respectively (26.04.2011)



ΔΕΟΝΤΟΛΟΓΙΑ ΚΑΙ ΕΜΠΙΣΤΕΥΤΙΚΟΤΗΤΑ ΤΗΣ ΠΛΗΡΟΦΟΡΙΑΣ ΣΤΙΣ ΔΕΙΓΜΑΤΟΛΗΠΤΙΚΕΣ ΕΡΕΥΝΕΣ

Τάσος Χριστοφίδης

Τμήμα Μαθηματικών και Στατιστικής
Πανεπιστήμιο Κύπρου
tasos@ucy.ac.cy

ΠΕΡΙΛΗΨΗ

Οι έννοιες της δεοντολογίας και της διαφάνειας εκ πρώτης όψεως φαίνεται να είναι διαμετρικά αντίθετες με την έννοια της εμπιστευτικότητας των πληροφοριών. Στις δειγματοληπτικές έρευνες πολλές φορές οι ερωτώμενοι έχουν να απαντήσουν ερωτήσεις που αφορούν την ιδιωτική τους ζωή και που πολλές φορές οι απαντήσεις σε αυτές μπορεί να τους κατατάξουν σε μια κοινωνικά στιγματισμένη ομάδα ανθρώπων. Στην εργασία αυτή θα δούμε πώς και σε ποιο βαθμό συμβιβάζονται οι έννοιες της δεοντολογίας και της διαφάνειας με την έννοια της ιδιωτικότητας (privacy) του πολίτη.

Λέξεις Κλειδιά: Τεχνικές τυχαιοποιημένης απάντησης, στιγματιστικά χαρακτηριστικά, προστασία ιδιωτικότητας.

1. ΕΙΣΑΓΩΓΗ

Οι έννοιες της δεοντολογίας και της διαφάνειας εκ πρώτης όψεως φαίνονται να είναι διαμετρικά αντίθετες με την έννοια της εμπιστευτικότητας των πληροφοριών. Στις δειγματοληπτικές έρευνες πολλές φορές οι ερωτώμενοι έχουν να απαντήσουν ερωτήσεις που αφορούν την ιδιωτική τους ζωή και που πολλές φορές οι απαντήσεις σε αυτές μπορεί να τους κατατάξουν σε μια κοινωνικά στιγματισμένη ομάδα ανθρώπων. Στην εργασία αυτή θα δούμε πώς και σε ποιο βαθμό συμβιβάζονται οι έννοιες της δεοντολογίας και της διαφάνειας με την έννοια της ιδιωτικότητας (privacy) του πολίτη. Επίσης θα εξετάσουμε το πόσο διαφανής και δεοντολογικός είναι ο τρόπος παρουσίασης των αποτελεσμάτων των δειγματοληπτικών ερευνών αλλά και ο τρόπος συλλογής των πληροφοριών.

Πολλές φορές χτυπά το τηλέφωνο κυρίως ενωρίς το βράδυ και στην άλλη γραμμή ένας νεαρός ή νεαρή συνήθως μιλώντας πολύ γρήγορα αρχίζει να κάνει διάφορες ερωτήσεις στα πλαίσια της διεξαγωγής μιας δειγματοληπτικής έρευνας. Βεβαίως το άτομο αυτό ζητά τη συγκατάθεσή μας για να κάνει τις ερωτήσεις και πολλοί από εμάς ίσως από ευγένεια δεχόμαστε. Το άτομο αυτό γνωρίζει ποιο αριθμό έχει καλέσει

και ίσως με ποιο άτομο μιλά, όμως εμείς όχι. Το άτομο στην άλλη γραμμή αναφέρει συνήθως το όνομα της εταιρείας που διενεργεί την έρευνα.

Στις μέρες μας όλο και πιο πολλές έρευνες γίνονται μέσω διαδικτύου. Πιστεύω ότι στο σύντομο μέλλον το διαδίκτυο και η κινητή τηλεφωνία θα έχουν τη μερίδα του λέοντος αναφορικά με τις δειγματοληπτικές έρευνες. Σημειώνεται ότι η υποβολή πληροφοριών μέσω διαδικτύου δεν είναι ανώνυμη. Ο παραλήπτης των πληροφοριών γνωρίζει τον αποστολέα. Μπορεί όχι προσωπικά, αλλά σίγουρα γνωρίζει την ηλεκτρονική του ταυτότητα.

Πόσο δεοντολογικό είναι αυτό, δηλαδή να μας ζητούν πληροφορίες ενώ δεν τους γνωρίζουμε και δεν τους βλέπουμε; Ακόμη και αν τους βλέπουμε ας πούμε όπως συμβαίνει στην περίπτωση προσωπικών συνεντεύξεων, πόσο δεοντολογικό είναι να εμπιστευόμαστε πληροφορία που μας αφορά σε άγνωστα άτομα; Αν η ερώτηση είναι αν χρησιμοποιούμε το Α ή Β σαπούνι, κανένα πρόβλημα. Όμως πόσο δεοντολογικό είναι να μας ερωτούν για παράδειγμα:

- 1) αν είμαστε ομοφυλόφιλοι
- 2) αν φοροδιαφεύγουμε
- 3) πόσο εισόδημα έχουμε
- 4) αν έχουμε εξωσυζυγικές σχέσεις
- 5) αν έχουμε διαπράξει μια αξιόποινη πράξη
- 6) αν έχουμε κάνει έκτρωση
- 7) αν είμαστε χρήστες ναρκωτικών
- 8) ακόμη και ποιο κόμμα θα ψηφίσουμε, αφού στο κάτω κάτω η ψήφος πρέπει να είναι μυστική

Πέρα από το γεγονός ότι δεν είναι δεοντολογικό να μας ερωτούν τέτοιου είδους ερωτήσεις αφού αυτές αφορούν την ιδιωτική μας ζωή, ένας λογικός άνθρωπος θα πρέπει να υποθέσει ότι οι περισσότεροι από αυτούς που ερωτώνται δεν θα είναι ειλικρινείς. Κα αφού δεν θα είναι, αντιλαμβάνεται κάποιος ότι και τα ευρήματα τα έρευνας θα είναι εν αμφιβόλω.

Από τη σκοπιά τώρα του ερευνητή, το θέμα της δεοντολογίας τον αφορά μόνο στο βαθμό που τα αποτελέσματα της έρευνάς του δεν θα είναι ορθά. Συνεπώς τι κάνουμε;

Είναι προς το συμφέρον του ερευνητή να χρησιμοποιήσει μεθοδολογία η οποία όχι μόνο να προστατεύει την ιδιωτικότητα του ερωτώμενου αλλά πολύ περισσότερο να πείθει των ερωτώμενο ότι πράγματι η ιδιωτικότητα του προστατεύεται.

2. ΤΕΧΝΙΚΕΣ ΤΥΧΑΙΟΠΟΙΗΜΕΝΗΣ ΑΠΑΝΤΗΣΗΣ

Τις τελευταίες δεκαετίες έχουν αναπτυχθεί τεχνικές που στόχο έχουν τη συλλογή πληροφοριών με σκοπό την εκτίμηση της σχετικής συχνότητας στον πληθυσμό ενός ποιοτικού χαρακτηριστικού ή τη εκτίμηση μιας παραμέτρου ενός ποσοτικού χαρακτηριστικού όπως αυτά που είδαμε πιο πάνω. Οι τεχνικές αυτές παράλληλα

προστατεύουν την ιδιωτικότητα του πολίτη, δηλαδή από την απάντηση και μόνο δεν είναι δυνατόν ο ερευνητής να βγάλει συμπεράσματα για τον ερωτώμενο. Οι πιο πολλές τεχνικές ανήκουν στην κατηγορία των τεχνικών τυχαιοποιημένης απάντησης (randomized response). Η θεμελίωση των τεχνικών αυτών ξεκινά από το Warner (1965).

Ας δούμε τη βασική αρχή των τεχνικών αυτών μέσα από τη μέθοδο του Warner.

Υποθέτουμε ότι θέλουμε να εκτιμήσουμε τη συχνότητα στον πληθυσμό του επίμαχου χαρακτηριστικού A . Έστω η συχνότητα αυτή ότι είναι π . Κάθε ερωτώμενος τοποθετείται (με πιθανότητα p) με ένα ναι ή ένα όχι στην δήλωση

(I) Έχω το χαρακτηριστικό A

και με πιθανότητα $1 - p$ στη δήλωση

(II) Έχω το χαρακτηριστικό A^c .

Η τοποθέτηση στη δήλωση (I) ή (II) αποφασίζεται με τη χρήση μιας «συσκευής τυχαιοποίησης» χωρίς την παρουσία του ερευνητή, δηλαδή όταν ο ερωτώμενος τοποθετηθεί, ο ερευνητής δεν γνωρίζει αν η τοποθέτηση αφορά τη δήλωση (I) ή τη δήλωση (II).

Έστω τώρα ότι \hat{g} είναι το δειγματικό ποσοστό των ερωτώμενων που απάντησαν «ναι».

Τότε

$$\hat{\pi} = \frac{\hat{g} - (1 - p)}{2p - 1}$$

είναι αμερόληπτη εκτιμήτρια του π με διασπορά

$$Var(\hat{\pi}) = \frac{\pi(1 - \pi)}{n} + \frac{p(1 - p)}{n(2p - 1)^2}. \quad (1)$$

Η πιο πάνω εκτιμήτρια προκύπτει ως εξής:

Έστω ότι η απάντηση στη δήλωση (I) καθορίζεται από μια τυχαία μεταβλητή X που ακολουθεί την κατανομή Bernoulli (p). Έστω επίσης ότι το αν έχει κάποιος το επίμαχο χαρακτηριστικό καθορίζεται από μια τυχαία μεταβλητή Y (ανεξάρτητη από τη X) που ακολουθεί την κατανομή Bernoulli (π). Τότε η απάντηση (ναι ή όχι) περιγράφεται από την τυχαία μεταβλητή

$$W = XY + (1 - X)(1 - Y) = (2X - 1)Y + 1 - X$$

Από την οποία προκύπτει ότι

$$E(W) = (2p - 1)\pi + 1 - p$$

Και συνεπώς

$$\pi = \frac{E(W) - (1 - p)}{2p - 1}.$$

Παρατηρήσεις:

- 1) Είναι προφανές ότι αν η απάντηση του ερωτώμενου είναι «ναι», δεν σημαίνει ότι αυτός έχει το επίμαχο χαρακτηριστικό. Μπορεί να σημαίνει ακριβώς το αντίθετο.
- 2) Η τοποθέτηση σε μια από τις δηλώσεις δεν γίνεται με την ίδια πιθανότητα δηλαδή δεν μπορεί να υπάρξει εκτίμηση αν $p = 0.5$.
- 3) Στην (1) παρατηρούμε ότι η διασπορά της εκτιμήτριας έχει δύο συνιστώσες, μια που οφείλεται στην τυχαιότητα και μια άλλη που οφείλεται στη χρήση της συσκευής τυχαιοποίησης.
- 4) Από την (1) συμπεραίνουμε ότι το p δεν πρέπει να είναι πολύ κοντά στο 0.5 γιατί αυτό εκτοξεύει τη διασπορά της εκτιμήτριας. Από την άλλη, το p δεν πρέπει να είναι πολύ μακριά από το 0.5 γιατί τότε οι ερωτώμενοι να μην είναι και πολύ συνεργάσιμοι.

Από το 1965 και μετά παρατηρείται μια αυξημένη δραστηριότητα στον τομέα των τεχνικών τυχαιοποιημένης απάντησης. Κάποιες τεχνικές έχουν αντικαταστήσει τη δεύτερη δήλωση με δήλωση που αφορά κάποιο άλλο μη επίμαχο χαρακτηριστικό (βλ. Greenberg et al. (1969)). Άλλες τεχνικές επιτρέπουν πολλαπλές απαντήσεις (βλ. Christofides (2003)), ενώ υπάρχουν τεχνικές που επιτρέπουν στον ερωτώμενο να απαντήσει ευθέως την ερώτηση ή να χρησιμοποιήσει τυχαιοποίηση (βλ. Chaudhuri and Saha (2005)).

Η πιο πάνω τεχνική αφορά την εκτίμηση της συχνότητας στον πληθυσμό ενός χαρακτηριστικού που συνοδεύεται από κοινωνικό στίγμα. Όμως τεχνικές τυχαιοποιημένης απάντησης υπάρχουν και για την εκτίμηση ενός ποσοτικού χαρακτηριστικού. Οι τεχνικές αυτές στηρίζονται περίπου στην εξής διαδικασία:

Αν Y είναι η τυχαία μεταβλητή που αφορά το επίμαχο χαρακτηριστικό, αυτό «μολύνεται» από μια άλλη τυχαία μεταβλητή X με γνωστή κατανομή και ο ερωτώμενος δίνει απάντηση που περιγράφεται από την τυχαία μεταβλητή

$$W = h(X, Y)$$

με h γνωστή συνάρτηση. Συνήθως

$$h(X, Y) = X + Y$$

ή

$$h(X, Y) = XY .$$

Ο ερευνητής, είναι σε θέση να εκτιμήσει την κατανομή (ή τουλάχιστον κάποιες παραμέτρους) της Y έχοντας γνώση και της κατανομής της «μολύνουσας» μεταβλητής X .

3. ΑΛΛΕΣ ΤΕΧΝΙΚΕΣ

Πέρα από τις τεχνικές τυχαιοποιημένης απάντησης υπάρχουν και άλλες τεχνικές για τη συλλογή πληροφοριών που αφορούν επίμαχα χαρακτηριστικά. Η πιο διαδεδομένη είναι η «τεχνική μέτρησης δηλώσεων» (item count technique) (βλ. Miller et al. (1986) και Chaudhuri and Christofides (2007)). Για την εφαρμογή της μεθόδου αυτής χρειάζονται δύο (ανεξάρτητα) δείγματα. Στο ένα δείγμα δίνεται ερωτηματολόγιο που περιλαμβάνει αριθμό δηλώσεων. Ο ερωτώμενος i δηλώνει πόσες από αυτές τις δηλώσεις τον αφορούν, χωρίς να λείει ποιες (έστω X_i ο αριθμός αυτός). Στο δεύτερο δείγμα δίνεται το ίδιο ερωτηματολόγιο στο οποίο όμως έχει συμπεριληφθεί και μια δήλωση που αφορά το επίμαχο χαρακτηριστικό. Και πάλι κάθε ερωτώμενος καλείται να δηλώσει πόσες από τις δηλώσεις του ερωτηματολογίου τον αφορούν χωρίς να πει ποιές. Έστω Y_j ο αριθμός που αναφέρει ο ερωτώμενος j . Τότε η ποσότητα

$$\hat{\pi} = \bar{Y} - \bar{X}$$

είναι εκτιμήτρια του πληθυσμιακού ποσοστού π που έχει το στιγματιστικό χαρακτηριστικό.

4. ΠΡΟΣΤΑΣΙΑ ΙΔΙΩΤΙΚΟΤΗΤΑΣ

Ένα μεγάλο ερώτημα στις τεχνικές τυχαιοποιημένης απάντησης και γενικά στις μεθόδους έμμεσης ερώτησης είναι το εξής:

Προστατεύουν την ιδιωτικότητα του ερωτωμένου και σε ποίο βαθμό;

Η απάντηση στο πιο πάνω ερώτημα δεν είναι εύκολη. Σίγουρα οι τεχνικές τυχαιοποιημένης απάντησης παρέχουν προστασία αλλά ο βαθμός ποικίλει από μέθοδο σε μέθοδο.

Και πώς μετράται ο βαθμός προστασίας; Κάποιοι χρησιμοποιούν την εξής μεθοδολογία για να καθορίσουν το βαθμό προστασίας.

Έστω R συμβολίζει μια απάντηση (ναι ή όχι) και A το ενδεχόμενο ο ερωτώμενος να ανήκει στην επίμαχη κατηγορία. Τότε οι ποσότητες $P(A | R)$ και $P(A^c | R)$ μπορούν να θεωρηθούν ως «πιθανότητες αποκάλυψης» (revealing probabilities). Αν

$P(A | R) > \pi$ τότε η απάντηση R εμπεριέχει κίνδυνο αναφορικά με το A ενώ αν $P(A^c | R) > 1 - \pi$ τότε η απάντηση R εμπεριέχει κίνδυνο αναφορικά με το A^c . Οι δύο αυτές ποσότητες συνδυάζονται στον πιο κάτω δείκτη επικινδυνότητας (measure of jeopardy):

$$J(R) = \frac{P(A | R) / \pi}{P(A^c | R) / (1 - \pi)}.$$

Όσο πιο μεγάλη είναι αυτή η τιμή, τόσο πιο μεγάλος θεωρείται ο κίνδυνος παραβίασης της ιδιωτικότητας του ερωτώμενου αν αυτός δώσει την απάντηση R .

Υπάρχει μεγάλη συζήτηση αναφορικά με τα μέτρα προστασίας της ιδιωτικότητας των ερωτωμένων στις τεχνικές τυχαιοποιημένης απάντησης. Δεν υπάρχουν μέτρα προστασίας που είναι αποδεκτά από όλους. Όμως εκείνο που σίγουρα ισχύει είναι ότι όσο πιο πολύ προστατεύεται η ιδιωτικότητα των ερωτωμένων τόσο πιο μεγάλη είναι η διασπορά της εκτιμήτριας.

5. ΑΙΣΘΗΣΗ ΤΗΣ ΠΡΟΣΤΑΣΙΑΣ ΤΗΣ ΙΔΙΩΤΙΚΟΤΗΤΑΣ

Είναι αυτονόητο ότι ένας άνθρωπος που έχει να επιλέξει αν θα συμμετάσχει ή όχι σε δειγματοληπτική έρευνα δεν γνωρίζει τίποτα για τους δείκτες προστασίας της ιδιωτικότητας. Ακόμη και αν είναι γνώστης της φιλοσοφίας που διέπει τις τεχνικές αυτές, ή ακόμη και αν μπορεί να αντιληφθεί τα μαθηματικά των τεχνικών αυτών δεν θα αποφασίσει με βάση τους δείκτες. Η απόφασή του θα στηριχθεί στη αίσθηση προστασίας της ιδιωτικότητας (perception of the protection of privacy). Θα συμμετάσχει μόνο αν πιστεύει ότι υπάρχει επαρκής προστασία της ιδιωτικότητάς του.

Στον τομέα αυτό πολύ λίγα πράγματα έχουν γίνει. Η έρευνα στον τομέα των δεικτών προστασίας προχωρεί διότι είναι σχετικά εύκολο να μαθηματικοποιήσει κάποιος την έννοια της προστασίας της ιδιωτικότητας και να κάνει συγκρίσεις μεταξύ των διαφόρων τεχνικών. Η μαθηματικοποίηση της αίσθησης της προστασίας της ιδιωτικότητας δεν είναι εύκολη υπόθεση και μάλλον είναι θέμα στο οποίο θα πρέπει να έχουν ενεργό συμβολή και επιστήμονες άλλων ειδικοτήτων. Όμως θεωρώ ότι είναι εξαιρετικής σημασίας διότι όσο καλή και να είναι μία μέθοδος, όσο ψηλός και να είναι ο δείκτης προστασίας της ιδιωτικότητας, όλα αυτά θα είναι χωρίς αξία αν οι ερωτώμενοι δείχνουν δυσπιστία στη μέθοδο. Πιστεύω ότι ένα μεγάλο μέρος της έρευνας στις τεχνικές συλλογής πληροφοριών με έμμεσο τρόπο θα επικεντρωθεί στον τομέα της μαθηματικοποίησης της αίσθησης προστασίας της ιδιωτικότητας.

6. ΠΑΡΟΥΣΙΑΣΗ ΑΠΟΤΕΛΕΣΜΑΤΩΝ ΚΑΙ ΔΕΟΝΤΟΛΟΓΙΑ

Τέλος θα ήθελα να θίξω το θέμα του πόσο δεοντολογικός και διαφανής είναι ο τρόπος παρουσίασης των αποτελεσμάτων, γενικά για τις δειγματοληπτικές έρευνες (και όχι μόνο αυτές που αφορούν στιγματιστικά χαρακτηριστικά) τουλάχιστον όπως παρουσιάζονται από τα MME. Ιδιαίτερα θα ήθελα να σταθώ στο θέμα της λεγόμενης

«ταυτότητας» της έρευνας. Συνήθως η ταυτότητα της έρευνας δίνεται μέσα από το εξής λεκτικό:

«Η έρευνα έγινε από (ημερομηνία) μέχρι (ημερομηνία) και κάλυψε αστικές και αγροτικές περιοχές. Το μέγεθος του δείγματος ήταν (αριθμός) και χρησιμοποιήθηκε η μέθοδος της τυχαίας στρωματοποιημένης πολυσταδιακής δειγματοληψίας. Η συλλογή των πληροφοριών έγινε τηλεφωνικώς (ή με προσωπικές συνεντεύξεις) με δομημένο ερωτηματολόγιο και στην έρευνα συμμετείχαν άντρες και γυναίκες ηλικίας άνω των 18. Το μέγιστο δειγματοληπτικό σφάλμα είναι τέσσερις (ή τρεις) εκατοστιαίες μονάδες».

Ας με συγχωρήσουν οι φίλοι δειγματολήπτες ή δημοσκόποι, αλλά η ταυτότητα τουλάχιστον όπως παρουσιάζεται δεν είναι καθόλου ταυτότητα. Δύο διαφορετικά άτομα δεν μπορούν να έχουν την ίδια ταυτότητα. Οι δειγματοληπτικές έρευνες, σχεδόν όλες έχουν την ίδια ταυτότητα., αν εξαιρέσει κανείς το χρόνο διεξαγωγής και το μέγεθος δείγματος. Από το πιο πάνω λεκτικό απουσιάζει πληροφόρηση αναφορικά με τους παράγοντες ως προς τους οποίους έγινε η στρωματοποίηση. Απουσιάζει επίσης η πληροφόρηση ως προς τα στάδια της λεγόμενης πολυσταδιακής δειγματοληψίας. Επίσης η πληροφόρηση αναφορικά με το δειγματοληπτικό σφάλμα είναι ελλιπής και ιδιαίτερα παραπλανητική, αφού ως γνωστό αυτό υπολογίζεται από τους δειγματολήπτες με βάση όλο το δείγμα ενώ στην παρουσίαση των αποτελεσμάτων αναφέρονται σε υποπληθυσμούς και σε συμπεράσματα που εξάγονται με βάση μέρος του δείγματος, πολλές φορές με υπερβολικά μικρό μέγεθος. Τέλος απουσιάζει η πληροφόρηση ως προς το ποσοστό άρνησης συμμετοχής στην έρευνα, κάτι που πολλές φορές αποδεικνύεται πολύ σημαντικό. Το τελευταίο δεν είναι ένδειξη διαφάνειας και σίγουρα υπάρχει και θέμα δεοντολογίας.

Γενικά, πόσο δεοντολογικό είναι να δίνεται η ταυτότητα με αυτό τον τρόπο; Είναι δεοντολογικό να χρησιμοποιούνται επιστημονικοί όροι χωρίς περισσότερη πληροφόρηση, όροι, που ακόμη και στους ειδικούς δεν λένε και πολλά πράγματα; Είναι λοιπόν δεοντολογικό, να δίνεται η εντύπωση μιας επιστημονικότητας της έρευνας που μπορεί και να μην υπάρχει;

ABSTRACT

Randomized response techniques first introduced by Warner (1965) and other indirect questioning techniques are methods to collect information on stigmatizing characteristics while at the same time the privacy of participants is protected. The issue of privacy protection is examined in this review paper. In addition, we discuss the issue of the perception of the protection of privacy from the respondents' point of view. Finally we examine the way opinion poll results are presented to the general public.

ΑΝΑΦΟΡΕΣ

- Chaudhuri, A. and Christofides, T.C. (2007): Item count technique in estimating the proportion of people with a sensitive feature. *J. Statist. Plan. Inf.*, **137**, 589-593.
- Chaudhuri, A. and Saha, A. (2005): Optional versus compulsory randomized response techniques in complex surveys. *J. Statist. Plan. Inf.*, **135**, 516-527.
- Christofides, T.C. (2003): A generalized randomized response technique. *Metrika*, **57**, 195-200.
- Greenberg, B.G., Abul-Ela, Abdel-Latif, A., Simmons, W.R., and Horvitz, D.G. (1969): The unrelated question RR model: theoretical framework. *J. Amer. Statist. Assoc.* **64**, 520-539.
- Miller, J.D., Cisin, I.H. and Harrel, A.V: (1986): A new technique for surveying deviant behavior: Item count estimates of marijuana, cocaine and heroin. Paper presented at the Annual Meeting of the American Association for Public Opinion Research, St. Petersburg, Florida
- Warner, S.L. (1965): Randomized Response: a survey technique to eliminate evasive answer bias. *J. Amer. Statist. Assoc.* **60**, 63-69.

εργασίες

στα ελληνικά

Η ΧΡΗΣΗ ΤΩΝ ΟΜΣ ΜΕ ΧΩΡΗΤΙΚΟΤΗΤΕΣ ΣΤΗ ΜΕΛΕΤΗ ΣΥΣΤΗΜΑΤΩΝ ΑΝΑΜΟΝΗΣ. ΕΦΑΡΜΟΓΗ ΣΕ ΣΥΣΤΗΜΑΤΑ ΑΝΑΜΟΝΗΣ ΜΕ ΠΕΠΕΡΑΣΜΕΝΗ ΠΗΓΗ

Γεώργιος Βασιλειάδης

Τμήμα Πληροφορικής και Τεχνολογίας Υπολογιστών,
Τ.Ε.Ι. Δυτικής Μακεδονίας, Παράρτημα Καστοριάς
gvasil@math.auth.gr

ΠΕΡΙΛΗΨΗ

Στην παρούσα εργασία περιγράφεται αρχικά το ομογενές Μαρκοβιανό σύστημα διακριτού χρόνου με χωρητικότητες στις καταστάσεις του (ΟΜΣ/ c) και προτεραιότητες στις μετακινήσεις. Πρόκειται για ένα ομογενές Μαρκοβιανό σύστημα του οποίου κάθε κατάσταση i του χώρου καταστάσεων $S = \{1, 2, \dots, k\}$ παρουσιάζει μία πεπερασμένη χωρητικότητα $c_i \in \mathbb{N}$, $i = 1, 2, \dots, k$. Επίσης θεωρείται ότι οι μετακινήσεις των μελών του συστήματος σε οποιαδήποτε χρονική στιγμή t δεν πραγματοποιούνται ταυτόχρονα από όλες τις καταστάσεις αλλά σύμφωνα με κάποια σειρά προτεραιότητας. Έτσι στην περίπτωση που ένα μέλος μετακινηθεί προς μία κατάσταση που έχει συμπληρώσει τη χωρητικότητά της, το μέλος αυτό εγκαταλείπει το σύστημα. Αφού υπολογιστεί ο πίνακας μετάβασης του μοντέλου, εξηγείται πώς αυτό το μοντέλο μπορεί να χρησιμοποιηθεί για τη μελέτη ενός συστήματος αναμονής διακριτού χρόνου με πεπερασμένη πηγή.

Λέξεις Κλειδιά: Μαρκοβιανά μοντέλα διακριτού χρόνου, ΟΜΣ με χωρητικότητες, Ουρές διακριτού χρόνου.

1. ΕΙΣΑΓΩΓΗ

Θεωρούμε ένα σύστημα του οποίου τα μέλη μπορούν να ταξινομηθούν σε k καταστάσεις σύμφωνα με κάποια χαρακτηριστικά τους. Τα μέλη του συστήματος έχουν τη δυνατότητα μετακίνησης από κατάσταση σε κατάσταση σύμφωνα με τη Μαρκοβιανή ιδιότητα. Οι μετακινήσεις γίνονται σε χρόνο διακριτό και σε κάθε χρονική στιγμή τα μέλη του συστήματος βρίσκονται αποκλειστικά σε μία μόνο κατάσταση. Οι πιθανότητες μετάβασης από μία κατάσταση σε άλλη είναι σταθερές (δεν εξαρτώνται από το χρόνο). Επίσης, θεωρούμε ότι το πλήθος των μελών του συστήματος είναι σταθερό, δηλαδή δεν υπάρχει η δυνατότητα εισόδου νέων μελών ούτε επιτρέπεται η έξοδος στους ήδη υπάρχοντες.

Ένα σύστημα με τα παραπάνω χαρακτηριστικά ονομάζεται κλειστό ομογενές Μαρκοβιανό σύστημα (ΟΜΣ) διακριτού χρόνου (Bartholomew, 1982) και για τη μελέτη του

χρησιμοποιούμε τους παρακάτω συμβολισμούς:

- $t = 1, 2, \dots$, η παράμετρος του χρόνου,
- $S = \{1, 2, \dots, k\}$, ο χώρος καταστάσεων,
- $\mathbf{P} = (p_{ij})$, $i, j = 1, 2, \dots, k$, ο πίνακας πιθανοτήτων μετάβασης, όπου p_{ij} η πιθανότητα μετάβασης από την κατάσταση i στη j σ' ένα βήμα,
- $\mathbf{n}(t) = (n_1(t), n_2(t), \dots, n_k(t))^T$, το διάνυσμα κατάστασης του συστήματος, όπου $n_i(t)$, $i = 1, 2, \dots, k$, το πλήθος των μελών που βρίσκονται στην i κατάσταση τη χρονική στιγμή t ,
- N , το πλήθος των μελών του συστήματος.

Για τις βασικές έννοιες στις Μαρκοβιανές αλυσίδες και τη θεμελίωση του ΟΜΣ αναφέρουμε ενδεικτικά τα Bartholomew (1982), Isaacson and Madsen (1976), Vassiliou (1997). Όσον αφορά στις εφαρμογές των Μαρκοβιανών συστημάτων, τα συστήματα αυτά εμφανίζονται στη μελέτη συστημάτων ανθρώπινου δυναμικού (Bartholomew, 1982, Gani, 1963, McClean et al., 1998, Vassiliou, 1997), στη δημογραφία (Bartholomew, 1982), στη βιολογία (Patoucheas and Stamou, 1993).

Γενίκευση του κλασικού ΟΜΣ διακριτού χρόνου αποτελεί ένα σύστημα στο οποίο κάθε κατάσταση i του χώρου καταστάσεων S παρουσιάζει πεπερασμένη χωρητικότητα $c_i \in \mathbb{N}$ (ΟΜΣ/ c). Αναλυτική περιγραφή και μελέτη συστημάτων με αυτό το χαρακτηριστικό παρουσιάζονται στις εργασίες Βασιλειάδης και Τσακλίδης (2010), Vasiliadis and Tsaklidis (2011).

Το γεγονός ότι κάθε κατάσταση i του χώρου καταστάσεων S ενός ΟΜΣ διακριτού χρόνου παρουσιάζει πεπερασμένη χωρητικότητα $c_i \in \mathbb{N}$, έχει ως αποτέλεσμα το μέγεθος οποιασδήποτε κατάστασης i να μην μπορεί να υπερβεί την τιμή c_i , δηλαδή

$$n_i(t) \leq c_i, \quad i \in S,$$

για κάθε χρονική στιγμή t , $t = 1, 2, \dots$

Όλα τα μέλη του συστήματος αποφασίζουν σε κάθε χρονική στιγμή t ταυτόχρονα σε ποια κατάσταση θα μετακινηθούν. Οι αποφάσεις αυτές λαμβάνονται σύμφωνα με τον πίνακα μετάβασης \mathbf{P} τον οποίο στο εξής θα ονομάζουμε *πίνακα απόφασης*. Αν το πλήθος των μελών του συστήματος που αποφασίσουν να μετακινηθούν προς μία κατάσταση i , $i \in S$, κάποια χρονική στιγμή t , δεν υπερβαίνει τη χωρητικότητα c_i , τότε όλα τα μέλη εισέρχονται σ' αυτήν, και το πλήθος τους αποτελεί το μέγεθος της κατάστασης για το χρονικό διάστημα $(t, t + 1]$. Αν όμως, σε κάποια χρονική στιγμή t , το πλήθος αυτό υπερβεί τη χωρητικότητα, τότε στην κατάσταση i εισέρχονται μόνο c_i από αυτά τα μέλη και λέμε ότι τη δεδομένη χρονική στιγμή το σύστημα παρουσιάζει υπερχειλίση. Θεωρούμε ότι τα μέλη του συστήματος που υπερχειλίζουν από κάθε κατάσταση εισέρχονται σε μία κατάσταση (κατάσταση υπερχειλίσης) την οποία συμβολίζουμε με $k + 1$. Τα μέλη που θα βρεθούν σε αυτή την κατάσταση λόγω του φαινομένου της υπερχειλίσης έχουν τη δυνατότητα να επανέλθουν στο σύστημα κατά την επόμενη μετακίνηση σύμφωνα με γνωστές πιθανότητες επανεισόδου.

Έτσι, με την παραπάνω θεώρηση ορίζουμε ένα νέο σύστημα με χώρο καταστάσεων $\tilde{S} = \{1, 2, \dots, k + 1\}$. Θα συμβολίζουμε το νέο, επεκτεταμένο, σύστημα με ΟΜΣ/ cp .

Ο πίνακας απόφασης αυτού του συστήματος είναι της μορφής

$$\tilde{\mathbf{P}} = \begin{pmatrix} \mathbf{P} & \mathbf{0} \\ \mathbf{p}^T & 0 \end{pmatrix},$$

όπου $\mathbf{p} = (p_{k+1,1}, p_{k+1,2}, \dots, p_{k+1,k})^T$, το διάνυσμα που περιέχει τις πιθανότητες με τις οποίες τα υπερχειλιζόντα μέλη αποφασίζουν σε ποια από τις k καταστάσεις θα μετακινηθούν.

Η πληθυσμιακή δομή αυτού του συστήματος τη χρονική στιγμή t δίνεται από το διάνυσμα κατάστασης

$$\tilde{\mathbf{n}}(t) = (n_1(t), n_2(t), \dots, n_k(t), n_{k+1}(t))^T.$$

Είναι φανερό ότι η συμπεριφορά του ΟΜΣ/ cp επηρεάζεται από τον τρόπο με τον οποίο πραγματοποιούνται οι μετακινήσεις των μελών. Στην περίπτωση του ΟΜΣ/ c (Vasiliadis and Tsaklidis, 2011) θεωρούμε ότι οι μετακινήσεις των μελών σε κάθε χρονική στιγμή t πραγματοποιούνται ταυτόχρονα από όλες τις καταστάσεις χωρίς να έχει οριστεί κάποια σειρά προτεραιότητας, και ότι τα μέλη που θα υπερχειλίσουν επιλέγονται τυχαία, χωρίς να υπάρχει κάποια εξάρτηση από την κατάσταση από την οποία έχει πραγματοποιηθεί η μετακίνηση. Γενικότερα, όμως, ενδέχεται οι μετακινήσεις των μελών να πραγματοποιούνται με κάποια σειρά προτεραιότητας, οπότε τα μέλη που θα υπερχειλίσουν είναι αυτά που είναι χαμηλότερα στη σειρά προτεραιότητας (Βασιλειάδης και Τσακλίδης, 2010). Έτσι, για να μπορέσουμε να μελετήσουμε τη συμπεριφορά ενός ΟΜΣ με χωρητικότητες θα πρέπει να καθοριστούν αρχικά οι προτεραιότητες με βάση τις οποίες συντελούνται οι μετακινήσεις των μελών στις καταστάσεις. Στην παρούσα εργασία θεωρούμε ότι τα μέλη που αποφασίζουν να αλλάξουν κατάσταση μετακινούνται με μία σειρά προτεραιότητας ανάλογα με την κατάσταση στην οποία βρίσκονται ξεκινώντας με αυτούς που βρίσκονται στην κατάσταση υπερχειλίσης και συνεχίζοντας με φθίνουσα σειρά. Στην περίπτωση που από την ίδια κατάσταση αποφασίσει ένας αριθμός μελών να μετακινηθεί προς μία κατάσταση της οποίας οι κενές θέσεις δεν επαρκούν, τότε οι κενές θέσεις συμπληρώνονται με τυχαία επιλογή.

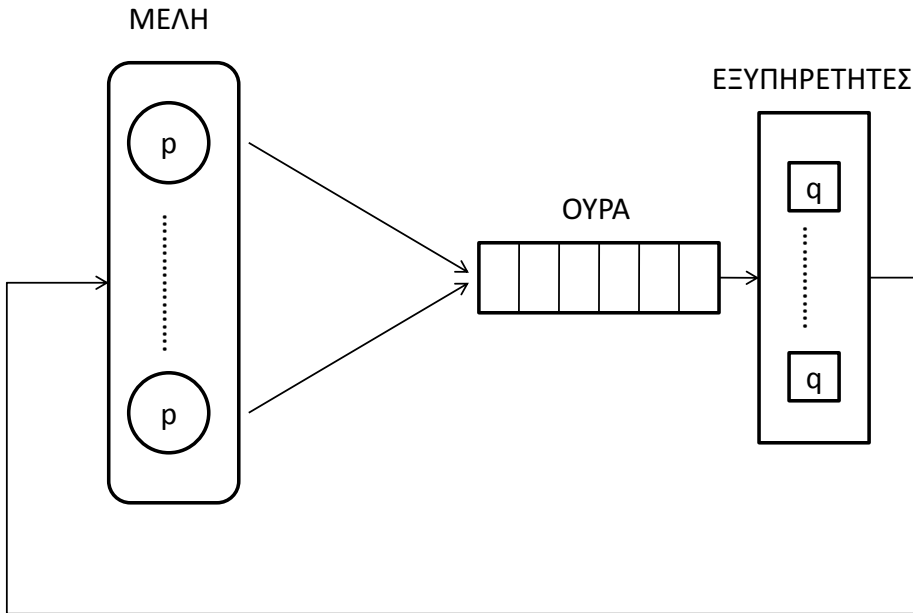
Το γενικό αυτό μοντέλο, εκτός των άλλων εφαρμογών, μπορεί να χρησιμοποιηθεί και για τη μελέτη συστημάτων ανομονής. Στη συνέχεια παρουσιάζεται ένα σύστημα ανομονής διακριτού χρόνου με πεπερασμένη πηγή το οποίο μπορεί να αναπαρασταθεί σαν ένα ΟΜΣ/ cp .

2. Η ΟΥΡΑ ΔΙΑΚΡΙΤΟΥ ΧΡΟΝΟΥ ΜΕ ΠΕΠΕΡΑΣΜΕΝΗ ΠΗΓΗ

Θεωρούμε ένα κλειστό σύστημα το οποίο αποτελείται από N μέλη, $N \in \mathbb{N}$, και m εξυπηρετητές, $m \in \mathbb{N}$, $m < N$. Υποθέτουμε ότι σε χρόνο διακριτό κάθε μέλος του συστήματος (το οποίο δεν βρίσκεται σε διαδικασία εξυπηρέτησης) έχει την ίδια πιθανότητα p να ζητήσει να εξυπηρετηθεί από κάποιον διαθέσιμο εξυπηρετητή. Στην περίπτωση που υπάρχει διαθέσιμος εξυπηρετητής, η εξυπηρέτηση πραγματοποιείται σύμφωνα με τη γεωμετρική κατανομή με παράμετρο q . Αν σε κάποια χρονική στιγμή

ένα μέλος ζητήσει εξυπηρέτηση και όλοι οι εξυπηρετητές είναι απασχολημένοι τότε το μέλος αυτό εισέρχεται σε μία ουρά. Θεωρούμε ότι οι αφίξεις στην εξυπηρέτηση του συστήματος καθώς επίσης και οι αναχωρήσεις απ' αυτήν πραγματοποιούνται ταυτόχρονα σε διακριτές χρονικές στιγμές $t, t = 1, 2, \dots$. Επίσης, θεωρούμε ότι δεν υπάρχει κάποια σειρά στην εξυπηρέτηση των μελών που βρίσκονται στην ουρά του συστήματος και αυτή γίνεται με τυχαία επιλογή. Το σύστημα που περιγράψαμε είναι ένα σύστημα αναμονής διακριτού χρόνου με πεπερασμένη πηγή (Σχήμα 1).

Σχήμα 1: Κλειστό σύστημα αναμονής.



Η μελέτη αυτού του κλειστού συστήματος αναμονής μπορεί να γίνει με τη βοήθεια ενός ΟΜΣ/ cp . Θεωρούμε ένα ΟΜΣ διακριτού χρόνου το οποίο αποτελείται από N μέλη και δύο καταστάσεις. Η πρώτη κατάσταση αντιπροσωπεύει την πηγή του συστήματος και τη χρονική $t = 0$ περιέχει τα N μέλη. Η δεύτερη κατάσταση αντιπροσωπεύει τους εξυπηρετητές του συστήματος. Αφού το πλήθος των εξυπηρετητών είναι m το μέγεθος αυτής της κατάστασης δεν μπορεί να υπερβαίνει την τιμή m . Έτσι η δεύτερη κατάσταση παρουσιάζει μία χωρητικότητα $c_2 = m$. Όσον αφορά τη χωρητικότητα της πρώτης κατάστασης δεν υπάρχει κάποιος περιορισμός επομένως μπορούμε να θεωρήσουμε ότι $c_1 = N$.

Η παρουσία της χωρητικότητας στη δεύτερη κατάσταση έχει σαν αποτέλεσμα τη θεώρηση του επεκταμένου συστήματος (ΟΜΣ/ cp) το οποίο περιέχει και μία τρίτη κατάσταση στην οποία εισέρχονται τα μέλη που υπερχειλίζουν από τη δεύτερη κατάσταση. Η κατάσταση αυτή αντιπροσωπεύει την ουρά του συστήματος. Είναι προφανές ότι τα μέλη του συστήματος που σε κάποια χρονική στιγμή t βρεθούν στην ουρά, την

επόμενη χρονική στιγμή $t + 1$ θα ζητήσουν να εξυπηρετηθούν (θα αποφασίσουν να μετακινηθούν στην κατάσταση 2) με πιθανότητα 1. Επίσης τα μέλη αυτά θα προηγούνται στην εξυπηρέτηση (αν υπάρχει διαθέσιμος εξυπηρετητής) σε σχέση με τα νέα μέλη που ζητούν εξυπηρέτηση την ίδια χρονική στιγμή. Έτσι, οι μετακινήσεις των μελών του συστήματος γίνονται σύμφωνα με τη σειρά προτεραιότητας που ορίστηκε στην προηγούμενη ενότητα και σύμφωνα με τον πίνακα απόφασης

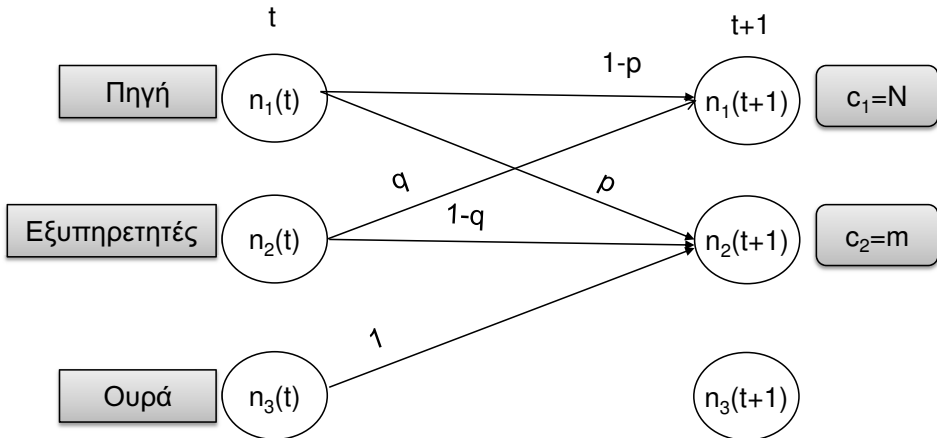
$$\tilde{\mathbf{P}} = \begin{pmatrix} 1-p & p & 0 \\ q & 1-q & 0 \\ 0 & 1 & 0 \end{pmatrix}.$$

Η πληθυσμιακή δομή του συστήματος τη χρονική στιγμή t δίνεται από το διάνυσμα κατάστασης

$$\tilde{\mathbf{n}}(t) = (n_1(t), n_2(t), n_3(t))^T,$$

όπου $n_i(t)$, $i = 1, 2, 3$, το πλήθος των μελών που βρίσκονται στην i κατάσταση του συστήματος τη χρονική στιγμή t . Παρουσιάζουμε στη συνέχεια (Σχήμα 2) γραφικά τη μορφή του συστήματος για δύο διαδοχικές χρονικές στιγμές.

Σχήμα 2: ΟΜΣ/ερ διακριτού χρόνου.



3. Ο ΠΙΝΑΚΑΣ ΜΕΤΑΒΑΣΗΣ ΤΟΥ ΟΜΣ/ερ

Στο ΟΜΣ οι πιθανότητες μετάβασης είναι ανεξάρτητες του χρόνου στον οποίο πραγματοποιείται η μετάβαση. Στην περίπτωση όμως του ΟΜΣ/ερ, που χρησιμοποιήσαμε για να περιγράψουμε το σύστημα αναμονής με πεπερασμένη πηγή, αφού υπάρχει περιορισμός στην χωρητικότητα των καταστάσεων, ένα μέλος που αποφασίζει να μετακινηθεί από την κατάσταση i στην j , $i, j \in \tilde{S}$, τη χρονική στιγμή t , είτε παραμένει στην j είτε υπερχειλίζει και εισέρχεται στην ουρά του συστήματος. Έτσι, η πιθανότητα ένα μέλος του συστήματος να μετακινηθεί από την κατάσταση i στην j τη χρονική

στιγμή t , εξαρτάται από τη χωρητικότητα c_j , την κατάσταση i και από το συνολικό αριθμό μελών του συστήματος που αποφάσισαν να μετακινηθούν από όλες τις καταστάσεις του συστήματος προς την j την ίδια χρονική στιγμή t . Επομένως οι πιθανότητες μετάβασης για το ΟΜΣ/ cp είναι εξαρτημένες από το διάνυσμα κατάστασης $\tilde{\mathbf{n}}(t)$.

Συμβολίζουμε με $\tilde{p}_{ij}(\tilde{\mathbf{n}}(t))$, $i, j = 1, 2, 3$, την πιθανότητα ένα μέλος του συστήματος αναμονής να μεταβεί από την κατάσταση i στην κατάσταση j τη χρονική στιγμή t και με $\tilde{\mathbf{P}}(\tilde{\mathbf{n}}(t)) = (\tilde{p}_{ij}(\tilde{\mathbf{n}}(t)))$ τον αντίστοιχο 3×3 πίνακα μετάβασης. Ο προσδιορισμός του πίνακα μετάβασης μπορεί να γίνει ακολουθώντας τη μεθοδολογία που παρουσιάζεται στην εργασία Βασιλειάδης και Τσακλίδης (2010), λαμβάνοντας υπόψη τις προτεραιότητες στις μετακινήσεις που έχουν οριστεί.

Έτσι, αν συμβολίσουμε με

$$p(b_{ij}; \tilde{\mathbf{n}}(t)) = P(\text{ένα μέλος εισέρχεται στην } j \text{ τη χρονική στιγμή } t \mid \text{το μέλος αυτό βρισκόταν στην κατάσταση } i \text{ τη χρονική στιγμή } t \text{ και αποφάσισε να μετακινηθεί στην } j),$$

και

$$p(\bar{b}_{ij}; \tilde{\mathbf{n}}(t)) = P(\text{ένα μέλος υπερχειλίζει από την } j \text{ κατά τη χρονική στιγμή } t \mid \text{το μέλος αυτό βρισκόταν στην κατάσταση } i \text{ τη χρονική στιγμή } t \text{ και αποφάσισε να μετακινηθεί στην } j),$$

$i = 1, 2, 3, j = 1, 2$, τότε ο πίνακας μετάβασης του συστήματος αναμονής για μία δεδομένη δομή του διανύσματος κατάστασης $\tilde{\mathbf{n}}(t)$ μπορεί να γραφεί στη μορφή (Βασιλειάδης και Τσακλίδης, 2010)

$$\tilde{\mathbf{P}}(\tilde{\mathbf{n}}(t)) = \begin{pmatrix} \mathbf{P} \Delta \mathbf{H}(\tilde{\mathbf{n}}(t)) & \mathbf{P} \Delta \bar{\mathbf{H}}(\tilde{\mathbf{n}}(t)) \mathbf{e} \\ \mathbf{p}^T p(b_{32}; \tilde{\mathbf{n}}(t)) & p(\bar{b}_{32}; \tilde{\mathbf{n}}(t)) \end{pmatrix}, \quad (1)$$

όπου

- $\mathbf{H}(\tilde{\mathbf{n}}(t))$, 2×2 πίνακας με στοιχεία τις πιθανότητες $p(b_{ij}; \tilde{\mathbf{n}}(t))$, $i, j = 1, 2$,
- $\bar{\mathbf{H}}(\tilde{\mathbf{n}}(t))$, 2×2 πίνακας με στοιχεία τις πιθανότητες $p(\bar{b}_{ij}; \tilde{\mathbf{n}}(t))$, $i, j = 1, 2$,
- \mathbf{e} διάνυσμα στήλη με όλα τα στοιχεία του μονάδες,
- $\mathbf{P} = \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix}$,
- $\mathbf{p} = (0, 1)^T$,
- " Δ " το σύμβολο του Hadamard γινομένου.

Η εύρεση του πίνακα μετάβασης $\tilde{\mathbf{P}}(\tilde{\mathbf{n}}(t))$ που περιέχει τις πιθανότητες μετάβασης για μία δεδομένη δομή του συστήματος $\tilde{\mathbf{n}}(t)$, καθιστά δυνατή τη μελέτη της συμπεριφοράς του συστήματος αναμονής με πεπερασμένη πηγή όχι μόνο ασυμπτωτικά αλλά και στην παροδική φάση. Η μελέτη αυτή μπορεί να επιτευχθεί με χρήση των παραγοντικών ροπών των μεγεθών των καταστάσεων κατ' αναλογία με το ΟΜΣ/ c (Vasiliadis and Tsaklidis, 2011).

ABSTRACT

In this paper we consider the discrete time HMS with finite state capacities $c_i \in \mathbb{N}$, $i = 1, 2, \dots, k$, for any state i of the state space $S = \{1, 2, \dots, k\}$. This system is a generalization of the classical HMS, where the states' capacities are infinite. Also, we assume that movements of the system's members occur according to some priority order. So, if a member which is in state i decides to move to state j with capacity c_j then it either enters state j or overflows from the state. Next, the HMS with finite state capacities and transition priorities is used in order to examine the behavior of a finite source discrete-time queuing system.

REFERENCES

- Bartholomew, D.J. (1982). *Stochastic Models for Social Processes*, 3rd edn., New York: John Wiley.
- Βασιλειάδης, Γ. και Τσακλίδης, Γ. (2010). Το ομογενές Μαρκοβιανό σύστημα διακριτού χρόνου με χωρητικότητες στις καταστάσεις και προτεραιότητες στις μετακινήσεις. *Πρακτικά 23ου Πανελληνίου Συνεδρίου Στατιστικής*.
- Gani, J. (1963). Formulae for projecting enrolments and degrees awarded in universities. *J. R. Statist. Soc. A*. **126**, 400-409.
- Isaacson, D.L. and Madsen, R.W. (1976). *Markov Chains. Theory and applications*, New York: John Wiley.
- McClellan, S.I., McAlea, B. and Millard, P. (1998). Using a Markov reward model to estimate spend-down costs for a geriatric department. *J. Operat. Res. Soc.* **10**, 1021-1025.
- Patoucheas, P.D. and Stamou, G. (1993). Non-homogeneous Markovian models in ecological modelling: a study of the zoobenthos dynamics in Thermaikos Gulf, Greece. *Ecological Modelling* **66**, 197-215.
- Vasiliadis, G. and Tsaklidis, G. (2011). On the distributions of the state sizes of the closed discrete-time Homogeneous Markov System with finite state capacities (HMS/c). *Markov Processes and Related Fields* **17**, 91-118.
- Vassiliou, P.-C. G. (1997). The evolution of the theory of non-homogeneous Markov systems. *Appl. Stochastic Models Data Anal.* **13**, no. **3-4**, 159-176.



Βελτίωση της εκτίμησης των παραμέτρων ενός γενικευμένου γραμμικού μοντέλου με τη χρήση της παλινδρόμησης Ridge

Β.Γ. Βασιλειάδης, Ε.Ν. Σπανού, Α.Γ. Ρήγας

Δημοκρίτειο Πανεπιστήμιο Θράκης
{bvasil, ispanou, rigas}@ee.duth.gr

ΠΕΡΙΛΗΨΗ

Η μοντελοποίηση ενός νευροφυσιολογικού συστήματος και συγκεκριμένα της μυϊκής ατράκτου πραγματοποιείται, προκειμένου να μελετηθεί η συμπεριφορά της στην περίπτωση που έχουμε μία είσοδο και δύο εξόδους. Η αναγνώριση του συστήματος υλοποιείται στο πεδίο του χρόνου, όπου τόσο η είσοδος όσο και οι εξόδοι του συστήματος μπορούν να θεωρηθούν ως δυαδικές χρονοσειρές και επομένως, είναι δυνατή η εφαρμογή της θεωρίας των γενικευμένων γραμμικών μοντέλων. Στη συγκεκριμένη εργασία εφαρμόζεται η μέθοδος της ποινικοποιημένης συνάρτησης πιθανοφάνειας για την εκτίμηση των παραμέτρων του συστήματος και συγκεκριμένα η μέθοδος της παλινδρόμησης Ridge, η οποία δίνει καλύτερες εκτιμήσεις. Το στοχαστικό μοντέλο το οποίο προτείνεται για την περιγραφή του συστήματος περιέχει το κατώφλι και την αθροιστική συνάρτηση, όπου η εκτίμηση της τελευταίας είναι σημαντική καθώς περιγράφει την επίδραση της εισόδου στο σύστημα. Επίσης, παρουσιάζεται ένας έλεγχος για την εγκυρότητα του προσαρμοσμένου μοντέλου που βασίζεται στα τυχοποιημένα ποσοστιαία υπόλοιπα που ακολουθούν κανονική κατανομή. Για τον έλεγχο χρησιμοποιείται η παράσταση Q-Q στην οποία έχουν προστεθεί 5% απορριπτικές περιοχές. Τα παραπάνω επεξηγούνται με ένα παράδειγμα από το πεδίο της Νευροφυσιολογίας στο οποίο μελετάται η συμπεριφορά του συστήματος της μυϊκής ατράκτου υπό την επίδραση ενός γάμμα κινητονεύρου.

Keywords: λογιστική παλινδρόμηση, παλινδρόμηση Ridge, τυχοποιημένα ποσοστιαία υπόλοιπα, γράφημα Q-Q, μυϊκή ατράκτος

1. ΕΙΣΑΓΩΓΗ

Η μοντελοποίηση ενός νευροφυσιολογικού συστήματος αφορά στη μελέτη των αλληλεπιδράσεων που παρατηρούνται μεταξύ των διαφόρων διεργασιών του και επιτυγχάνεται με τη βοήθεια γενικευμένων γραμμικών μοντέλων και εφαρμογή τεχνικών μέγιστης πιθανοφάνειας. Η λογιστική παλινδρόμηση αποτελεί την πιο ευρέως διαδεδομένη μέθοδο μοντελοποίησης δυαδικών δεδομένων, ωστόσο, η ύπαρξη συσχέτισης μεταξύ των ανεξάρτητων μεταβλητών ή το μεγάλο πλήθος αυτών οδηγεί σε ανακριβή και αναξιόπιστα αποτελέσματα. Το πρόβλημα αυτό μπορεί να αντιμετωπισθεί

χρησιμοποιώντας την παλινδρόμηση Ridge, η οποία προτάθηκε από τον Hoerl (1962).

Η μέθοδος αυτή αποτελεί μία εναλλακτική μέθοδο εκτίμησης των αγνώστων παραμέτρων του συστήματος και ανήκει στο σύνολο των μεροληπτικών τεχνικών παλινδρόμησης. Σε γενικές γραμμές, η μέθοδος αυτή διαφοροποιείται από την αντίστοιχη της μέγιστης πιθανοφάνειας στο γεγονός ότι προστίθεται ένας ποινικοποιημένος όρος (penalty term) και μία παράμετρος ομαλοποίησης (regularization parameter).

Μετά την προσαρμογή ενός γραμμικού παλινδρομικού μοντέλου στα δεδομένα, είναι σύνηθες να τυποποιηθούν τα υπόλοιπα, που ακολουθούν κανονική κατανομή και να χρησιμοποιηθούν για τον έλεγχο της εγκυρότητας του μοντέλου. Στην περίπτωση των γενικευμένων γραμμικών μοντέλων, τα συνηθισμένα υπόλοιπα δεν έχουν πρακτική χρησιμότητα, αφού τις περισσότερες φορές δεν ακολουθούν την κανονική κατανομή και η διασπορά τους δεν είναι σταθερή.

2. ΠΑΛΙΝΔΡΟΜΗΣΗ RIDGE

Έστω y_i , $i = 1, \dots, n$, οι δυαδικές αποκρίσεις n τυχαίων μεταβλητών Y_i , όπου $Y_i \sim B(1, \pi_i)$ και \mathbf{x}_i είναι το διάνυσμα των ανεξάρτητων μεταβλητών. Τότε το μοντέλο λογιστικής παλινδρόμησης δίνεται ως εξής

$$\pi_i = \{1 + \exp(-\boldsymbol{\beta}\mathbf{x}_i)\}^{-1}, \quad (1)$$

όπου $\boldsymbol{\beta}$ το διάνυσμα των παραμέτρων (McCullagh and Nelder, 1989). Οι εκτιμητές μέγιστης πιθανοφάνειας των παραμέτρων β_j , $j = 1, \dots, k$, και κατά συνέπεια των πιθανοτήτων π_i , επιτυγχάνονται μεγιστοποιώντας το λογάριθμο της συνάρτησης πιθανοφάνειας $l(\boldsymbol{\beta}|\mathbf{y}) = \log L(\boldsymbol{\beta}|\mathbf{y})$,

$$l(\boldsymbol{\beta}|\mathbf{y}) = \sum_{i=1}^n \left\{ y_i \log \left[\frac{1}{1 + \exp(-\mathbf{x}_i^T \boldsymbol{\beta})} \right] + (1 - y_i) \log \left[1 - \frac{1}{1 + \exp(-\mathbf{x}_i^T \boldsymbol{\beta})} \right] \right\}. \quad (2)$$

Χρησιμοποιώντας τον αλγόριθμο Newton-Raphson ή τη μέθοδο scoring, οι εκτιμητές των παραμέτρων β_j είναι οι λύσεις των εξισώσεων score

$$\frac{\partial l(\boldsymbol{\beta}|\mathbf{y})}{\partial \beta_j} \equiv U(\beta_j) = \sum_{i=1}^n (y_i - \pi_i) x_{ij} = 0 \quad (j = 1, \dots, k). \quad (3)$$

Στην παλινδρόμηση Ridge χρησιμοποιείται η ποινικοποιημένη συνάρτηση μέγιστης πιθανοφάνειας και η οποία δίδεται από τη σχέση (LeCessie and Van Houwelingen, 1992):

$$l^\lambda(\boldsymbol{\beta}|\mathbf{y}) = l(\boldsymbol{\beta}|\mathbf{y}) - \lambda \|\boldsymbol{\beta}\|^2 = l(\boldsymbol{\beta}|\mathbf{y}) - \lambda R, \quad (4)$$

όπου R είναι ο ποινικοποιημένος όρος ο οποίος δίδεται από τη σχέση (Brillinger et al., 2009):

$$R = \sum_{j=0}^{k-1} (\beta_{j+1} - \beta_j)^2. \quad (5)$$

Η διαδικασία υπολογισμού των ποινικοποιημένων εκτιμητών $\hat{\beta}^\lambda$ μπορεί να υλοποιηθεί με τη βοήθεια του αλγορίθμου Newton-Raphson και απαιτείται ο μετασχηματισμός του γραμμικού εκτιμητή του λογιστικού μοντέλου, προκειμένου ο ποινικοποιημένος όρος R να έρχεται σε συμφωνία με τη σχέση (4). Επομένως

$$\beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik} = \beta_1 z_{i1} + (\beta_2 - \beta_1) z_{i2} + \dots + (\beta_k - \beta_{k-1}) z_{ik} \quad (6)$$

$$= \gamma_1 z_{i1} + \gamma_2 z_{i2} + \dots + \gamma_k z_{ik},$$

όπου $\gamma_1 = \beta_1$, $\gamma_j = \beta_j - \beta_{j-1}$, $j = 2, \dots, k$ και $z_{ij} = \sum_{u=j}^k x_{iu}$. Άρα, η νέα ποινικοποιημένη συνάρτηση μέγιστης πιθανοφάνειας θα δίδεται από τη σχέση:

$$l^\lambda(\gamma|\mathbf{y}) = l(\gamma|\mathbf{y}) - \lambda \|\gamma\|^2. \quad (7)$$

Η πρώτη παράγωγος της σχέσης (7) θα ισούται με:

$$U^\lambda(\gamma) = \sum_{i=1}^n z_i \{y_i - \pi_i\} - 2\lambda\gamma = U(\gamma) - 2\lambda\gamma. \quad (8)$$

Υπολογίζοντας την αρνητική δεύτερη παράγωγο προκύπτει:

$$\Omega^\lambda(\gamma) = \Omega(\gamma) + 2\lambda I, \quad (9)$$

όπου $\Omega(\gamma) = \mathbf{z}^T \mathbf{W} \mathbf{z}$ και \mathbf{W} είναι ένας $n \times n$ διαγώνιος πίνακας με στοιχεία $W_{ii} = \pi_i(1 - \pi_i)$. Χρησιμοποιώντας τη σειρά Taylor στην πρώτη παράγωγο μπορούν να επιτευχθούν οι ιδιότητες που ισχύουν για μεγάλα δείγματα στην ποινικοποιημένη συνάρτηση μέγιστης πιθανοφάνειας. Επομένως, θα ισχύει:

$$U^\lambda(\hat{\gamma}^\lambda) = U^\lambda(\gamma_0) - (\hat{\gamma}^\lambda - \gamma_0) \Omega^\lambda(\gamma_0) + o(\|\hat{\gamma}^\lambda - \gamma_0\|). \quad (10)$$

Χρησιμοποιώντας τις σχέσεις (8) και (9) και θέτοντας τη σχέση (10) ίση με 0 προκύπτει:

$$\hat{\gamma}^\lambda = \{\Omega(\gamma) + 2\lambda I\}^{-1} \{U(\gamma_0) + \gamma_0 \Omega(\gamma_0)\}. \quad (11)$$

Η παράμετρος ομαλοποίησης λ είναι ένας θετικός αριθμός και ουσιαστικά ο ρόλος της είναι η μετατροπή της σημαντικότητας των ποινικοποιημένων όρων. Όταν η παράμετρος $\lambda = 0$, τότε προκύπτουν οι ίδιοι εκτιμητές με αυτούς της μέγιστης πιθανοφάνειας. Η βέλτιστη παράμετρος λ επιλέγεται από ένα σύνολο διαφόρων τιμών ως εκείνη που δίνει το ελάχιστο μέσο σφάλμα και το οποίο υπολογίζεται με βάση τα 2 παρακάτω κριτήρια (LeCessie and Van Houwelingen, 1992):

1. Προσεγγιστικό κριτήριο διεπικύρωσης

$$MSE_{ACV} = n^{-1} \sum_{i=1}^n \frac{(y_i - \hat{\pi}_i)^2}{(1 - h_{ii})^2}. \quad (12)$$

2. Γενικευμένο κριτήριο διεπικύρωσης

$$MSE_{GCV} = n^{-1} \frac{\sum_{i=1}^n (y_i - \hat{\pi}_i)^2}{(1 - n^{-1} \sum_{i=1}^n h_{ii})^2}, \quad (13)$$

όπου $h_{ii} = W_{ii} z_i \{ \mathbf{\Omega}(\boldsymbol{\gamma}) + 2\lambda \mathbf{I} \}^{-1} z_i^T$.

Προκειμένου να υπολογιστούν τα διαστήματα εμπιστοσύνης πρέπει να ληφθεί υπόψη η μεροληψία που παρατηρείται στην εκτιμώμενη παράμετρο. Γι' αυτό το λόγο, εφαρμόζεται η μέθοδος sandwich, ώστε να υπολογισθούν συνεπείς εκτιμητές του πίνακα συνδιακυμάνσεων καθώς αποτελεί μία γενικευμένη τεχνική υπολογισμού του συγκεκριμένου πίνακα των εκτιμώμενων αγνώστων παραμέτρων. Το σημαντικό χαρακτηριστικό της είναι ότι παράγει ασυμπτωτικά συνεπείς εκτιμητές του πίνακα αυτού χωρίς να γίνονται υποθέσεις που αφορούν στην κατανομή (Kauermann and Carroll, 2001).

Η γενική μορφή του εκτιμητή sandwich της διασποράς υπολογίζεται από τη σχέση:

$$\hat{\mathbf{V}}_s = \hat{\mathbf{V}}_H^{-1} \hat{\mathbf{B}}_s \hat{\mathbf{V}}_H^{-1}, \quad (14)$$

όπου $\hat{\mathbf{V}}_H^{-1}$ είναι ο Hessian πίνακας και υπολογίζεται από τη σχέση:

$$\hat{\mathbf{V}}_H^{-1} = E \left(- \frac{\partial^2 l}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}^T} \right). \quad (15)$$

Στη συγκεκριμένη εργασία χρησιμοποιείται ο πίνακας πληροφορίας με αποτέλεσμα η σχέση (15) να μετατρέπεται στην ακόλουθη μορφή (Hardin and Hilbe, 2007):

$$\left[\hat{\mathbf{V}}_H \right]_{jk} = \sum_{i=1}^n \frac{1}{\alpha(\phi)} \frac{1}{V(\pi_i)} \left(\frac{\partial \pi_i}{\partial \eta_i} \right)^2 z_{ji} z_{ki}. \quad (16)$$

Καθώς $Y_i \sim B(1, \pi_i)$ ισχύουν τα παρακάτω:

$$\begin{aligned} V(\pi_i) &= \pi_i(1 - \pi_i), & \pi_i &= \frac{\exp\{\eta_i\}}{1 + \exp\{\eta_i\}}, & \alpha(\phi) &= 1, \\ \eta_i &= \sum_{j=1}^k \gamma_j z_{ij}, & \frac{\partial \pi_i}{\partial \eta_i} &= \pi_i(1 - \pi_i). \end{aligned} \quad (17)$$

Αντικαθιστώντας τις παραπάνω σχέσεις στην (16) προκύπτει:

$$\left[\hat{\mathbf{V}}_H \right]_{jk} = \sum_{i=1}^n \pi_i(1 - \pi_i) z_{ji} z_{ki}. \quad (18)$$

Ο όρος $\hat{\mathbf{B}}_s$ είναι ένας διορθωτικός όρος και υπολογίζεται από την εξίσωση:

$$\hat{\mathbf{B}}_s = \sum_{i=1}^n \mathbf{z}_i^T \left\{ \frac{y_i - \pi_i}{V(\pi_i)} \left(\frac{\partial \pi_i}{\partial \eta_i} \right) \right\}^2 \mathbf{z}_i. \quad (19)$$

Λαμβάνοντας υπόψη τη σχέση (17) η σχέση (20) θα γίνει:

$$\hat{B}_s = \sum_{i=1}^n \mathbf{z}_i^T (y_i - \pi_i)^2 \mathbf{z}_i. \quad (20)$$

Τέλος, υπολογίζεται το 95% διάστημα εμπιστοσύνης, όπου τα τυπικά σφάλματα εκτιμούνται ως η τετραγωνική ρίζα των στοιχείων της διαγωνίου του εκτιμητή sandwich.

3. ΤΥΧΑΙΟΠΟΙΗΜΕΝΑ ΠΟΣΟΣΤΙΑΙΑ ΥΠΟΛΟΙΠΑ

Στην εργασία αυτή εξετάζεται επίσης αν η προσαρμογή ενός γενικευμένου γραμμικού μοντέλου στην περιγραφή του συστήματος της μυϊκής ατράκτου είναι έγκυρο και κατάλληλο. Για το σκοπό αυτό χρησιμοποιούνται τα συνήθη υπόλοιπα που εκφράζουν τη διαφορά μεταξύ της παρατηρούμενης και της προσαρμοζόμενης τιμής.

Έστω $F(y_i; \pi_i) = \Pr(Y_i \leq y_i) = \sum_{m=0}^{\lfloor y_i \rfloor} \pi_i^m (1 - \pi_i)^{1-m}$ η αθροιστική συνάρτηση κατανομής της i -οστής διωνυμικής απόκρισης, όπου $\lfloor y_i \rfloor$ είναι ο μεγαλύτερος ακέραιος που είναι μικρότερος ή ίσος του y_i . Τότε τα τυχαιοποιημένα ποσοστιαία υπόλοιπα ορίζονται ως

$$r_{rq,i} = \Phi^{-1}(u_i),$$

όπου u_i είναι μια ομοιόμορφη τυχαία μεταβλητή στο διάστημα

$$(a_i, b_i] = \left(\lim_{y \uparrow y_i} F(y; \hat{\pi}_i), F(y_i; \hat{\pi}_i) \right] \approx [F(y_i - 1; \hat{\pi}_i), F(y_i; \hat{\pi}_i)],$$

και $\Phi(x)$ είναι η αθροιστική συνάρτηση της τυπικής κανονικής κατανομής. Η κατανομή των ποσοστιαίων υπολοίπων καθώς και των τυχαιοποιημένων ποσοστιαίων υπολοίπων $r_{rq,i}$ είναι τυπική κανονική αν οι εκτιμητές των β_1, \dots, β_k είναι συνεπείς (Dunn and Smyth, 1996). Έτσι η εγκυρότητα του προσαρμοσθέντος μοντέλου μπορεί να διαπιστωθεί χρησιμοποιώντας ελέγχους καλής προσαρμογής για την κανονικότητα των $r_{rq,i}$.

Ένας εύκολος τρόπος για να ελεγχθεί η μηδενική υπόθεση $H_0: \mathbf{r}_{rq} \sim \mathbf{N}(\mathbf{0}, \mathbf{I})$, που θεωρεί ότι το διάλυμα των τυχαιοποιημένων ποσοστιαίων υπολοίπων ακολουθεί πολυμεταβλητή κανονική κατανομή με μηδενική μέση τιμή και μοναδιαίο πίνακα διασποράς—συνδιασποράς, είναι η κατασκευή μιας 5% απορριπτικής περιοχής (R) στο n -διάστατο χώρο. Στην περίπτωση μας τα τυχαιοποιημένα ποσοστιαία υπόλοιπα είναι ανεξάρτητα και μία απορριπτική περιοχή μπορεί να αποτελείται από τα εξωτερικά σημεία μιας n -διάστατης υπερσφαίρας. Τότε,

$$(r_{rq,1}, \dots, r_{rq,n}) \in R \Leftrightarrow S = \|(r_{rq,1}, \dots, r_{rq,n})\| > \epsilon, \quad (21)$$

όπου το ϵ είναι τέτοιο ώστε

$$\text{Prob}\{(r_{rq,1}, \dots, r_{rq,n}) \in R | H_0\} = 0.05 \Leftrightarrow \text{Prob}\left\{\left[\sum_{i=1}^n r_{rq,i}^2\right]^{1/2} > \epsilon\right\} = 0.05.$$

Από τα παραπάνω προκύπτει ότι $\epsilon^2 = \chi_{n;0.95}^2$, όπου $\chi_{n;0.95}^2$ είναι το κάτω 95% ποσοστιαίο σημείο της κατανομής χ_n^2 .

Ένας τρόπος για τον έλεγχο των αποκλίσεων από τη μηδενική υπόθεση είναι η κατασκευή της γραφικής παράστασης Q-Q. Στο παράδειγμά μας χρησιμοποιείται η βελτιωμένη γραφική παράσταση Q-Q στην οποία έχουν προστεθεί 5% απορριπτικές περιοχές με τη βοήθεια του θεωρήματος του David (1981) για την ασυμπτωτική κατανομή των ποσοστιαίων σημείων (Βασιλειάδης και Ρήγας, 2010). Λεπτομέρειες σχετικά με τη βελτιωμένη γραφική παράσταση Q-Q δίνονται στις εργασίες (Βασιλειάδης και συν., 2008) και (Vassiliadis and Rigas, 2009).

4. ΤΟ ΠΑΡΑΔΕΙΓΜΑ ΤΗΣ ΜΥΪΚΗΣ ΑΤΡΑΚΤΟΥ

Στην παράγραφο αυτή κατασκευάζουμε ένα λογιστικό μοντέλο που περιγράφει τις αποκρίσεις του πρωτεύοντα άξονα Ια και του δευτερεύοντα άξονα ΙΙ της μυϊκής ατράκτου υπό την επίδραση ενός γάμμα κινητονεύρου. Το μοντέλο που χρησιμοποιείται για την περιγραφή του συστήματος βασίζεται στις εργασίες του Brillinger (1988) και των Karavasilis et al. (2005). Στη συγκεκριμένη εργασία το μοντέλο που περιγράφει για κάθε έξοδο τη συμπεριφορά της μυϊκής ατράκτου περιλαμβάνει δύο παραμέτρους: το κατώφλι και την αθροιστική συνάρτηση. Θεωρούμε ότι y_t είναι η διεργασία πυροδότησης του συστήματος, για την οποία ισχύει $y_t \sim B(1, \pi_t)$. Το μοντέλο της λογιστικής παλινδρόμησης παίρνει την ακόλουθη μορφή:

$$\log \left(\frac{\pi_{i,t}}{1 - \pi_{i,t}} \right) = \sum_{u \leq t} \alpha_{i,u} x_{t-u} - \theta_0, \quad i = 1, 2. \quad (22)$$

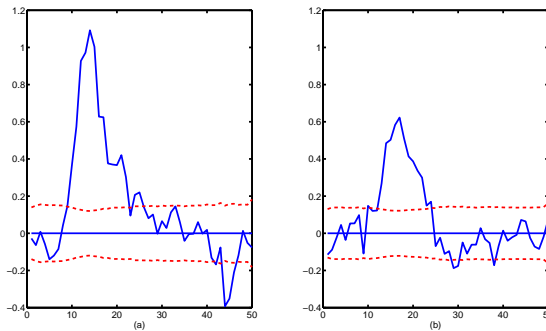
όπου x_{t-u} είναι οι διεργασίες πυροδότησης του γάμμα κινητονεύρου και θ_0 είναι μία παράμετρος κατωφλίου. Ο συνολικός χρόνος καταγραφής των δεδομένων είναι $T = 60000 \text{ msec}$. Η είσοδος (γάμμα κινητονεύρο) περιλαμβάνει 3771 παλμούς, η απόκριση του Ια αισθητήριου άξονα 1231 παλμούς και η απόκριση του ΙΙ αισθητήριου άξονα 1712 παλμούς. Στο Σχήμα 1(a) παρουσιάζονται οι εκτιμητές των συντελεστών της αθροιστικής συνάρτησης στην Ια έξοδο υπό την επίδραση του γάμμα κινητονεύρου. Αυτό που διαπιστώνεται είναι ότι το σύστημα αρχίζει και επιταχύνει γύρω στα 10 msec και συνεχίζει μέχρι το 23 msec. Επίσης, στο Σχήμα 1(b) παρουσιάζονται οι εκτιμητές των συντελεστών της αθροιστικής συνάρτησης στην ΙΙ έξοδο, όπου και σε αυτή την περίπτωση το σύστημα επιταχύνει στο διάστημα 12-25 msec, ενώ παράλληλα παρατηρείται και μία μικρή μετατόπιση στην κορυφή. Στο Σχήμα 2(a) παρουσιάζεται η παράσταση Q-Q τυχοποιημένων ποσοστιαίων υπολοίπων για την Ια έξοδο. Οι 5% απορριπτικές περιοχές υπολογίζονται με τη μέθοδο του David. Από τον έλεγχο προκύπτει ότι το 4.52% των 60000 υπολοίπων βρίσκονται εκτός των 5% απορριπτικών περιοχών και γενικά το γράφημα Q-Q δεν παρουσιάζει αποκλίσεις από την κανονικότητα. Η απόσταση S της σχέσης (21) στην περίπτωση αυτή γίνεται $S = 242.93 < \epsilon = \sqrt{\chi_{60000;0.95}^2} = 246.11$ που σημαίνει ότι η μηδενική υπόθεση δεν μπορεί να απορριφθεί και το μοντέλο είναι έγκυρο. Παρόμοια αποτελέσματα

προκύπτουν και για την II απόκριση, τα οποία απεικονίζονται στο Σχήμα 2(b) και 2.13% των 60000 υπολοίπων βρίσκονται εκτός των 5% απορριπτικών περιοχών και το $S = 244.40 < \epsilon = \sqrt{\chi_{60000;0.95}^2} = 246.11$.

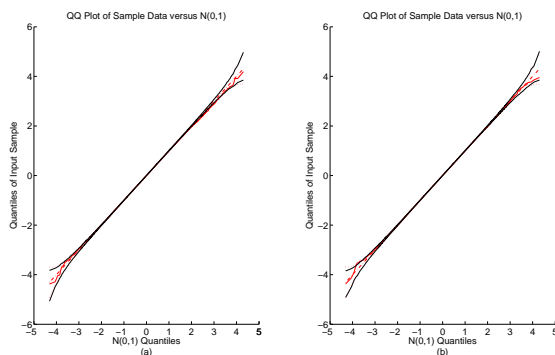
ABSTRACT

The identification of a neurophysiological system, called muscle spindle, is studied when it is affected by a gamma motoneuron and produces two responses in order to describe its behavior. A time domain analysis is used, where the input and the outputs of the system can be considered as binary time series and therefore the theory of generalized linear models (GLMs) can be applied. The advantage of this approach is based on the fact that estimates of the system's parameters can be obtained by using the maximum likelihood function. In this paper, we use the penalized likelihood function and more specific the Ridge regression. The stochastic model which is proposed for the description of the system involves the threshold and the summation function, which is of great interest as it describes whether the system is inhibitory or excitatory. Furthermore, a test for the validity of the fitted model is presented based on the randomized quantile residuals which follow a Gaussian distribution. The Q-Q plot is used with the addition of 5% rejection regions evaluated with the help of David's Theorem for the asymptotic distribution of the quantiles. An illustrative example from the field of Neurophysiology is presented in which the behavior of muscle spindle is studied under the influence of a gamma motoneuron.

Σχήμα 1: Η αθροιστική συνάρτηση της (a) Ia εξόδου και (b) II εξόδου υπό την επίδραση του γάμμα κινητονεύρου



Σχήμα 2: Η παράσταση Q-Q για την (a) Ia έξοδο και (b) για την II έξοδο. Οι 5% απορριπτικές περιοχές υπολογίζονται με τη μέθοδο του David.



ΑΝΑΦΟΡΕΣ

- Brillinger, D.R. (1988). Maximum likelihood analysis of spike trains of interacting nerve cells. *Biol. Cybern.* **59**, 189-200.
- Brillinger, D.R., Lindsay, K.A. and Rosenberg J.R. (2009). Combining frequency and time domain approaches to systems with multiple spike train input. *Biological Cybernetics* **100**, 459-474.
- David, H.A. (1981). *Order Statistics*, New York: Wiley.
- Dunn, P.K. and Smyth, G.K. (1996). Randomized quantile residuals. *J. Comput. Graph. Stat.* **5(3)**, 236-244.
- Hardin, W.J., Hilbe, M.J. 2007. *Generalized Linear Models and Extensions*. Stata Press, Texas.
- Hoerl, A.E. (1962). Application of Ridge analysis to regression problems. *Chem. Eng. Pr.* **55**, 69-78.
- Karavasilis, G.J., Kotti, V.K., Tsitsis, D.S., Vassiliadis, V.G., Rigas, A.G. (2005). Statistical methods and software for risk assessment: applications to a neurophysiological data set. *Computational Statistics and Data Analysis* **49**, 243-263.
- Kauermann, G. and Carroll, R.J. (2001). A note on the efficiency of sandwich covariance matrix estimation. *Journal of the American Statistical Association* **96**, 1387-1396.
- McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models, Second Edition*, London: Chapman and Hall.
- Le Cessie, S. and Van Houwelingen, D. (1992). Ridge estimators in logistic regression. *Appl. Statist.* **41**, 191-201.
- Vassiliadis, V.G. and Rigas, A.G. (2009). A new formulation of the Hinich's bispectral test for linearity based on a novel Q-Q plot for testing distributional hypotheses. In: Kitsos, C.P. and Caroni C. (Eds.), *Proc. International Conference on Cancer Risk Assessment 3*.

- Βασιλειάδης, Β.Γ., Κώττη, Β.Κ. και Ρήγας, Α.Γ (2008). Ο έλεγχος της εγκυρότητας ενός γενικευμένου γραμμικού μοντέλου με τη χρήση των τυχοποιημένων ποσοστιαίων υπολοίπων. *Πρακτικά 21ου Πανελληνίου Συνεδρίου Στατιστικής*, 81-88.
- Βασιλειάδης και Ρήγας, Α.Γ (2010). Επίλυση του διαχωρισμού σε λογιστικά μοντέλα με τη χρήση της ποινικοποιημένης συνάρτησης πιθανοφάνειας. *Πρακτικά 23ου Πανελληνίου Συνεδρίου Στατιστικής*.



Κρυφά Μαρκοβιανά και ημι-Μαρκοβιανά μοντέλα για τον εντοπισμό του υποκείμενου κατά τη γένεση σεισμών πεδίου τάσεων

Ε. Βότση^{1,2}, Ν. Λημνιός¹, Γ. Τσακλίδης², Ε. Παπαδημητρίου³

¹Laboratoire de Mathématiques Appliquées de Compiègne

Université de Technologie de Compiègne

{irene.votsi, nikolaos.limnios}@utc.fr

²Τμήμα Μαθηματικών, Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης

tsaklidi@math.auth.gr

³Τμήμα Γεωλογίας, Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης

ritsa@geo.auth.gr

ΠΕΡΙΛΗΨΗ

Η αναγκαιότητα της εφαρμογής των κρυφών Μαρκοβιανών μοντέλων στη Σεισμολογία οφείλεται στο γεγονός ότι οι απόλυτες τιμές του πεδίου των τάσεων δεν είναι προσβάσιμες σε παρατήρηση. Με σκοπό την όσο το δυνατόν ακριβέστερη προσέγγιση των τιμών του πεδίου των τάσεων σε συγκεκριμένο χώρο και χρόνο, εφαρμόζουμε κρυφά Μαρκοβιανά και ημι-Μαρκοβιανά μοντέλα τύπου SM-M0, διακριτού χρόνου και χώρου καταστάσεων.

Για την εφαρμογή του κρυφού Μαρκοβιανού μοντέλου το ενδιαφέρον εστιάζεται στη μη παραμετρική περίπτωση, ενώ οι κατανομές των χρόνων παραμονής σχετίζονται με τις καταστάσεις. Δεδομένης μιας ακολουθίας παρατηρήσεων που σχετίζεται με τη γένεση σεισμών, ο κύριος στόχος είναι ο εντοπισμός των κρυφών καταστάσεων που έπεται της εκτίμησης των πιθανοτήτων μετάβασης της Μαρκοβιανής αλυσίδας και των παραμέτρων της κατανομής των παρατηρήσεων. Επιλύονται τα τρία βασικά προβλήματα της θεωρίας και των εφαρμογών των κρυφών Μαρκοβιανών μοντέλων, δηλαδή το πρόβλημα εκτίμησης, υπολογισμού και αποκωδικοποίησης. Το βέλτιστο κρυφό Μαρκοβιανό για διαφορετικό πλήθος καταστάσεων επιλέγεται βάσει των τιμών AIC και BIC. Στη συνέχεια το βέλτιστο μοντέλο προσομοιώνεται προσφέροντας ένα σύνολο καταστάσεων και παρατηρήσεων που συνεισφέρει στην εκτίμηση της σεισμικής επικινδυνότητας.

Ο βασικός περιορισμός των κρυφών Μαρκοβιανών μοντέλων διακριτού χρόνου είναι η υπόθεση της γεωμετρικής κατανομής ως κατανομής των χρόνων παραμονής. Ο ανωτέρω σχετικός περιορισμός αίρεται με την εφαρμογή των κρυφών ημι-Μαρκοβιανών μοντέλων. Ειδικότερα θεωρούνται μη παραμετρικά εκτιμώμενοι χρόνοι παραμονής ενώ η κατανομή Poisson επιλέγεται ως η κατανομή των παρατηρήσεων με παράμετρο που καθορίζεται από την αντίστοιχη κατάσταση της ημι-Μαρκοβιανής αλυσίδας.

Λέξεις Κλειδιά: Μαρκοβιανή αλυσίδα, ημι-Μαρκοβιανή αλυσίδα, ημι-Μαρκοβιανός πυρήνας, διαδικασία παρατηρήσεων, κατανομή Poisson.

1. Εισαγωγή

Τα κρυφά Μαρκοβιανά μοντέλα (hidden Markov models (HMMs)) εισήχθησαν από τους Baum και Petrie (1966) κι έκτοτε εφαρμόστηκαν ευρέως σε πολλές επιστήμες. Ειδικότερα τα HMMs εφαρμόστηκαν για πρώτη φορά στη Σεισμολογία από τους Granat και Donnellan (2002) για τον καθορισμό διακριτών συστάδων σεισμών από σύνολο δεδομένων της νότιας Καλιφόρνιας. Οι Ebel et al. (2007) εφάρμοσαν HMMs υποθέτοντας εκθετική κατανομή παρατηρήσεων κι εκτίμησαν τη θέση και το χρόνο γένεσης του επόμενου σεισμού. Τέλος, οι Orfanogiannaki et al. (2010) εντόπισαν χρονικές μεταβολές σεισμικότητας με χρήση Poisson HMMs.

Στην παρούσα μελέτη εφαρμόζουμε HMMs με στόχο τον καθορισμό των απόλυτων τιμών του πεδίου των τάσεων. Είναι δυνατόν να υπολογίσουμε τις μεταβολές του πεδίου οι οποίες οφείλονται στη σεισμική ολίσθηση καθώς και στη διαρκή τεκτονική φόρτιση, όμως η δυνατότητα καθορισμού των μεταβολών αυτών περιορίζεται χρονικά λόγω της μη-πληρότητας των ιστορικών καταλόγων. Στη συνέχεια εφαρμόζουμε κρυφά ημι-Μαρκοβιανά μοντέλα (hidden semi-Markov models (HSMMs)) θεωρώντας την κατανομή Poisson ως κατανομή των παρατηρήσεων κι εκτιμώντας μη παραμετρικά την κατανομή των χρόνων παραμονής στις καταστάσεις της ημι-Μαρκοβιανής αλυσίδας (M.A).

2. Κρυφό Μαρκοβιανό Μοντέλο

Ένα κρυφό Μαρκοβιανό μοντέλο διακριτού χρόνου είναι μια διπλή στοχαστική διαδικασία που περιλαμβάνει μια μη παρατηρήσιμη (“κρυφή”) Μαρκοβιανή αλυσίδα (hidden Markov chain (HMC)) και μια διαδικασία εκπεμπόμενων από την κρυφή Μαρκοβιανή αλυσίδα, παρατηρήσεων (observation process).

Θεωρούμε μια ομογενή Μαρκοβιανή αλυσίδα $(J_n)_{n \in \mathbb{N}}$ με πεπερασμένο χώρο καταστάσεων E και πίνακα πιθανοτήτων μετάβασης $P = (P(i, j))$, $i, j \in E$. Η αρχική κατανομή πιθανοτήτων (initial law) της M.A. συμβολίζεται με α . Συμβολίζουμε τη διαδικασία παρατηρήσεων ή εκπομπών με $(Y_n)_{n \in \mathbb{N}}$ και τον αντίστοιχο χώρο καταστάσεων με A . Η σχέση που συνδέει τις δύο διαδικασίες (J_n) , (Y_n) είναι η ακόλουθη: $R_{i;a} = P(Y_n = a \mid J_n = i)$, $i \in E, a \in A$. Ο στοχαστικός πίνακας $R = (R_{i;a})$, $i \in E, a \in A$ ονομάζεται πίνακας πιθανοτήτων εκπομπής (emission probability matrix). Συμβολίζουμε επίσης με Y_l^k το διάνυσμα $Y_l^k := (Y_l, \dots, Y_k)$ κι όμοια συμβολίζουμε με J_l^k το διάνυσμα $J_l^k = (J_l, \dots, J_k)$. Θεωρούμε ότι το HMM είναι τύπου SM-M0, δηλαδή ισχύει $P(Y_n = a \mid Y_0^{n-1} = \cdot, J_0^{n-1} = \cdot, J_n = i) = P(Y_n = a \mid J_n = i) = R_{i;a} \forall a \in A, i \in E, n \in \mathbb{N}$.

Για λόγους συντομίας το σύνολο των παραμέτρων του μοντέλου συμβολίζεται με $\theta = (P, R, \alpha)$. Για να προσδιορίσουμε πλήρως ένα HMM είναι απαραίτητος ο καθορισμός του πλήθους των καταστάσεων της M.A. και του συνόλου των παρατηρήσεων καθώς επίσης και η εκτίμηση του συνόλου των παραμέτρων θ . Η στατιστική

συμπερασματολογία που αφορά τις δύο διαδικασίες προκύπτει αποκλειστικά μέσω της διαθέσιμης πληροφορίας, δηλαδή αποκλειστικά μέσω των παρατηρήσεων.

Προκειμένου να εφαρμόσουμε τη θεωρία των HMMs σε σύγχρονα προβλήματα, συμπεριλαμβανομένου του καθορισμού του υποκείμενου κατά τη γένεση σεισμών πεδίου των τάσεων, κρίνεται αναγκαία η επίλυση των επόμενων προβλημάτων.

1. *Πρόβλημα Εκτίμησης (Estimation Problem)*: Δοθείσας μιας ακολουθίας παρατηρήσεων και της δομής του HMM, καθορίζουμε το σύνολο των παραμέτρων του, θ , μέσω του αλγορίθμου Baum-Welch.
2. *Πρόβλημα Υπολογισμού (Evaluation Problem)*: Δοθέντος του συνόλου των παραμέτρων, θ , υπολογίζουμε την πιθανότητα πραγματοποίησης μιας συγκεκριμένης ακολουθίας παρατηρήσεων μέσω του αλγορίθμου Forward-Backward.
3. *Πρόβλημα Αποκωδικοποίησης (Decoding Problem)*: Δοθέντος του θ και της ακολουθίας των παρατηρήσεων, υπολογίζουμε τη βέλτιστη ακολουθία καταστάσεων ως αυτή που αντιστοιχεί σε μεγαλύτερη πιθανότητα συνδυασμού καταστάσεων και παρατηρήσεων με εφαρμογή του αλγορίθμου Viterbi.

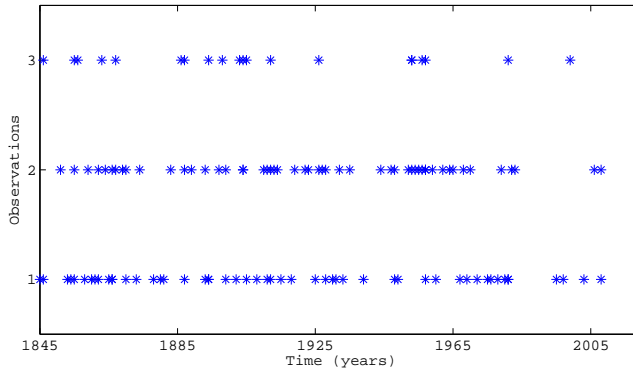
Παρατηρώντας το σύνολο των εκπομπών της M.A., ο αρχικός στόχος είναι η εκτίμηση του συνόλου των παραμέτρων, θ . Η μοντελοποίηση με την εφαρμογή των HMMs οδηγεί στην επίτευξη του τελικού στόχου, δηλαδή στην αποκωδικοποίηση των καταστάσεων της κρυφής M.A. (Ephraim and Merhav 2002).

3. Εφαρμογή του HMM

3.1 Καθορισμός Συνόλου Δεδομένων

Το σύνολο δεδομένων προέρχεται από κατάλογο που εκπονήθηκε στον Τομέα Γεωφυσικής του Αριστοτέλειου Πανεπιστημίου Θεσ/νίκης βάσει των καταγραφών στο εθνικό σεισμολογικό δίκτυο (<http://geophysics.geo.auth.gr/ss/>). Οι σεισμοί έχουν γίνει στον ευρύτερο ελληνικό χώρο κατά το χρονικό διάστημα 1845-2008. Ο κατάλογος καλύπτει μεγάλο χρονικό διάστημα και είναι πλήρης, δηλαδή περιέχει πληροφορίες για όλους τους σεισμούς με μέγεθος $M \geq 6.5$ οι οποίοι έχουν γίνει από το 1845 έως το 2008.

Το ενδιαφέρον μας επικεντρώνεται στην περίπτωση όπου οι παρατηρήσεις καθορίζονται βάσει των μεγεθών των σεισμών. Το σύνολο των παρατηρήσεων χωρίζεται σε τρία υποσύνολα βάσει της μεθόδου ομαδοποίησης κατά συστάδες k-means : το 1^ο υποσύνολο περιλαμβάνει σεισμούς μεγέθους $M \in [6.5, 6.7]$ (62 παρατηρήσεις), το 2^ο υποσύνολο περιλαμβάνει σεισμούς μεγέθους $M \in [6.8, 7.1]$ (61 παρατηρήσεις) και το 3^ο υποσύνολο περιλαμβάνει σεισμούς μεγαλύτερου μεγέθους (21 παρατηρήσεις) (Σχ. 1).



Σχήμα 1: Καθορισμός του συνόλου των παρατηρήσεων βάσει των μεγεθών των σεισμών.

3.2 Διδιάστατος και Τριδιάστατος Χώρος Καταστάσεων

Αρχικά θεωρούμε ότι ο χώρος καταστάσεων της κρυφής Μαρκοβιανής αλυσίδας είναι διδιάστατος και οι αρχικές τιμές των παραμέτρων επιλέγονται από την ομοιόμορφη κατανομή $U(0, 1)$, εξασφαλίζοντας τη στοχαστικότητα των πινάκων P και R . Ο εκτιμώμενος πίνακας πιθανοτήτων μετάβασης, \hat{P} , είναι:

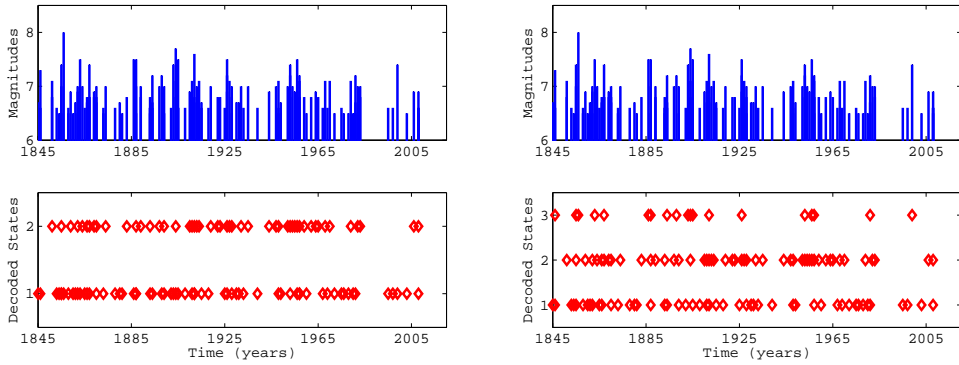
$$\begin{array}{cc} & \begin{array}{cc} \text{Κατάσταση 1} & \text{Κατάσταση 2} \end{array} \\ \begin{array}{c} \text{Κατάσταση 1} \\ \text{Κατάσταση 2} \end{array} & \left(\begin{array}{cc} 0.355 & 0.645 \\ 0.535 & 0.465 \end{array} \right). \end{array}$$

Στη συνέχεια παρουσιάζουμε τον πίνακα πιθανοτήτων εκπομπής, \hat{R} , όπως αυτός προκύπτει από την επίλυση του προβλήματος εκτίμησης:

$$\begin{array}{ccc} & \text{Παρατ. Τύπου 1} & \text{Παρατ. Τύπου 2} & \text{Παρατ. Τύπου 3} \\ \begin{array}{c} \text{Κατάσταση 1} \\ \text{Κατάσταση 2} \end{array} & \left(\begin{array}{ccc} 0.679 & 0.002 & 0.319 \\ 0.222 & 0.778 & 0.000 \end{array} \right). \end{array}$$

Ο Baum-Welch αλγόριθμος συγκλίνει σε 146 βήματα στην τελική τιμή $\log L = -143.89$, καθώς η διαφοροποίηση των διαδοχικών τιμών της $\log L$ είναι μηδενική (προσέγγιση 5 δεκαδικών ψηφίων). Στο σχήμα 2 (αριστερά) παρουσιάζεται η βέλτιστη ακολουθία των κρυφών καταστάσεων σύμφωνα με τον αλγόριθμο Viterbi.

Παρατηρούμε ότι η πρώτη κατάσταση της Μ.Α. εκπέμπει κυρίως παρατηρήσεις τύπου 1 ενώ οι παρατηρήσεις τύπου 2 εκπέμπονται κατά κανόνα από τη δεύτερη κατάσταση η οποία παράγει και παρατηρήσεις τύπου 1. Σύμφωνα με το HMM οι παρατηρήσεις τύπου 3 πηγάζουν αποκλειστικά από την πρώτη κατάσταση της Μ.Α.



Σχήμα 2: Βέλτιστη ακολουθία καταστάσεων με βάση τον αλγόριθμο Viterbi για διδιάστατο (αριστερά) και τριδιάστατο (δεξιά) χώρο καταστάσεων.

Στη συνέχεια θεωρούμε την ύπαρξη τριών κρυφών καταστάσεων που γεννούν το σύνολο των παρατηρήσεων. Υποθέτοντας ότι οι αρχικές τιμές των παραμέτρων του μοντέλου ακολουθούν ομοιόμορφη κατανομή, επιλύονται τα τρία προβλήματα που σχετίζονται με τη θεωρία των κρυφών Μαρκοβιανών μοντέλων. Ο εκτιμώμενος πίνακας πιθανοτήτων μετάβασης, \hat{P} , είναι

$$\begin{array}{c}
 \text{Κατάσταση 1} \quad \text{Κατάσταση 2} \quad \text{Κατάσταση 3} \\
 \text{Κατάσταση 1} \left(\begin{array}{ccc} 0.414 & 0.498 & 0.088 \\ \text{Κατάσταση 2} & 0.403 & 0.396 & 0.201 \\ \text{Κατάσταση 3} & 0.297 & 0.559 & 0.144 \end{array} \right).
 \end{array}$$

Ο πίνακας των πιθανοτήτων εκπομπής των παρατηρήσεων, \hat{R} , είναι

$$\begin{array}{c}
 \text{Παρατ. Τύπου 1} \quad \text{Παρατ. Τύπου 2} \quad \text{Παρατ. Τύπου 3} \\
 \text{Κατάσταση 1} \left(\begin{array}{ccc} 0.998 & 0.000 & 0.002 \\ \text{Κατάσταση 2} & 0.074 & 0.926 & 0.000 \\ \text{Κατάσταση 3} & 0.007 & 0.001 & 0.992 \end{array} \right).
 \end{array}$$

Χρησιμοποιώντας ως κριτήριο σύγκλισης το μηδενισμό των διαδοχικών τιμών της $\log L$ (προσέγγιση 5 δεκαδικών ψηφίων), ο αλγόριθμος Baum-Welch συγκλίνει σε 401 βήματα στην τελική τιμή $\log L = -142.33$. Επιλύοντας το πρόβλημα αποκωδικοποίησης η βέλτιστη ακολουθία κρυφών καταστάσεων με βάση τον αλγόριθμο Viterbi δίνεται στο σχήμα 2 (δεξιά).

Σύμφωνα με τις εκτιμήσεις και την αποκωδικοποίηση που προηγήθηκε, το μοντέλο υποδεικνύει σε γενικές γραμμές ότι οι παρατηρήσεις τύπου 1 ($M \in [6.5, 6.7]$) εκπέμπονται από την κρυφή κατάσταση 1, παρατηρήσεις τύπου 2 ($M \in [6.8, 7.1]$) εκπέμπονται από την κρυφή κατάσταση 2, ενώ οι υπόλοιπες παρατηρήσεις προκύπτουν από την κρυφή κατάσταση 3.

3.3 Επιλογή Βέλτιστου Μοντέλου

Προκειμένου να επιλέξουμε το βέλτιστο ως προς το πλήθος καταστάσεων μοντέλο, χρησιμοποιούμε το κριτήριο πληροφορίας του Akaike (Akaike 1974), $AIC = -2 \cdot \log L + 2 \cdot m^2$, όπου $\log L$ η εκτιμώμενη λογαριθμική πιθανοφάνεια και m το πλήθος των καταστάσεων του μοντέλου. Οι τιμές του κριτηρίου υποδεικνύουν ότι το μοντέλο των δύο καταστάσεων ($AIC=295.76$) προσαρμόζεται καλύτερα στα δεδομένα από το μοντέλο των τριών καταστάσεων ($AIC=302.65$). Για την επιλογή του βέλτιστου μοντέλου εφαρμόζουμε επίσης το κριτήριο πληροφορίας του Bayes (Schwarz 1978), $BIC = -2 \cdot \log L + m^2 \cdot \log n$, όπου n συμβολίζει το πλήθος των παρατηρήσεων. Το μοντέλο των τριών καταστάσεων λαμβάνει μεγαλύτερη τιμή του BIC (329.378) συγκριτικά με εκείνο των δύο καταστάσεων (307.639). Σύμφωνα με τα δύο κριτήρια ως βέλτιστο μοντέλο υποδεικνύεται το HMM με διδιάστατο χώρο καταστάσεων.

3.4 Προσομοίωση Βέλτιστου Μοντέλου

Προσομοιώνοντας το βέλτιστο HMM με εφαρμογή της μεθόδου προσομοίωσης Markov chain Monte Carlo, προκύπτει ένα σύνολο 1000 προσομοιωμένων καταστάσεων και παρατηρήσεων. Αρχικά προσομοιώνουμε την M.A. κι έπειτα μετασχηματίζουμε την πιθανότητα επίσκεψης σε μια κατάσταση (όπως έχει καθοριστεί από την προσομοίωση της M.A.) σε πιθανότητα εκπομπής κάποιου τύπου παρατηρήσεων βάσει του εκτιμώμενου πίνακα πιθανοτήτων εκπομπής. Στον Πίνακα 1 παρουσιάζονται οι πιθανότητες επίσκεψης στις καταστάσεις της κρυφής M.A. για το σύνολο δεδομένων και το προσομοιωμένο σύνολο δεδομένων.

Πίνακας 1: Πιθανότητες επίσκεψης στις καταστάσεις της κρυφής M.A. για το πραγματικό και το προσομοιωμένο σύνολο δεδομένων.

Καταστάσεις	Σύνολο Δεδομένων	Προσομοιωμένο Σύνολο Δεδομένων
1	0.576	0.568
2	0.424	0.432

Ο δίπλευρος έλεγχος για τη διαφορά των πιθανοτήτων επίσκεψης σε μια κατάσταση της M.A. μεταξύ του συνόλου των δεδομένων και του προσομοιωμένου συνόλου δεδομένων, οδηγεί στο συμπέρασμα ότι δεν μπορούμε να απορρίψουμε σε επίπεδο εμπιστοσύνης 95% πως οι πιθανότητες αυτές είναι ίσες. Όμοια μπορούμε να συγκρίνουμε τις πιθανότητες των παρατηρήσεων του συνόλου των δεδομένων με τις αντίστοιχες των προσομοιωμένων δεδομένων. Προκειμένου να ελέγξουμε το μοντέλο, το εφαρμόζουμε εκ νέου στο σύνολο των προσομοιωμένων παρατηρήσεων κι ακολουθεί σύγκριση των εκτιμήσεων του θ όπως αυτές προκύπτουν για το σύνολο των δεδομένων και το αντίστοιχο προσομοιωμένο σύνολο. Επιπλέον οι εκτιμήσεις που προηγήθηκαν καθιστούν εφικτό τον υπολογισμό του μέσου πλήθους βημάτων καθώς και της αντίστοιχης διασποράς για την μετάβαση σε οποιαδήποτε κατάσταση της M.A. ή ακόμη για μετάβαση σε οποιοδήποτε συνδυασμό κατάστασης και παρατήρησης.

4. Κρυφό Ημι-Μαρκοβιανό Μοντέλο

Ο κύριος περιορισμός των κρυφών Μαρκοβιανών μοντέλων διακριτού χρόνου είναι η υπόθεση της γεωμετρικής κατανομής ως κατανομής των χρόνων παραμονής. Ο περιορισμός αυτός αίρεται με την εισαγωγή των κρυφών ημι-Μαρκοβιανών μοντέλων (Ferguson 1980). Συμβολίζουμε με $(S_n)_{n \in \mathbb{N}}$ τους χρόνους των αλμάτων της Μαρκοβιανής αλυσίδας $(J_n)_{n \in \mathbb{N}}$ και με $(X_n)_{n \in \mathbb{N}}$ τους διαδοχικούς χρόνους παραμονής. Θεωρούμε την αντίστοιχη Μαρκοβιανή αλυσίδα ανανέωσης $(J_n, S_n)_{n \in \mathbb{N}}$ με ημι-Μαρκοβιανό πυρήνα $q_{ij}(k) := P(J_{n+1} = j, X_{n+1} = k | J_n = i)$, $i, j \in E$, $k, n \in \mathbb{N}$. Ορίζουμε την ημι-Μαρκοβιανή αλυσίδα (semi-Markov chain (SMC)) ως $Z_k = J_{N(k)}$, $k \in \mathbb{N}$, όπου $N(k) := \max\{n \geq 0 : S_n \leq k\}$ η διακριτού χρόνου απαριθμητή διαδικασία του αριθμού των αλμάτων στο διάστημα $[1, k] \subset \mathbb{N}$. Επιπλέον η κατανομή των χρόνων παραμονής σχετίζεται με τις καταστάσεις υπό την έννοια ότι $q_{ij}(k) = P(Z_{S_1} = j | Z_0 = i)P(S_1 = k | Z_0 = i) = P(i, j)f_i(k)$, $\forall i, j \in E, k \in \mathbb{N}$. Ένα κρυφό ημι-Μαρκοβιανό μοντέλο ορίζεται από μια διδιάστατη διαδικασία $(Z_n, Y_n)_{n \in \mathbb{N}}$, όπου Z_n είναι μια κρυφή SMC και Y_n η διαδικασία των παρατηρήσεων. Θεωρούμε ότι το υπό μελέτη HSMM είναι τύπου SM-M0.

Έστω $\mathcal{H}(M) := (J_0, X_1, \dots, J_{N(M)-1}, X_{N(M)}, J_{N(M)}, U_M)$ ένα δειγματικό μονοπάτι της κρυφής SMC λογοκριμένο σε σταθερό χρόνο $M \in \mathbb{N}$, δηλαδή μια ακολουθία καταστάσεων και χρόνων παραμονής, όπου $U_M = M - S_{N(M)}$ ο λογοκριμένος χρόνος παραμονής στην τελευταία κατάσταση $J_{N(M)}$ και $\overline{H}_i(\cdot) = P(X_1 > \cdot | J_0 = i)$ η συνάρτηση επιβίωσης του χρόνου παραμονής στην κατάσταση $i \in E$. Χρησιμοποιούμε το συμβολισμό $\theta = (q, R)$ για τις παραμέτρους του μοντέλου.

Η συνάρτηση λογαριθμικής πιθανοφάνειας για το πλήρες σύνολο δεδομένων είναι (Barbu and Limnios 2008):

$$L_M(Z_0^M, Y_0^M | \theta) = \mu(J_0, Y_0) \prod_{k=1}^{N(M)} [q_{J_{n-1}J_n}(X_k) \prod_{l=S_{k-1}}^{S_k-1} R_{J_{k-1}; Y_l}] \\ \times \overline{H}_{J_{N(M)}}(U_M) \prod_{l=S_{N(M)}}^M R_{J_{N(M)}; Y_l},$$

όπου $\mu(i, a)$ παριστά την αρχική κατανομή της κατάστασης $(i, a) \in E \times A$.

Ο κύριος στόχος είναι η εκτίμηση του θ αποκλειστικά και μόνο μέσω των παρατηρήσεων στο $[0, M]$. Υποθέτουμε ότι οι παρατηρήσεις δοθέντων των καταστάσεων της κρυφής SMC ακολουθούν κατανομή *Poisson* και στοχεύουμε στην εκτίμηση της παραμέτρου (ρυθμού) της κατανομής. Προκειμένου να εκτιμήσουμε τις παραμέτρους του μοντέλου, θ , εφαρμόζουμε τον αλγόριθμο EM (Bulla 2006, 2010). Θεωρώντας τρεις τύπους παρατηρήσεων (παρ. 3.1) και διδιάστατο χώρο καταστάσεων, η εκτιμώμενη κατανομή των χρόνων παραμονής στην κατάσταση $i \in E$, \hat{f}_i , παρουσιάζεται στον Πίνακα 2. Για παράδειγμα, η πιθανότητα ο χρόνος παραμονής στην πρώτη κρυφή κατάσταση να ισούται με ένα έτος, είναι ίση με 0.016. Σημειώνεται ότι από το σύνολο δεδομένων ο μέγιστος χρόνος παραμονής προκύπτει να είναι 12 έτη. Θεωρώντας ότι $P(i, j) = 0$, $i \neq j$, από τον επόμενο πίνακα προκύπτει άμεσα η εκτίμηση

του ημι-Μαρκοβιανού πυρήνα.

Η κατανομή των παρατηρήσεων δοθείσας της κατάστασης 1 είναι Poisson με εκτιμώμενο ρυθμό $\hat{\lambda}_1 = 1.739$, ενώ δοθείσας της κατάστασης 2 είναι $\hat{\lambda}_2 = 1.669$. Τέλος εφαρμόζοντας τον αλγόριθμο Viterbi μπορεί να αποκωδικοποιηθεί το σύνολο των κρυφών καταστάσεων της Η.Μ.Α.

Πίνακας 2: Εκτίμηση της κατανομής των χρόνων παραμονής, $\hat{f}_i, i \in E$.

Χρόνος (έτη)	1	2	3	4	5	6
Κατάσταση 1	0.016	0.023	0.039	0.049	0.050	0.044
Κατάσταση 2	0.079	0.147	0.366	0.086	0.049	0.021
Χρόνος (έτη)	7	8	9	10	11	12
Κατάσταση 1	0.043	0.053	0.083	0.125	0.187	0.288
Κατάσταση 2	0.027	0.039	0.088	0.077	0.018	0.003

ABSTRACT

A discrete-time hidden Markov model (HMM) is applied in order to reveal the actual stress field underlying the earthquake generation mechanism. The three basic problems of the theory of HMMs are presented and solved in order to determine the optimal model. The best HMM, in terms of the number of states, is selected through the criteria AIC and BIC. Simulated states and observations are derived for the selected model.

In order to relax the assumption of the geometric sojourn time distribution met in discrete time HMM, hidden semi-Markov models are applied. The distribution of sojourn times is estimated non-parametrically and Poisson distribution is selected as the emission time distribution with a rate parameter depending on the corresponding hidden state.

ΑΝΑΦΟΡΕΣ

- Akaike, H. (1974). A New Look at the Statistical Model Identification. *IEEE Trans. Autom. Control* **19**, 716-723.
- Barbu, V. and Limnios, N. (2008). Semi-Markov Chains and Hidden Semi-Markov Models toward Applications, Springer.
- Baum, L.E. and Petrie, T. (1966). Statistical Inference for probabilistic functions of finite state Markov chains. *Ann. Math. Stat.* **37**, 1554-1563.
- Bulla, J. (2006). Application of hidden Markov models and hidden semi-Markov models to financial time series. Thesis. Georg-August-Univer. Gottingen.
- Bulla, J. (2010). HSMM - an R package for analyzing hidden semi-Markov models. *Comp. Stat. and Data Anal.* **54(3)**, 611-619.
- Ebel, J.E., Chambers, D.W., Kafka, A.L. and Baglivo, J.A. (2007). Non-Poissonian earthquake clustering and the hidden Markov model as bases for earthquake forecasting in California. *Seismol. Res. Lett.* **78**, 57-65.
- Ephraim, Y. and Merhav, N. (2002). Hidden Markov Processes. *IEEE Trans. Inf. Theory* **48(6)**, 1518-1569.
- Ferguson, J. (1980). Variable duration models for speech. *Proc. Symp. Appl. HMMs Text Speech. Princeton, New Jersey*, 143-179.
- Granat, R. and Donnellan, A. (2002). A hidden Markov model based tool for geophysical data exploration. *Pure Appl. Geophys.* **159**, 2271-2283.
- Orfanogiannaki, K., Karlis, D. and Papadopoulos, G. (2010). Identifying seismicity levels via Poisson hidden Markov models. *Pure Appl. Geophys.* **167**, 919-931.
- Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Stat.* **6**, 461-464.



ΠΟΛΥΔΙΑΣΤΑΤΟ ΣΥΝΑΡΤΗΣΙΑΚΟ ΚΕΝΤΡΙΚΟ ΟΡΙΑΚΟ ΘΕΩΡΗΜΑ ΓΙΑ ΤΟΝ ΕΜΠΕΙΡΙΚΟ ΕΚΤΙΜΗΤΗ ΕΝΟΣ ΗΜΙ-ΜΑΡΚΟΒΙΑΝΟΥ ΠΥΡΗΝΑ

Σ. Γεωργιάδης, Ν. Λημνιός

Université de Technologie de Compiègne, France

{stylianos.georgiadis, nikolaos.limnios}@utc.fr

ΠΕΡΙΛΗΨΗ

Στην παρούσα εργασία, θεωρούμε τον εμπειρικό εκτιμητή του πυρήνα μίας ημι-Μαρκοβιανής ανελίξης σε συνεχή χρόνο με πεπερασμένο χώρο καταστάσεων και εξετάζουμε ένα πολυδιάστατο συναρτησιακό Κεντρικό Οριακό Θεώρημα (ΚΟΘ) για αυτόν τον εκτιμητή. Στη συνέχεια, παρουσιάζουμε το αντίστοιχο θεώρημα στην περίπτωση των ημι-Μαρκοβιανών ανελίξεων διακριτού χρόνου.

Λέξεις Κλειδιά: Ημι-Μαρκοβιανές ανελίξεις, συναρτησιακό κεντρικό οριακό θεώρημα, εμπειρικός εκτιμητής.

1 ΕΙΣΑΓΩΓΗ

Οι ημι-Μαρκοβιανές ανελίξεις είναι μία κατηγορία στοχαστικών ανελίξεων που γενικεύουν τις Μαρκοβιανές ανελίξεις άλματος και τις Μαρκοβιανές ανανεωτικές ανελίξεις (Pyke (1961a)). Αποτελούν ένα χρήσιμο εργαλείο στη στοχαστική μοντελοποίηση με ποικίλες εφαρμογές ειδικά οι ανελίξεις με πεπερασμένο χώρο καταστάσεων (Pyke (1961b)). Η μη-παραμετρική εμπειρική εκτίμηση για τον ημι-Μαρκοβιανό πυρήνα έχει παρουσιαστεί σε παλαιότερες εργασίες. Οι Gill (1980) και Ouhbi, Limnios (1996) έχουν μελετήσει τους μη-παραμετρικούς εκτιμητές του ημι-Μαρκοβιανού πυρήνα και τις ασυμπτωτικές ιδιότητές τους. Οι Limnios, Oprisan (2001) παρουσιάζουν μια συνολική μελέτη των ημι-Μαρκοβιανών ανελίξεων σε συνεχή χρόνο.

Το πολυδιάστατο συναρτησιακό ΚΟΘ ενός ημι-Μαρκοβιανού πυρήνα που παρουσιάζεται στη συγκεκριμένη εργασία γενικεύει τα αποτελέσματα που παρουσιάζονται σε παλαιότερη εργασία (Limnios (2004)), όπου οι καταστάσεις ήταν σταθεροποιημένες. Αυτή είναι μια ενδιαφέρουσα γενίκευση για μετέπειτα χρήση στην μη-παραμετρική εκτίμηση προβλημάτων συναρτησιακών του εμπειρικού εκτιμητή. Επιπλέον, επεκτείνουμε τα αποτελέσματα αυτά στις ημι-Μαρκοβιανές ανελίξεις διακριτού χρόνου.

2 ΗΜΙ-ΜΑΡΚΟΒΙΑΝΕΣ ΑΝΕΛΙΞΕΙΣ

Στο εξής, θα χρησιμοποιήσουμε την ακόλουθη σημειογραφία: $\mathbb{N}^* = \mathbb{N} \setminus \{0\}$, $\mathbb{R}_+ = [0, \infty)$ και $\mathbb{R}_+^* = (0, \infty)$.

Θεωρούμε ένα πεπερασμένο σύνολο E και μία ανέλιξη άλματος $(Z_t)_{t \in \mathbb{R}_+}$ με τιμές στο E . Έστω $(J_n)_{n \in \mathbb{N}}$ οι διαδοχικές καταστάσεις που επισκέφθηκε η (Z_t) με χώρο καταστάσεων E και $(S_n)_{n \in \mathbb{N}}$ οι χρόνοι άλματος της (Z_t) με τιμές στο \mathbb{R}_+ και $0 = S_0 \leq S_1 \leq \dots \leq S_n \leq S_{n+1} \leq \dots$. Επίσης, δηλώνουμε με $X_n := S_n - S_{n-1}$, $n \in \mathbb{N}^*$, τους χρόνους παραμονής σε αυτές τις καταστάσεις με τιμές στο \mathbb{R}_+ .

Η στοχαστική ανέλιξη $(J_n, S_n)_{n \in \mathbb{N}}$, με χώρο καταστάσεων E , καλείται Μαρκοβιανή ανανεωτική ανέλιξη (ΜΑΑ), εάν, για κάθε $j \in E$, $t \in \mathbb{R}_+$ και $n \in \mathbb{N}$, ικανοποιεί σ.β. την ακόλουθη ισότητα

$$\mathbb{P}(J_{n+1} = j, X_{n+1} \leq t | J_0, \dots, J_n; S_1, \dots, S_n) = \mathbb{P}(J_{n+1} = j, X_{n+1} \leq t | J_n).$$

Σε αυτή την περίπτωση, η (Z_t) καλείται ημι-Μαρκοβιανή ανέλιξη (ΗΜΑ).

Υποθέτουμε ότι η ΜΑΑ είναι χρονικά ομογενής, δηλαδή, η παραπάνω πιθανότητα είναι ανεξάρτητη των n και S_n . Η ανέλιξη (J_n) είναι μία Μαρκοβιανή αλυσίδα με χώρο καταστάσεων E , καλούμενη ενσωματωμένη Μαρκοβιανή αλυσίδα (ΕΜΑ) της (Z_t) . Δηλώνουμε με $N(t)$, $t \in \mathbb{R}_+$, την ανέλιξη που απαριθμεί τον αριθμό αλμάτων της (Z_t) στο διάστημα $(0, t]$, ορισμένη ως $N(t) = \sup\{n \geq 0 : S_n \leq t\}$. Η ΗΜΑ (Z_t) συνδέεται με την ΜΑΑ (J_n, S_n) από τη σχέση

$$Z_t := J_{N(t)}, \quad t \in \mathbb{R}_+.$$

Έστω $N_i(t)$ ο αριθμός των επισκέψεων της (Z_t) στην κατάσταση $i \in E$ έως το χρόνο t . Συγκεκριμένα,

$$N_i(t) := \sum_{k=1}^{N(t)} \mathbf{1}_{\{J_{k-1}=i\}}, \quad (1)$$

όπου $\mathbf{1}_{\{\cdot\}}$ είναι η δείκτρια συνάρτηση. Θεωρώντας την ανανεωτική ανέλιξη $(S_n^i)_{n \in \mathbb{N}}$ των διαδοχικών επισκέψεων στην κατάσταση i , μ_{ii} είναι ο μέσος χρόνος επαναφοράς της (S_n^i) , δηλαδή, $\mu_{ii} := \mathbb{E}[S_2^i - S_1^i]$.

Ο πυρήνας μετάβασης $Q(t) := (Q_{ij}(t); i, j \in E)$, $t \in \mathbb{R}_+$, καλείται ημι-Μαρκοβιανός πυρήνας (ΗΜΠ) της ΗΜΑ (Z_t) και ορίζεται ως

$$Q_{ij}(t) := \mathbb{P}(J_{n+1} = j, X_{n+1} \leq t | J_n = i), \quad t \in \mathbb{R}_+, \quad i, j \in E. \quad (2)$$

Δηλώνουμε με $\nu = (\nu_i; i \in E)$, τη στάσιμη κατανομή της ΕΜΑ (J_n) . Έστω, επίσης, ο μέσος χρόνος παραμονής της (Z_t) , $\bar{m} := \sum_{i \in E} \nu_i m_i$, όπου m_i ο μέσος χρόνος παραμονής στην κατάσταση $i \in E$. Υπενθυμίζουμε την ιδιότητα $\mu_{ii} = m/\nu_i$. Καθ'ολη την παρούσα εργασία, υποθέτουμε ότι η ΕΜΑ (J_n) είναι μη-υποβιβάσιμη και γνήσια επαναληπτική, δηλαδή, $\bar{m} < \infty$.

3 ΠΟΛΥΔΙΑΣΤΑΤΟ ΣΥΝΑΡΤΗΣΙΑΚΟ ΚΟΘ

Θεωρούμε το πεπερασμένο σύνολο $E = \{1, \dots, d\}$, $d \in \mathbb{N}^*$, και μία παρατήρηση μιας ΗΜΑ (Z_t) , με χώρο καταστάσεων E , στο διάστημα $[0, t]$, $t \in \mathbb{R}_+$ σταθερό, με $\mathcal{H}_t := \{Z_u, 0 \leq u \leq t\} = \{J_0, X_1, \dots, J_{N(t)-1}, X_{N(t)}, J_{N(t)}, U_t\}$, όπου $U_t := t - S_{N(t)}$ ο χρόνος οπισθοδρομικής επαναφοράς.

Ο εμπειρικός εκτιμητής $\hat{Q}(x, t) := (\hat{Q}_{ij}(x, t); i, j \in E)$, $t \in \mathbb{R}_+$, του ΗΜΠ $Q(x)$ ορίζεται από τη σχέση

$$\hat{Q}_{ij}(x, t) := \frac{1}{N_i(t)} \sum_{k=1}^{N(t)} \mathbf{1}_{\{J_{k-1}=i, J_k=j, X_k \leq x\}}, \quad 0 \leq x \leq t, \quad i, j \in E. \quad (3)$$

Σημειώνουμε ότι οι χρόνοι οπισθοδρομικής επαναφοράς παραλείπονται από τους εμπειρικούς εκτιμητές καθώς όταν το t τείνει στο άπειρο, ο χρόνος U_t δεν προσθέτει επιπλέον πληροφορία στην εκτίμηση.

Θεωρούμε το χώρο πιθανοτήτων $(\Omega, \mathcal{F}, \mathbb{P})$ και τη διύλιση διακριτού χρόνου $\mathcal{F}_k = \sigma(J_0, J_i, X_i; 1 \leq i \leq k)$, $k \in \mathbb{N}^*$. Για όλες τις καταστάσεις $i, j \in E$, κάθε σταθεροποιημένο χρόνο $x \in \mathbb{R}_+$, θέτουμε $Y_k = (Y_k^{ij}; i, j \in E) \in \mathbb{R}^{d^2}$, όπου $(Y_k^{ij})_{k \in \mathbb{N}^*}$ είναι τυχαίες ακολουθίες ορισμένες ως

$$Y_k^{ij} := \mathbf{1}_{\{J_{k-1}=i, J_k=j, X_k \leq x\}} - \mathbf{1}_{\{J_{k-1}=i\}} Q_{ij}(x), \quad k \in \mathbb{N}^*. \quad (4)$$

Στην περίπτωση που εξετάζουμε στην παρούσα εργασία, υποθέτουμε ότι

$$Y_k^n := \frac{Y_k}{\sqrt{n}}, \quad k, n \in \mathbb{N}^*. \quad (5)$$

Επιπλέον, για κάθε σταθεροποιημένο $t \in \mathbb{R}_+$, ορίζουμε τις στοχαστικές ανελίξεις

$$S_t^{n,ij} := \sum_{k=1}^{N(nt)} Y_k^{n,ij} = \frac{1}{\sqrt{n}} \sum_{k=1}^{N(nt)} Y_k^{ij}, \quad n \in \mathbb{N}^*. \quad (6)$$

Χρησιμοποιώντας την παραπάνω σημειογραφία, έχουμε την ακολουθία στοχαστικών ανελίξεων

$$S_t^n := (S_t^{n,ij}; i, j \in E), \quad t \in \mathbb{R}_+, \quad n \in \mathbb{N}^*. \quad (7)$$

Τα αποτελέσματά μας ισχύουν στο χώρο Skorohod $D[0, \infty)$. Δηλώνουμε με $\xrightarrow{\sigma.\beta.}$ τη σχεδόν βεβαία σύγκλιση, με \Rightarrow την ασθενή σύγκλιση στο χώρο αυτόν και με \xrightarrow{D} τη σύγκλιση κατά κατανομή μιας ακολουθίας τυχαίων μεταβλητών. Επιπλέον, W_t είναι μία d^2 -διάστατη τυποποιημένη Wiener ανελίξη.

Παραθέτουμε τα επόμενα Λήμματα, τα οποία αποτελούν ουσιαστικά και την απόδειξη των κύριων αποτελεσμάτων.

Λήμμα 3.1. *Οι τυχαίες ακολουθίες $(Y_k^n)_{k \in \mathbb{N}^*}$, $n \in \mathbb{N}^*$, όπως ορίστηκαν από τις (4) και (5), είναι martingale διαφορές, δηλαδή $\mathbb{E}[Y_k^n | \mathcal{F}_{k-1}] = 0$, για κάθε $n \in \mathbb{N}^*$. Συνεπώς, η ακολουθία S_t^n , σύμφωνα με τις (6) και (7), είναι ένα martingale.*

Απόδειξη. Για κάθε $i, j \in E$ και $x \in \mathbb{R}_+$ σταθεροποιημένο, μπορούμε εύκολα να δείξουμε ότι $\mathbb{E}[Y_k | \mathcal{F}_{k-1}] = 0$. Οπότε, όλα τα στοιχεία της $(Y_k)_{k \in \mathbb{N}^*}$ είναι martingale διαφορές και συνεπώς $(Y_k^n)_{k \in \mathbb{N}^*}$, $n \in \mathbb{N}^*$, είναι μία martingale διαφορά. Βάσει ορισμού, το δεύτερο μέρος της πρότασης επάγεται άμεσα. \square

Λήμμα 3.2. (Limnios (2004)) *Εάν η EMA (J_n) είναι γνήσια επαναληπτική με πεπερασμένη στάσιμη κατανομή ν , τότε, για κάθε $i, j \in E$, έχουμε*

1. $\frac{N_i(t)}{t} \xrightarrow{\sigma, \beta} \frac{1}{\mu_{ii}}$,
2. $\frac{N(t)}{t} \xrightarrow{\sigma, \beta} \frac{1}{\bar{m}}$,

καθώς $t \rightarrow \infty$.

Λήμμα 3.3. *Για τη martingale διαφορά $(Y_k^n)_{k \in \mathbb{N}^*}$, $n \in \mathbb{N}^*$, ισχύει η υπό όρους Lindeberg συνθήκη, δηλαδή, για κάθε $t \in \mathbb{R}_+^*$,*

$$\frac{1}{n} \sum_{k=1}^{N(nt)} \mathbb{E}[\|Y_k\|^2 \mathbf{1}_{\{\|Y_k\| > \epsilon \sqrt{n}\}} | \mathcal{F}_{k-1}] \xrightarrow{\mathbf{P}} 0, \quad (8)$$

καθώς $n \rightarrow \infty$, όπου $\|\cdot\|$ μία νόρμα στον \mathbb{R}^{d^2} .

Απόδειξη. Για κάθε $i, j \in E$ και κάθε $k \in \mathbb{N}^*$, $|Y_k^{ij}| \leq 1$. Τότε, $\|Y_k\| \leq d^2$. Χρησιμοποιώντας την ανισότητα Markov, για κάθε $\epsilon > 0$, έχουμε

$$\frac{1}{n} \sum_{k=1}^{N(nt)} \mathbb{E}[\|Y_k\|^2 \mathbf{1}_{\{\|Y_k\| > \epsilon \sqrt{n}\}} | \mathcal{F}_{k-1}] \leq \frac{d^6}{\epsilon n \sqrt{n}} N(nt) \xrightarrow{\sigma, \beta} 0,$$

για κάθε $t \in \mathbb{R}_+^*$, καθώς $n \rightarrow \infty$, εφόσον από Λήμμα 3.2, $N(nt)/nt \xrightarrow{\sigma, \beta} 1/\bar{m}$. \square

Λήμμα 3.4. *Για κάθε $t \in \mathbb{R}_+^*$, η ακόλουθη σύγκλιση ισχύει για τη martingale διαφορά $(Y_k^n)_{k \in \mathbb{N}^*}$, $n \in \mathbb{N}^*$,*

$$\frac{1}{n} \sum_{k=1}^{N(nt)} \mathbb{E}[Y_k^{ij} Y_k^{lr} | \mathcal{F}_{k-1}] \xrightarrow{\sigma, \beta} t C^{ij,lr}, \quad (9)$$

καθώς $n \rightarrow \infty$, όπου $C^{ij,lr} = \delta_{il} \frac{1}{\mu_{ii}} Q_{ij}(x) (\delta_{ir} - Q_{ir}(x))$, $(i, j), (l, r) \in E \times E$.

Απόδειξη. Διακρίνουμε τις ακόλουθες περιπτώσεις:

1. Για $i \neq l$, έχουμε $\mathbb{E}[Y_k^{ij} Y_k^{lr} | \mathcal{F}_{k-1}] = 0$.
2. Για $i = l$ και $j \neq r$, για κάθε $x \in [0, nt]$,

$$\mathbb{E}[Y_k^{ij} Y_k^{lr} | \mathcal{F}_{k-1}] = -\mathbf{1}_{\{J_{k-1}=i\}} Q_{ij}(x) Q_{ir}(x).$$

3. Για $i = l$ και $j = r$, για κάθε $x \in [0, nt]$,

$$\mathbb{E}[Y_k^{ij} Y_k^{lr} | \mathcal{F}_{k-1}] = \mathbf{1}_{\{J_{k-1}=i\}} Q_{ij}(x)(1 - Q_{ij}(x))$$

Από τη σχέση (1), το Λήμμα 3.2 και το Λήμμα Slutsky, καταλήγουμε ότι

$$C^{ij,lr} = \begin{cases} \frac{1}{\mu_{ii}} Q_{ij}(x)(1 - Q_{ij}(x)) & , \text{αν } i = l \text{ και } j = r \\ -\frac{1}{\mu_{ii}} Q_{ij}(x) Q_{ir}(x) & , \text{αν } i = l \text{ και } j \neq r \\ 0 & , \text{αν } i \neq l. \end{cases}$$

□

Δηλώνουμε με C τον $d^2 \times d^2$ -διάστατο συμμετρικό πίνακα $(C^{ij,lr}; (i, j), (l, r) \in E \times E)$.

Λήμμα 3.5. *Ο πίνακας C είναι ένας θετικά ημι-ορισμένος πίνακας. Ως εκ τούτου, ο C είναι ένας πίνακας συνδιακύμανσης.*

Απόδειξη. Έστω ένα διάνυσμα $z^\top = [z_{11} z_{12} \cdots z_{1d} z_{21} z_{22} \cdots z_{2d} \cdots z_{d1} z_{d2} \cdots z_{dd}] \in \mathbb{R}^{d^2}$. Αρχεί να δείξουμε ότι $z^\top C z \geq 0$, ώστε ο C να είναι ένας πίνακας συνδιακύμανσης. □

Θεώρημα 1. *Έστω $S_t^n, t \in \mathbb{R}_+^*, n \in \mathbb{N}^*$, οι τυχαίες ακολουθίες όπως ορίστηκαν στις (6) και (7). Τότε, ισχύει :*

$$S_t^n \Rightarrow S_t, \quad (10)$$

καθώς $n \rightarrow \infty$, όπου S_t είναι ένα d^2 -διάστατο συνεχές Γκαουσιανό martingale με προβλέψιμα χαρακτηριστικά $(0, t C, 0)$.

Απόδειξη. Ακολουθώντας τα Λήμματα 3.1, 3.3, 3.4 και 3.5, και από το VIII Theorem 3.33, Jacod, Shiryaev (2003), το τελικό αποτέλεσμα ισχύει. □

Παρατήρηση. Ένα Γκαουσιανό martingale με προβλέψιμα χαρακτηριστικά $(0, t C, 0)$ είναι μία Wiener ανέλιξη (II Theorem 4.36, Jacod, Shiryaev (2003)). Επιπλέον, η συνάρτηση διακύμανσής της είναι $\sigma^2(x, t) = \langle W, W \rangle_t = t C$. Συνεπώς, μπορούμε να γράψουμε $S_t^n \Rightarrow C^{1/2} W_t$, καθώς $n \rightarrow \infty$.

Για κάθε $i, j \in E$ και $x \in [0, nt], t \in \mathbb{R}_+^*, n \in \mathbb{N}^*$, θέτουμε $\Delta Q_{ij}(x, nt) := \hat{Q}_{ij}(x, nt) - Q_{ij}(x)$. Πλέον, μπορούμε να προχωρήσουμε στο πολυδιάστατο συναρτησιακό ΚΟΘ για τον εκτιμητή (3), όπως εκφράζεται από το ακόλουθο Θεώρημα.

Θεώρημα 2. *Για κάθε $i, j \in E$ και $x \in [0, nt], t \in \mathbb{R}_+^*, n \in \mathbb{N}^*$, ισχύει η ασθενής σύγκλιση :*

$$(\sqrt{n} \Delta Q_{ij}(x, nt); i, j \in E) \Rightarrow (\mathcal{G}_{t,ij}; i, j \in E), \quad (11)$$

καθώς $n \rightarrow \infty$, όπου $\mathcal{G}_{t,ij} = \mu_{ii} C^{1/2} \frac{W_t}{t}$ είναι μία συνεχής Γκαουσιανή ανέλιξη.

Απόδειξη. Οι ανελίξεις $(\sqrt{n}\Delta Q_{ij}(x, nt); i, j \in E)$, $t \in \mathbb{R}_+^*$, $n \in \mathbb{N}^*$, μπορούν να γραφούν

$$(\sqrt{n}\Delta Q_{ij}(x, nt); i, j \in E) = \left(\sqrt{n} \frac{1}{N_i(nt)} \sum_{k=1}^{N(nt)} Y_k^{ij}; i, j \in E \right) = \frac{n}{N_i(nt)} S_t^n.$$

Λαμβάνοντας υπόψιν το Θεώρημα 1, το Λήμμα 3.2 και το Λήμμα Slutsky, για κάθε $t \in \mathbb{R}_+^*$, συμπεραίνουμε ότι

$$(\sqrt{n}\Delta Q_{ij}(x, nt); i, j \in E) \Rightarrow (\mu_{ii} C^{1/2} \frac{W_t}{t}; i, j \in E),$$

καθώς $n \rightarrow \infty$. Εύκολα διαπιστώνουμε ότι, για $t \in \mathbb{R}_+^*$, $\mathcal{G}_{t,ij} = \mu_{ii} C^{1/2} \frac{W_t}{t}$ είναι μία συνεχής Γκαουσιανή ανελίξη. \square

Πόρισμα 3.1. Για κάθε $i, j \in E$ και $x \in [0, t]$, $t \in \mathbb{R}_+^*$, έχουμε

$$(\sqrt{n}\Delta Q_{ij}(x, t); i, j \in E) \xrightarrow{\mathcal{D}} \mathcal{N}(\mathbf{0}, \Sigma_Q), \quad (12)$$

καθώς $t \rightarrow \infty$, όπου $\mathcal{N}(\mathbf{0}, \Sigma)$ είναι μία d^2 -διάστατη κανονική τυχαία μεταβλητή με μέση τιμή $\mathbf{0}$ και πίνακα συνδιακύμανσης $\Sigma_Q = (\delta_{il} \mu_{ii} Q_{ij}(x) (\delta_{ir} - Q_{ir}(x)))_{(i,j),(l,r) \in E \times E}$.

Απόδειξη. Θέτοντας $t = 1$, έχουμε ότι $W_1 \sim \mathcal{N}(0, 1)$ και αντικαθιστώντας το n με t , ισχύει $\mathcal{G}_{1,ij} = \mu_{ii} C^{1/2} W_1$ και έχουμε το τελικό αποτέλεσμα. \square

4 ΗΜΙ-ΜΑΡΚΟΒΙΑΝΕΣ ΑΛΥΣΙΔΕΣ

Στην παράγραφο αυτή, παρουσιάζουμε τις ημι-Μαρκοβιανές αλυσίδες, οι οποίες είναι οι ΗΜΑ σε διακριτό χρόνο, και το αντίστοιχο πολυδιάστατο συναρτησιακό ΚΟΘ για τον ΗΜΠ τους.

Γενικά, μπορούμε να χειριστούμε τις ημι-Μαρκοβιανές αλυσίδες πιο εύκολα σε σχέση με τις ανελίξεις συνεχούς χρόνου. Μία ημι-Μαρκοβιανή αλυσίδα πραγματοποιεί μόνο ένα φραγμένο αριθμό μεταβάσεων σε ένα πεπερασμένο διάστημα. Το γεγονός αυτό ευνοεί πιο ακριβή αριθμητικά αποτελέσματα στις εφαρμογές. Μια λεπτομερής παρουσίαση των διακριτού-χρόνου ΗΜΑ δίνεται από τους Barbu, Limnios (2008).

Έστω η διακριτού-χρόνου ΜΑΑ $(J_n, S_n)_{n \in \mathbb{N}}$ με χώρο καταστάσεων το πεπερασμένο σύνολο $E = \{1, \dots, d\}$. Οι ανελίξεις (S_n) και X_n , τώρα, λαμβάνουν τιμές στο \mathbb{N} . Ορίζουμε τις ανελίξεις απαρίθμησης $N(k) := \max\{n \geq 0 : S_n \leq k\}$, $k \in \mathbb{N}^*$, και

$$N_i(k) := \sum_{n=1}^{N(k)} \mathbf{1}_{\{J_{n-1}=i\}}. \quad (13)$$

Η ημι-Μαρκοβιανή αλυσίδα $(Z_k)_{k \in \mathbb{N}}$, τότε, ορίζεται ως

$$Z_k := J_{N(k)}, k \in \mathbb{N},$$

η οποία είναι μία ομογενής αλυσίδα με τιμές στο E . Ο διακριτού-χρόνου ΗΜΠ $q(k) := (q_{ij}(k); i, j \in E)$, $k \in \mathbb{N}$, ορίζεται ως

$$q_{ij}(k) := \mathbb{P}(J_{n+1} = j, X_{n+1} = k | J_n = i), k \in \mathbb{N}, i, j \in E. \quad (14)$$

Η παρατήρηση στο διάστημα $[0, k]$, $k \in \mathbb{N}$, δίνεται πλέον $\mathcal{H}_k := \{Z_u, 0 \leq u \leq k\}$. Ο εμπειρικός εκτιμητής $\hat{q}(x, t) := (\hat{q}_{ij}(x, t); i, j \in E)$, $k \in \mathbb{N}^*$, του διακριτού-χρόνου ΗΜΠ $q(x)$ ορίζεται από τη σχέση

$$\hat{q}_{ij}(x, k) := \frac{1}{N_i(k)} \sum_{n=1}^{N(k)} \mathbf{1}_{\{J_{n-1}=i, J_n=j, X_n=x\}}, 0 \leq x \leq k, i, j \in E. \quad (15)$$

Δηλώνουμε με $\lfloor \alpha \rfloor$, το ακέραιο μέρος ενός θετικού πραγματικού αριθμού α . Για όλες τις καταστάσεις $i, j \in E$, κάθε σταθεροποιημένο χρόνο $x \in \mathbb{N}$, θέτουμε $Y_m := (Y_m^{ij}; i, j \in E) \in \mathbb{R}^{d^2}$, όπου $(Y_m^{ij})_{m \in \mathbb{N}^*}$ είναι τυχαίες ακολουθίες ορισμένες ως

$$Y_m^{ij} := \mathbf{1}_{\{J_{m-1}=i, J_m=j, X_m=x\}} - \mathbf{1}_{\{J_{m-1}=i\}} q_{ij}(x), m \in \mathbb{N}^*. \quad (16)$$

Στην περίπτωση του διακριτού χρόνου, υποθέτουμε τις ίδιες συνθήκες της Παραγράφου 2. Δηλαδή,

$$Y_m^n := \frac{Y_m}{\sqrt{n}}, m, n \in \mathbb{N}^* \quad (17)$$

και, για κάθε σταθεροποιημένο $t \in \mathbb{R}_+^*$,

$$S_t^{n,ij} := \sum_{m=1}^{N(\lfloor nt \rfloor)} Y_m^{n,ij} = \frac{1}{\sqrt{n}} \sum_{m=1}^{N(\lfloor nt \rfloor)} Y_m^{ij}, n \in \mathbb{N}^*, \quad (18)$$

και

$$S_t^n := (S_t^{n,ij}; i, j \in E), n \in \mathbb{N}^*. \quad (19)$$

Μπορούμε εύκολα να επαληθεύσουμε ότι τα Λήμματα 3.1-3.5 ισχύουν για τις ημι-Μαρκοβιανές αλυσίδες. Επιπλέον, η αντίστοιχη σύγκλιση του Λήμματος 3.4 δίνεται από τη σχέση $\mathcal{C}^{ij,lr} = \delta_{il} \frac{1}{\mu_{ii}} q_{ij}(x) (\delta_{ir} - q_{ir}(x))$, $(i, j), (l, r) \in E \times E$.

Θεώρημα 3. Έστω $(Y_m^n)_{m \in \mathbb{N}^*}$, $n \in \mathbb{N}^*$, οι *martingale* διαφορές και S_t^n , $t \in \mathbb{R}_+^*$, $n \in \mathbb{N}^*$, οι *στοχαστικές ανελίξεις* όπως ορίστηκαν στις (16), (17) και (18), (19), αντιστοίχως. Τότε

$$S_t^n \Rightarrow S_t, \quad (20)$$

καθώς $n \rightarrow \infty$, όπου S_t είναι ένα d^2 -διάστατο συνεχές Γκαουσιανό *martingale* με προβλέψιμα χαρακτηριστικά $(0, t \mathcal{C}, 0)$.

Το πολυδιάστατο συναρτησιακό ΚΟΘ στο διακριτό χρόνο ακολουθεί :

Θεώρημα 4. Για κάθε $i, j \in E$ και $x \in [0, [nt]]$, $t \in \mathbb{R}_+^*$, $n \in \mathbb{N}^*$, ισχύει η σύγκλιση

$$(\sqrt{n}\Delta q_{ij}(x, [nt]); i, j \in E) \Rightarrow (\mathcal{G}_{t,ij}; i, j \in E), \quad (21)$$

καθώς $n \rightarrow \infty$, όπου $\mathcal{G}_{t,ij} = \mu_{ii}C^{1/2}\frac{W_t}{t}$ είναι μία συνεχής Γκαουσιανή ανέλιξη.

Πόρισμα 4.1. Για κάθε $i, j \in E$ και $x \in [0, [t]]$, $t \in \mathbb{R}_+^*$, έχουμε

$$(\sqrt{n}\Delta q_{ij}(x, t); i, j \in E) \xrightarrow{\mathcal{D}} \mathcal{N}(\mathbf{0}, \Sigma_q), \quad (22)$$

καθώς $t \rightarrow \infty$, όπου $\mathcal{N}(\mathbf{0}, \Sigma)$ είναι μία d^2 -διάστατη κανονική τυχαία μεταβλητή με μέση τιμή $\mathbf{0}$ και πίνακα συνδιακύμανσης $\Sigma_q = (\delta_{il}\mu_{ii}q_{ij}(x)(\delta_{ir} - q_{ir}(x)))_{(i,j),(l,r) \in E \times E}$.

ABSTRACT

In the present work, we consider the empirical estimator of the kernel of a semi-Markov process on continuous time with finite state space and we examine a multidimensional functional central limit theorem for this estimator. Next, we present the corresponding theorem in the case of discrete-time semi-Markov processes.

Αναφορές

- Barbu, V. S. and Limnios, N. (2008). *Semi-Markov Chains and Hidden Semi-Markov Models toward Applications*, Springer.
- Gill, R.D. (1980). Nonparametric estimation based on censored observations of a markov renewal process. *Z. Wahrscheinlichkeitstheorie verw. Gebiete*, **53**, 97-116.
- Jacod, J. and Shiryaev, A. (2003). *Limit Theorems for Stochastic Processes*, 2nd Edition, Springer.
- Limnios, N. (2004). A functional central limit theorem for the empirical estimator of a semi-Markov kernel. *Nonparametric Stat.* **16**(1-2), 13-18.
- Limnios, N. and Oprisan, G. (2001). *Semi-Markov Processes and Reliability*, Birkhäuser, Boston.
- Ouhbi, B. and Limnios, N. (1996). Non-parametric estimation for semi-Markov kernels with application to reliability analysis. *Applied Stochastic Models and Data Analysis*, **12**, 209-220
- Pyke, R. (1961a). Markov renewal processes: definitions and preliminary properties. *Annals of Mathematical Statistics*, **32**, 1231-1242
- Pyke, R. (1961b). Markov renewal processes with finitely many states. *Annals of Mathematical Statistics*, **32**, 1243-1259.



ΕΚΤΙΜΗΣΗ ΤΗΣ ΑΠΟ ΚΟΙΝΟΥ ΚΑΤΑΝΟΜΗΣ ΣΤΗΘΙΑΙΑΣ ΔΙΑΜΕΤΡΟΥ ΚΑΙ ΥΨΟΥΣ ΔΕΝΔΡΩΝ ΜΕ ΤΗ ΧΡΗΣΗ ΣΥΖΕΥΞΕΩΝ (COPULAS)

Α. Ν. Γκιούλος¹, Γ. Τσακλίδης¹, Γ. Σταματέλλος²

¹Τμήμα Μαθηματικών Α.Π.Θ.

tsaklidi@math.auth.gr

²Τμήμα Δασολογίας και Φυσικού Περιβάλλοντος Α.Π.Θ.

stamatel@for.auth.gr

ΠΕΡΙΛΗΨΗ

Στην παρούσα εργασία προσδιορίζονται διδιάστατες κατανομές με τη βοήθεια των συζεύξεων (copulas), οι οποίες να μπορούν με επιτυχία να χρησιμοποιηθούν για την περιγραφή της από κοινού κατανομής της διαμέτρου και του ύψους δέντρων υβριδογενούς ελάτης του Δάσους Περτουλίου. Η γνώση της κοινής συνάρτησης κατανομής μπορεί να βοηθήσει στην εκτίμηση μεγεθών που έχουν ιδιαίτερο πρακτικό ενδιαφέρον, όπως η εκτίμηση του όγκου δασικής ξυλείας.

Λέξεις Κλειδιά: συζεύξεις, διδιάστατες κατανομές, υβριδογενής ελάτη

1. ΣΥΖΕΥΞΕΙΣ. ΒΑΣΙΚΕΣ ΕΝΝΟΙΕΣ

Οι συζεύξεις είναι συναρτήσεις, που συνδέουν την πολυδιάστατη συνάρτηση κατανομής (σ.κ.) με τις μονοδιάστατες περιθώριες συναρτήσεις. Επίσης, οι συζεύξεις μπορούν να θεωρηθούν ως πολυδιάστατες σ.κ., των οποίων οι περιθώριες είναι ομοιόμορφες κατανομές στο διάστημα $[0,1]$.

Προκειμένου να ορίσουμε τη σύζευξη, συμβολίζουμε με $I=[0,1]$, $R = (-\infty, \infty)$, και S_1, S_2 μη κενά υποσύνολα του R . Τότε (Nelsen, 2005):

Ορισμός 1.1

Η συνάρτηση $C : I^2 \rightarrow I$ καλείται *σύζευξη* αν ισχύουν οι ακόλουθες ιδιότητες:

1. Για κάθε u, v στο I ,

$$C(u, 0) = 0 = C(0, v) \text{ και } C(u, 1) = u, C(1, v) = v.$$

2. Για κάθε u_1, u_2, v_1, v_2 του I έτσι ώστε $u_1 \leq u_2$ και $v_1 \leq v_2$,

$$C(u_2, v_2) - C(u_2, v_1) - C(u_1, v_2) + C(u_1, v_1) \geq 0.$$

Το Θεώρημα που ακολουθεί είναι το κεντρικό θεώρημα της θεωρίας των συζεύξεων, καθώς και η βάση για τις περισσότερες εφαρμογές των συζεύξεων στη στατιστική.

Θεώρημα 1.1 (Θεώρημα του Sklar)

Ας είναι X και Y τυχαίες μεταβλητές με από κοινού συνάρτηση κατανομής H και περιθώριες σ.κ. F και G , αντίστοιχα. Τότε, υπάρχει σύζευξη C , τέτοια ώστε

$$H(x, y) = C(F(x), G(y)), \quad (1)$$

για όλα τα $x, y \in \mathbb{R}$.

Αν F και G είναι συνεχείς, τότε η C είναι μοναδική. Διαφορετικά, η σύζευξη C είναι μονότιμα ορισμένη στο σύνολο $Ran(F) \times Ran(G)$, όπου με Ran συμβολίζουμε το πεδίο τιμών.

Αντίστροφα, αν C είναι μια σύζευξη, και F, G συναρτήσεις κατανομών, τότε η συνάρτηση H , που ορίζεται με την (1), είναι η από κοινού σ.κ. των τυχαίων μεταβλητών X και Y με περιθώριες τις F και G .

Οι *Αρχιμήδειες συζεύξεις* αποτελούν μία από τις πιο γνωστές και πιο σημαντικές κλάσεις συζεύξεων (Nelsen, 2006).

Για τη συνέχεια είναι απαραίτητος ο ορισμός της ψευδο-αντίστροφης συνάρτησης: Αν είναι $\phi_\theta : I \rightarrow [0, \infty)$ μια συνεχής και αυστηρά φθίνουσα συνάρτηση έτσι ώστε $\phi_\theta(1) = 0$, τότε η *ψευδο-αντίστροφη* συνάρτηση της ϕ_θ είναι μια συνάρτηση $\phi_\theta^{[-1]}$ με $Dom \phi_\theta^{[-1]} = [0, \infty)$ και $Ran \phi_\theta^{[-1]} = I$, που δίνεται από τη σχέση

$$\phi_\theta^{[-1]} = \begin{cases} \phi_\theta^{-1}(t), & 0 \leq t \leq \phi_\theta(0) \\ 0, & \phi_\theta(0) \leq t \leq \infty \end{cases}.$$

Ορισμός 1.2

Έστω ϕ_θ μία συνεχής και αυστηρά φθίνουσα συνάρτηση με πεδίο ορισμού το σύνολο I και πεδίο τιμών το $[0, \infty)$, έτσι ώστε $\phi_\theta(1) = 0$, και ϕ_θ κυρτή. Επίσης, έστω $\phi_\theta^{[-1]}$ η *ψευδο-αντίστροφη* συνάρτηση της ϕ_θ . Τότε, η συνάρτηση $C : I^2 \rightarrow I$, η οποία ορίζεται ως

$$C(u, v) = \phi_\theta^{[-1]}(\phi_\theta(u) + \phi_\theta(v)), \quad \theta \in \mathbb{R},$$

ονομάζεται *Αρχιμήδεια σύζευξη*.

2. ΠΡΟΣΕΓΓΙΣΗ ΤΩΝ ΠΡΑΓΜΑΤΙΚΩΝ ΔΕΔΟΜΕΝΩΝ ΜΕ ΒΑΣΗ ΤΙΣ ΑΡΧΙΜΗΔΕΙΕΣ ΣΥΖΕΥΞΕΙΣ

Η προσέγγιση των πραγματικών δεδομένων από μια Αρχιμήδεια σύζευξη βασίζεται στην παρακάτω κατασκευή (Genest and Rivest, 1993, και Aderman and Pihlak, 2005) (για το πρόβλημα της εκτίμησης των παραμέτρων στην μοντελοποίηση με τη βοήθεια των συζεύξεων βλ. και Joe, 1997).

Έστω ένα τυχαίο δείγμα από παρατηρήσεις: $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$. Τότε, ορίζουμε τις τυχαίες μεταβλητές (τ.μ.) Z_i ως

$$Z_i = \frac{\#\{(x_j, y_j) : x_j < x_i \& y_j < y_i\}}{n-1}, \quad i=1, 2, \dots, n.$$

Από ένα δείγμα (x_i, y_i) , $i=1, 2, \dots, n$, παίρνουμε τις παρατηρήσεις z_1, z_2, \dots, z_n ως τιμές για τις ανεξάρτητες και ομοιόμορφα κατανεμημένες τ.μ. Z_1, Z_2, \dots, Z_n , και κατασκευάζουμε την εμπειρική συνάρτηση

$$K_n(z) = \frac{\#\{z_i : z_i < z\}}{n}.$$

Έπειτα, ορίζουμε τη θεωρητική συνάρτηση κατανομής μιας Αρχιμήδειας σύζευξης ως συνάρτηση της ϕ_θ από τη σχέση

$$K(z) = z - \frac{\phi_\theta(z)}{\phi'_\theta(z)}.$$

Στη συνέχεια, με τη βοήθεια του Kolmogorov-Smirnov στατιστικού τέστ με

$$D = \max |K_n(z) - K(z)|,$$

όπου n είναι το μέγεθος του δείγματος, μπορούμε να ελέγξουμε κατά πόσο η σύζευξη $C(u, v)$ προσεγγίζει την εμπειρική συνάρτηση κατανομής για τις τ.μ. X, Y .

3. ΕΦΑΡΜΟΓΗ. ΕΚΤΙΜΗΣΗ ΤΗΣ ΑΠΟ ΚΟΙΝΟΥ ΚΑΤΑΝΟΜΗΣ ΣΤΗΘΙΑΙΑΣ ΔΙΑΜΕΤΡΟΥ ΚΑΙ ΥΨΟΥΣ ΤΩΝ ΔΕΝΤΡΩΝ

Τα δεδομένα προέρχονται από το Πανεπιστημιακό Δάσος Πετρουλίου, το οποίο βρίσκεται στην Κεντρική Πίνδο (Ν. Τρικάλων). Το δάσος καλύπτεται σχεδόν εξ ολοκλήρου από υβριδογενή ελάτη και είναι ανομήλικο. Η συλλογή των στοιχείων, που έγινε για τις ανάγκες άλλης έρευνας. (Σταματέλλος, 1991, 1995) πραγματοποιήθηκε ως εξής : Σε χάρτη της περιοχής έρευνας εφαρμόστηκε πλέγμα οριζόντιων και κάθετων γραμμών με τα σημεία τομής να αποτελούν τα κέντρα των επιφανειών

δειγματοληψίας $((20 \times 25)m^2)$. Η συλλογή των δεδομένων έγινε με απλή τυχαία δειγματοληψία με επανάθεση. Τα τελικά 48 σύνολα δεδομένων προέκυψαν με ενώσεις ανά 4 γειτονικές επιφάνειες από τις αρχικές 192 επιφάνειες.

Στην παράγραφο αυτή παρουσιάζονται τα αποτελέσματα της μοντελοποίησης της από κοινού κατανομής της διαμέτρου X του κορμού των δένδρων, και του ύψους τους, Y . Αν f, g οι συναρτήσεις πυκνότητας (σ.π.) των τ.μ. X, Y αντίστοιχα και F, G οι αντίστοιχες σ.κ., τότε σύμφωνα με το θεώρημα του Sklar υπάρχει σύζευξη $C(F(x), G(y))$ για την από κοινού σ.κ..

Προκειμένου να εξετάσουμε την προσαρμογή αρχιμήδειων συζεύξεων στα δεδομένα, αρχικά μετασχηματίζουμε τα δεδομένα με τους μετασχηματισμούς (Karian et al., 1996),

$$X' = (X - \min(x)) / (\max(x) - \min(x)), \quad Y' = (Y - \min(y)) / (\max(y) - \min(y)),$$

και λαμβάνουμε υπόψη ότι (για τα ανομήλικα δέντρα υβριδογενούς ελάτης του πανεπιστημιακού δάσους Περτουλίου) οι μετασχηματισμένες τ.μ. X', Y' (μπορεί να θεωρηθεί ότι) ακολουθούν Βήτα κατανομή για όλες σχεδόν τις επιφάνειες (Στυλιανάκη κ.α. (2006)). Έτσι, οι σ.π. δίνονται αντίστοιχα ως

$$f(x') = \frac{\Gamma(\beta_{31} + \beta_{41} + 2)}{\Gamma(\beta_{31} + 1)\Gamma(\beta_{41} + 1)} x'^{\beta_{31}} (1 - x')^{\beta_{41}}, \text{ όπου } x' \in [0,1] \text{ και } \beta_{31}, \beta_{41} > -1,$$

και

$$g(y') = \frac{\Gamma(\beta_{32} + \beta_{42} + 2)}{\Gamma(\beta_{32} + 1)\Gamma(\beta_{42} + 1)} y'^{\beta_{32}} (1 - y')^{\beta_{42}}, \text{ όπου } y' \in [0,1] \text{ και } \beta_{31}, \beta_{41} > -1.$$

Για την εκτίμηση των αγνώστων παραμέτρων των f και g χρησιμοποιήθηκαν οι προτεινόμενοι από τους Evans et al. (2000) εκτιμητές (οι οποίοι βρέθηκαν με τη μέθοδο των ροπών)

$$\hat{\beta}_{3i} = \hat{\alpha}_{1i} \{[\hat{\alpha}_{1i}(1 - \hat{\alpha}_{1i}) / \hat{\alpha}_{2i}] - 1\}, \quad \hat{\beta}_{4i} = (1 - \hat{\alpha}_{1i}) \{[\hat{\alpha}_{1i}(1 - \hat{\alpha}_{1i}) / \hat{\alpha}_{2i}] - 1\}, \quad i=1,2,$$

όπου $\hat{\alpha}_{1i}, \hat{\alpha}_{2i}, \hat{\alpha}_{3i}$ και $\hat{\alpha}_{4i}$ είναι οι δειγματικοί εκτιμητές του μέσου, της διακύμανσης, της λοξότητας και της κύρτωσης.

Οι Αρχιμήδειες συζεύξεις που θα εξετάσουμε στη συνέχεια είναι οι ακόλουθες:

1) Συζεύξεις *Clayton*, με

$$C(u, v) = [\max(u^{-\theta} + v^{-\theta} - 1, 0)]^{-\frac{1}{\theta}}.$$

2) Συζεύξεις Gumbel, με

$$C_{\theta}(u,v) = \exp(-[(-\ln u)^{\theta} + (-\ln v)^{\theta}]^{\frac{1}{\theta}}).$$

3) Συζεύξεις $C(***)$, με

$$C_{\theta}(u,v) = \max(1 - [(1-u)^{\theta} + (1-v)^{\theta}]^{\frac{1}{\theta}}, 0)$$

και γεννήτρια συνάρτηση την $\phi_{\theta}(t) = (1-t)^{\theta}$.

Οι τιμές της παραμέτρου θ , για τις συζεύξεις Clayton, Gumbel και $C(***)$ εκτιμήθηκαν αντίστοιχα από τις σχέσεις

$$\theta = \frac{2\tau_{X',Y'}}{1 - \tau_{X',Y'}}, \quad \theta = \frac{1}{1 - \tau_{X',Y'}} \quad \text{και} \quad \theta = \frac{2}{1 - \tau_{X',Y'}},$$

όπου $\tau_{X',Y'}$ το τ του Kendall που πρέπει να βρίσκεται, για τις συζεύξεις Clayton και $C(***)$ στο διάστημα $[-1, 1)$ και για τις Gumbel στο $[0, 1)$ (Nelsen, 2006).

Έγινε αναλυτική επεξεργασία των δεδομένων σε μια από τις 48 επιφάνειες, την επιφάνεια 39 (56 δέντρα):

Οι παράμετροι των περιθώριων Βήτα κατανομών βρέθηκαν με τη μέθοδο των ροπών να είναι για τις μετασχηματισμένες τιμές των διαμέτρων $\hat{\beta}_{31} = 1.17$, $\hat{\beta}_{41} = 1.34$, ενώ των υψών $\hat{\beta}_{32} = 1.06$, $\hat{\beta}_{42} = 0.7$. Στη συνέχεια υπολογίστηκαν οι παράμετροι θ για τις τρεις Αρχιμήδειες συζεύξεις και εξετάστηκε η προσαρμογή των τριών συζεύξεων στα δεδομένα με τη βοήθεια του Kolmogorov-Smirnov στατιστικού τεστ. Τα αποτελέσματα της προσαρμογής για την επιφάνεια 39, δίνονται στον Πίνακα 3.1:

Πίνακας 3.1. Εκτίμηση της παραμέτρου θ και της τιμής D του Kolmogorov-Smirnov στατιστικού τεστ για τις Clayton, Gumbel και $C(***)$ συζεύξεις

Επιφάνεια	Μοντέλο	Παράμετρος θ	D
39	Clayton	3.8496	0.1610
	Gumbel	2.9248	0.1766
	$C(***)$	5.8496	0.3474

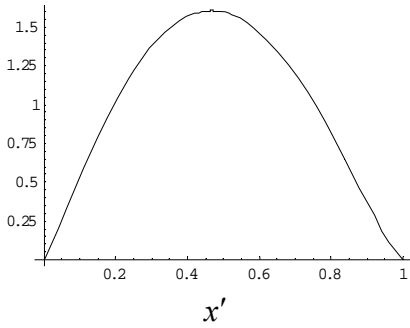
Από τον Πίνακα 3.1 προκύπτει ότι για την επιφάνεια 39 οι συζεύξεις Clayton και Gumbel προσαρμόζονται καλύτερα στα δεδομένα. Εργαζόμενοι με τον ίδιο τρόπο διαπιστώσαμε ότι το συμπέρασμα αυτό γενικεύεται για όλες τις επιφάνειες.

Οι γραφικές παραστάσεις των σ.π. για την επιφάνεια 39 έχουν ως εξής:

Σχήμα 3.1. Η σ.π. της διαμέτρου X' ,

$$f(x') = 9.099x'^{1.17}(1-x')^{1.34}.$$

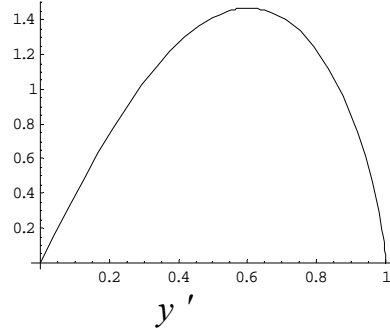
$f(x')$



Σχήμα 3.2. Η σ.π. του ύψους Y' ,

$$g(y') = 4.798y'^{1.06}(1-y')^{0.7}$$

$g(y')$



Ολοκληρώνοντας τις f και g προκύπτουν οι αντίστοιχες σ.κ.

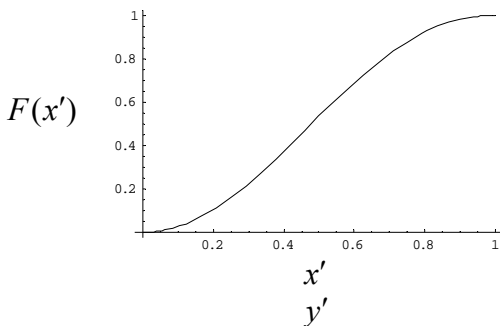
$$F(x') = 4.2x'^{2.17} \text{Hypergeom } 2F1 [2.17, -1.34, 3.17, x'] ,$$

$$G(y') = 2.3y'^{2.06} \text{Hypergeom } 2F1 [2.06, -0.7, 3.06, y'] ,$$

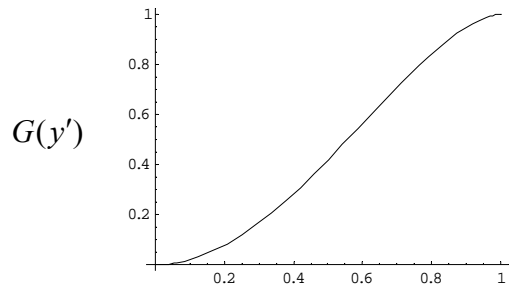
όπου ορίσαμε: $\text{Hypergeom } 2F1[a, b; c; z] = \sum_{k=0}^{\infty} \frac{(a)_k (b)_k}{(c)_k} \cdot \frac{z^k}{k!}.$

Οι γραφικές παραστάσεις των σ.κ. F και G δίνονται στα ακόλουθα σχήματα:

Σχήμα 3.3 Η σ.κ. της διαμέτρου X' ,



Σχήμα 3.4. Η σ.κ. του ύψους Y' ,



Στη συνέχεια υπολογίζεται η από κοινού σ.κ. για την επιφάνεια 39:

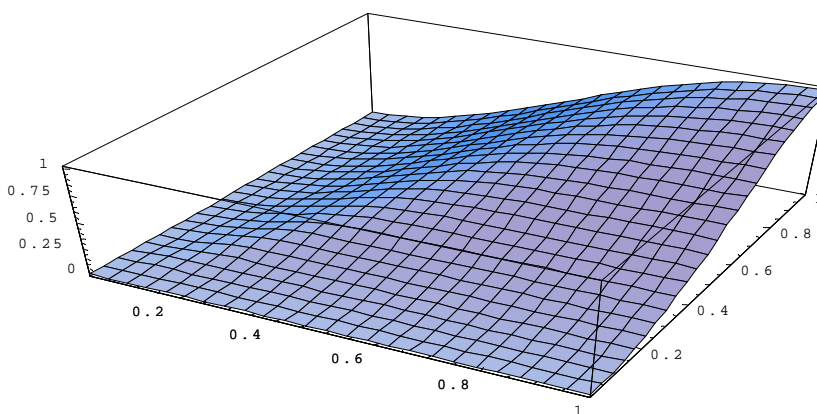
1) Εκτίμηση (προσέγγιση) της από κοινού κατανομής με τις Clayton συζεύξεις

Για τις Clayton συζεύξεις, όπου $C(u, v) = [\max(u^{-\theta} + v^{-\theta} - 1, 0)]^{-\frac{1}{\theta}}$, προκύπτει ότι

$$H(x', y') = C(F(x'), G(y')) \\ = [\max((4.19309x'^{2.17} \text{Hypergeom2F1}[2.17, -1.34, 3.17, x'])^{(-\theta)} + (2.32913y'^{2.06} \text{Hypergeom2F1}[2.06, -0.7, 3.06, y'])^{(-\theta)} - 1, 0)]^{-\frac{1}{\theta}}.$$

Στο σχήμα 3.5 που ακολουθεί δίνεται το τριδιάστατο διάγραμμα της από κοινού σ.κ. $H(x', y')$

Σχήμα 3.5. Η από κοινού κατανομή $H(x', y')$ διαμέτρων και υψών.



Με το X^2 τεστ διαπιστώθηκε η προσαρμογή της $H(x', y')$ στα δεδομένα σε σ.σ. 0.05.

2) Εκτίμηση της από κοινού κατανομής με τις Gumbel συζεύξεις

Για τις Gumbel συζεύξεις, όπου $C_{\theta}(u, v) = \exp[-(-\ln u)^{\theta} + (-\ln v)^{\theta}]^{\frac{1}{\theta}}$, έχουμε ότι

$$H(x', y') = \exp[-(-\ln(4.19309x'^{2.17} \text{Hypergeom2F1}[2.17, -1.34, 3.17, x']))^{\theta} + (-\ln(2.32913y'^{2.06} \text{Hypergeom2F1}[2.06, -0.7, 3.06, y']))^{\theta}]^{\frac{1}{\theta}}.$$

3) Εκτίμηση της από κοινού κατανομής με τις συζεύξεις της μορφής $C(***)$

Για τις συζεύξεις $C(***)$, με $C_{\theta}(u, v) = \max(1 - [(1-u)^{\theta} + (1-v)^{\theta}]^{\frac{1}{\theta}}, 0)$, προκύπτει ότι

$$H(x', y') = \max(1 - [(1 - (4.19309x'^{2.17} \text{Hypergeom}2F1[2.17, -1.34, 3.17, x'])^\theta)^\theta + (1 - (2.32913y'^{2.06} \text{Hypergeom}2F1[2.06, -0.7, 3.06, y'])^\theta)^\theta]^\frac{1}{\theta}, 0).$$

4. ΣΥΜΠΕΡΑΣΜΑ

Από τις συζεύξεις Clayton, Gumbel και C(***) που εξετάστηκαν για την επιφάνεια 39, καλύτερη στην προσαρμογή πρόκυψε να είναι η Clayton και ακολούθως η Gumbel. Το συμπέρασμα αυτό ισχύει για όλες τις επιφάνειες. Όμως, ενώ το συμπέρασμα ισχύει για (την έκταση συνολικά της) κάθε μια από τις επιφάνειες, εν τούτοις τόσο για την επιφάνεια 39 όσο και για τις υπόλοιπες, η εκτίμηση της από κοινού κατανομής με τη βοήθεια και των τριών εξεταζόμενων συζεύξεων δεν είναι αρκετά ικανοποιητική για τα μετασχηματισμένα δεδομένα που βρίσκονται στο κάτω αριστερά άκρο της κατανομής (δηλαδή για μικρές διαμέτρους και ύψη), ενώ για $x', y' > 0.4$ τα εκτιμώμενα αποτελέσματα προσεγγίζουν πολύ τα παρατηρούμενα.

ABSTRACT

In this paper two-dimensional distributions are estimated by means of copulas, which can be successfully used in the description of the joint distribution of breast height diameter and height of a hybrid fir in the University forest of Pertuli. The knowledge of the joint distribution function may help in estimating values with practical interest, such as the estimation of the timber volume.

ΑΝΑΦΟΡΕΣ

- Aderman. V., Pihlak. M. (2005). Using Copulas for Modeling the Dependence Between the Height and Diameter at Breast Height. *Commentationes Universitatis Tartuensis de Mathematica*. 9,77-83.
- Evans. M., Hastings. N., Peacock. B. (2000). *Statistical Distributions*. Third Edition. John Wiley & Sons.
- Genest, C. and Rivest, L.-P. (1993). Statistical inference procedures for bivariate Archimedean copulas. *Journal of the American Statistical Association* 88:1034-1043.
- Joe, H. (1997). *Multivariate Models and Dependence Concepts*. Chapman & Hall, London.
- Karian. Z. A., Dudewicz. E. J., McDonald. P. (1996). The extended Generalized Lamda Distribution System for fitting distributions to data: History, completion of theory, tables, applications, the “final word” on moment fits. *Communication in Statistics: Simul. Comput.* 25:611-642.
- Nelsen. R. B. (2006). *An Introduction to Copulas*. Springer Verlag. New York

- Σταματέλλος Γ. (1991). Έρευνα των Δυνατοτήτων Ογκομέτρησης των Δασών με Δισταδιακά Δειγματοληπτικά Σχέδια. *Διδακτορική Διατριβή. Α.Π.Θ.*
- Stamatellos G. (1995). Comparison of point-3P sampling estimation with cost analysis. *Forest ecology and management* 74, 75-79.
- Στυλιανάκη Ε., Σταματέλλος Γ. και Τσακλίδης Γ. (2006). Κατανομές διαμέτρων και υψών των δέντρων στη διαχείριση δασικών οικοσυστημάτων. *Πρακτικά 19ου Πανελληνίου Συνεδρίου Στατιστικής*, σελ 441-448.



ΕΝΑ ΣΥΣΤΗΜΑ ΜΕ ΕΠΑΝΑΛΑΜΒΑΝΟΜΕΝΟΥΣ ΠΕΛΑΤΕΣ, ΠΡΟΤΕΡΑΙΟΤΗΤΕΣ, ΕΞΑΡΤΗΜΕΝΗ ΡΟΗ ΑΦΙΞΕΩΝ ΚΑΙ ΒΛΑΒΕΣ ΥΠΑΛΛΗΛΟΥ

Ιωάννης Χ. Δημητρίου

Dept. of Electrical and Electronic Engineering

Imperial College, London, SW7 2BT, UK.

i.dimitriou@imperial.ac.uk

ΠΕΡΙΛΗΨΗ

Θεωρούμε ένα σύστημα εξυπηρέτησης επαναλαμβανόμενων αφίξεων με δύο κατηγορίες πελατών, προτεραιότητες, αναξιόπιστο εξυπηρετητή (υπάλληλο), αρνητικές αφίξεις και διακοπές. Το κυρίαρχο χαρακτηριστικό του μοντέλου είναι ότι ο ρυθμός αφίξεων των πελατών δεν εξαρτάται μόνο από τον τύπο του, εισερχόμενου στο σύστημα, πελάτη αλλά και από την κατάσταση του εξυπηρετητή. Μελετάται το σύστημα σε στατιστική ισορροπία ενώ δίνονται αποτελέσματα για διάφορα χαρακτηριστικά μέτρα, τη στοχαστική διάσπαση καθώς και πραγματοποιείται ανάλυση αξιοπιστίας.

Λέξεις κλειδιά: Σύστημα επαναλαμβανόμενων αφίξεων, εξηρητημένη ροή αφίξεων, αρνητικοί πελάτες, διακόπτοσα μη επαναληπτική προτεραιότητα, διακοπές εξυπηρετητή.

1. ΕΙΣΑΓΩΓΗ

Το χαρακτηριστικό γνώρισμα των συστημάτων εξυπηρέτησης με επαναλαμβανόμενους πελάτες είναι το εξής: ένας πελάτης που φθάνοντας στο σύστημα δεν βρίσκει διαθέσιμο τον εξυπηρετητή, αναχωρεί προσωρινά από αυτό και επαναλαμβάνει την άφιξή του ύστερα από τυχαίο χρονικό διάστημα, ανεξάρτητα από τους άλλους πελάτες που βρίσκονται στο σύστημα, με σκοπό να συνδεθεί με τον εξυπηρετητή.

Εφαρμογές τέτοιων συστημάτων εμφανίζονται συχνά σε συστήματα επικοινωνιών και τηλεπικοινωνιών και σε δίκτυα Η/Υ (LAN, Local Area Network) κάτω από διάφορα πρωτόκολλα επικοινωνίας. Για μια πλήρη επισκόπηση σχετικά με την ανάλυση και τις εφαρμογές των εν λόγω συστημάτων βλέπε Artalejo and Gomez-Corral (2008), Falin and Templeton (1997).

Στο κύριο όγκο της βιβλιογραφίας περι συστημάτων εξυπηρέτησης με επαναλαμβανόμενους πελάτες, ο ρυθμός αφίξεων παραμένει σταθερός. Ιδιαίτερη περίπτωση αποτελεί αυτή στην οποία πελάτες διαφόρων κατηγοριών φθάνουν στο σύστημα, εισάγοντας ταυτόχρονα και την έννοια της προτεραιότητας. Τότε ο ρυθμός αφίξεων εξαρτάται συνήθως από τον τύπο του πελάτη. Στην παρούσα λόγω εργασία θεω-

ρούμε για πρώτη φορά στη σχετική βιβλιογραφία ότι ο ρυθμός άφιξεων εξαρτάται τόσο από τον τύπο του πελάτη όσο και από την κατάσταση του εξυπηρετητή. Τα αποτελέσματα της παρούσης εργασίας περιλαμβάνονται στην πρόσφατη εργασία του Dimitriou (2011).

Το παρόν μοντέλο μπορεί να περιγράψει περιπτώσεις στις οποίες οι πελάτες έχουν γνώση της κατάστασης του εξυπηρετητή και ανάλογα με αυτή προσαρμόζουν την ροή τους προς το σύστημα, ώστε να αποφευχθεί συνωστισμός.

Τα τελευταία 15 χρόνια εμφανίζεται πλήθος εργασιών με το χαρακτηριστικό των αρνητικών αφίξεων. Το χαρακτηριστικό αυτό εισήχθη από τον Gelenbe (1991) στη προσπάθειά του να μοντελοποιήσει τη λειτουργία νευρωνικών δικτύων. Στην απλούστερη μορφή μία αρνητική άφιξη διαγράφει έναν πελάτη από το σύστημα. Οι αρνητικές αφίξεις στη μελέτη συστημάτων με επαναλαμβανόμενους πελάτες έχουν εισαχθεί από τους Artalejo and Gomez-Corral (1997, 1998).

Στον κύριο όγκο της σχετικής βιβλιογραφίας, ο εξυπηρετητής θεωρείται αξιόπιστος και πάντα διαθέσιμος. Στην πραγματικότητα όμως, δεν είναι πάντα διαθέσιμος αλλά υπόκειται σε βλάβες που συμβαίνουν τυχαία και επηρεάζουν την απόδοση του. Συστήματα με επαναλαμβανόμενους πελάτες και βλάβες εξυπηρετητή έχουν μελετηθεί από Aissani and Artalejo (1998), Dimitriou and Langaris (2010) ενώ σχετική μελέτη της αξιοπιστίας του έδωσαν οι Wang et al. (2001).

Όσον αφορά συστήματα με τη των ροή αφίξεων να εξαρτάται από την κατάσταση του υπαλλήλου, η βιβλιογραφία είναι περιορισμένη σε όγκο και αναφέρεται κυρίως σε Μαρκοβιανά συστήματα. Βλέπε σχετικά Shogan (1979), Boon et al. (2010).

Στη παρούσα εργασία γενικεύουμε την βιβλιογραφία συστημάτων με επαναλαμβανόμενους πελάτες και προτεραιότητες συνδέοντας για πρώτη φορά στη βιβλιογραφία όλα τα παραπάνω χαρακτηριστικά στη περίπτωση της διακόπτουσας μη επαναληπτικής προτεραιότητας και δίνοντας χρήσιμα χαρακτηριστικά μέτρα.

2. ΤΟ ΜΟΝΤΕΛΟ

Θεωρούμε ένα σύστημα εξυπηρέτησης με έναν εξυπηρετητή στο οποίο καταφθάνουν δύο τύποι πελατών, έστω P_i , $i = 1, 2$. Οι P_1 πελάτες που φθάνοντας δεν βρουν διαθέσιμο τον εξυπηρετητή, τοποθετούνται σε μία κανονική ουρά απείρου μήκους αναμένοντας τη σειρά τους να εξυπηρετηθούν. Αντίστοιχα, οι P_2 αν κατά την άφιξη τους δεν βρουν διαθέσιμο τον εξυπηρετητή εισέρχονται στο κουτί επαναλαμβανόμενων αφίξεων (retrial box) και επαναλαμβάνουν την προσπάθειά τους να συνδεθούν με τον εξυπηρετητή μετά από εκθετικό χρονικό διάστημα παραμέτρου α , ανεξάρτητα από τους υπόλοιπους πελάτες που βρίσκονται στο retrial box.

Οι P_1 έχουν διακόπτουσα μη-επαναληπτική προτεραιότητα έναντι των P_2 πελατών. Δηλαδή, αν κατά τη διάρκεια της εξυπηρέτησης ενός P_2 πελάτη, φθάσει στο σύστημα ένας P_1 , τότε τον διακόπτει και αναγκάζει τον εξυπηρετητή να αρχίσει να τον εξυπηρετεί. Ο διακοπτόμενος P_2 παραμένει στον χώρο εξυπηρέτησης αναμένοντας να ελευθερωθεί ο εξυπηρετητής από τους P_1 πελάτες για να συνεχίσει την εξυπηρέτησή του από το σημείο διακοπής.

Ο χρόνος εξυπηρέτησης ενός P_i πελάτη προέρχεται από μία αυθαίρετη γενική κατανομή με αθροιστική συνάρτηση κατανομής (*d.f.*) $B_i(x)$, συνάρτηση πυκνότητας πιθανότητας (*p.d.f.*) $b_i(x)$, Laplace – Stieltjes μετασχηματισμό (*LST*) $\beta_i^*(s)$, ροπές πρώτης και δεύτερης τάξης περί το μηδέν $\bar{b}_i, \bar{b}_i^{(2)}$ αντίστοιχα.

Το σύστημά υπόκειται σε βλάβες μετά από εκθετικό χρονικό διάστημα παραμέτρου μ . Αν συμβεί μία βλάβη, ο υπάλληλος στέλνεται άμεσα προς επισκευή, ενώ ο πελάτης που βρισκόταν στην εξυπηρέτηση την στιγμή της βλάβης, εισέρχεται στο retrieval box. Οι χρόνοι επισκευής προέρχονται από μία αυθαίρετη κατανομή με *d.f.* $B_3(x)$, *p.d.f.* $b_3(x)$, *LST* $\beta_3^*(s)$, $\bar{b}_3, \bar{b}_3^{(2)}$.

Όταν ο υπάλληλος μείνει ελεύθερος, δηλαδή όταν δεν υπάρχουν πελάτες στη κανονική ουρά ή στον χώρο εξυπηρέτησης μετά απο την ολοκλήρωση μιας εξυπηρέτησης ή μιας επισκευής, αναχωρεί για διακοπές τυχαίου μήκους με *d.f.* $B_4(x)$, *p.d.f.* $b_4(x)$, *LST* $\beta_4^*(s)$, \bar{b}_4 και $\bar{b}_4^{(2)}$. Επιστρέφοντας, αν υπάρχουν P_1 πελάτες να αναμένουν, αρχίζει να τους εξυπηρετεί ενώ σε αντίθετη περίπτωση παραμένει άεργος αναμένοντας την πρώτη άφιξη.

Επιπροσθέτως αρνητικές αφίξεις λαμβάνουν χώρα μετά από εκθετικό χρονικό διάστημα παραμέτρου ν και ' διαγράφουν' τον πελάτη στην εξυπηρετήση. Οι αρνητικές αφίξεις δεν έχουν καμία άλλη επίδραση στο μοντέλο.

Έστω $N_i(t)$, $i = 1, 2$, ο αριθμός των P_i πελατών στην ουρά και στο retrieval box αντίστοιχα την στιγμή t . Έστω επίσης η στοχαστική διαδικασία που περιγράφει την κατάσταση του υπαλλήλου,

$$\xi_t = \begin{cases} 0, & \text{άεργος την } t, \\ i, & \text{απασχολημένος με } P_i, \quad i = 1, 2 \text{ την } t, \\ 3, & \text{σε επισκευή την } t, \\ 4, & \text{σε διακοπές την } t, \end{cases}$$

$$u_t = \begin{cases} 0, & \text{κανένας } P_2 \text{ δεν έχει διακοπεί προηγουμένως την } t, \\ 1, & \text{ενας } P_2 \text{ έχει διακοπεί την } t. \end{cases}$$

Το στοιχείο που διακρίνει το εν λόγω μοντέλο είναι ότι η διαδικασία αφίξεων των πελατών είναι Poisson και εξαρτάται τόσο από την κατηγορία του πελάτη όσο και από την κατάσταση του εξυπηρετητή. Δηλαδή, δεδομένου ότι ο εξυπηρετητής είναι στην κατάσταση j , $j = 0, 1, 2, 3, 4$, οι P_i , $i = 1, 2$ πελάτες καταφθάνουν με ρυθμό λ_{ij} .

Οι πιθανότητες των καταστάσεων του συστήματος δίνονται από

$$\begin{aligned} p_i(k_1, k_2, x, t)dx &= P(N_1(t) = k_1, N_2(t) = k_2, x < \bar{B}_i(t) \leq x + dx, \\ &\quad \xi_t = i, u_t = 0), \quad i = 1, 3, 4, \\ p_2(k_2, x, t)dx &= P(N_2(t) = k_2, x < \bar{B}_2(t) \leq x + dx, \xi_t = 2, u_t = 0), \\ p_{i1}(k_1, k_2, x, y, t)dxdy &= P(N_1(t) = k_1, N_2(t) = k_2, x < \bar{B}_i(t) \leq x + dx, \\ &\quad y < \bar{B}_2(t) \leq y + dy, \xi = i, u_t = 1), \quad i = 1, 3, \\ q(k_2, t) &= P(N_2(t) = k_2, \xi_t = 0, u_t = 0), \end{aligned}$$

όπου \bar{X} ο χρόνος που έχει παρέλθει για την τυχαία μεταβλητή (τ.μ.) X .

3. ΒΟΗΘΗΤΙΚΑ ΑΠΟΤΕΛΕΣΜΑΤΑ

Στη συνέχεια παραθέτουμε ορισμένα σημαντικά για την μετέπειτα ανάλυση βοηθητικά αποτελέσματα. Από τώρα και στο εξής θα συμβολίζουμε με $y^*(s)$ τον LST της $y(t)$ για $Re(s) > 0$. Έστω S_i , $i = 1, 2$ το χρονικό διάστημα από τη στιγμή που αρχίζει να εξυπηρετείται ένας P_i πελάτης μέχρι τη στιγμή που ο εξυπηρετητής είναι έτοιμος για μια νέα εξυπηρέτηση. Στη περίπτωση ενός P_1 πελάτη, ο LST της απο κοινού κατανομής του S_1 και του αριθμού των νέων P_i που φθάνουν στο σύστημα δίνεται από

$$a_1^*(z_1, z_2, s) = \beta_1^*(\sigma_1(z_1, z_2, s)) + \frac{1 - \beta_1^*(\sigma_1(z_1, z_2, s))}{\sigma_1(z_1, z_2, s)} [\nu + \mu z_2 \beta_3^*(\sigma_3(z_1, z_2, s))],$$

όπου

$$\sigma_i(z_1, z_2, s) = s + \lambda_{1i}(1 - z_1) + \lambda_{2i}(1 - z_2) + (\mu + \nu)\delta_{\{i=1\}}, \quad i = 1, 3, 4,$$

και $\delta_{\{i\}}$ το δέλτα του Kronecker. Αν

$$\rho_1 = \frac{1 - \beta_1^*(\mu + \nu)}{\mu + \nu} (\lambda_{11} + \lambda_{13} \mu \bar{b}_3),$$

τότε αποδεικνύεται με τη βοήθεια του Takacs (1962) ότι,

Λήμμα 3.1. Για (i) $|z_2| < 1$, $\Re(s) \geq 0$, ή (ii) $|z_2| \leq 1$, $\Re(s) > 0$, ή (iii) $|z_2| \leq 1$, $\Re(s) \geq 0$ και $\rho_1 > 1$, η σχέση

$$z_1 - a_1^*(z_1, z_2, s), \quad (1)$$

έχει μοναδική ρίζα, $z_1 = x(s, z_2)$, εντός του $|z_1| < 1$. Για $s = 0$ και $z_2 = 1$, $x(0, 1)$ είναι η μικρότερη θετική πραγματική ρίζα της (1) με $x(0, 1) < 1$ αν $\rho_1 > 1$ και $x(0, 1) = 1$ για $\rho_1 \leq 1$.

Αν $\Theta^{(i)}$ η περίοδος συνεχούς απασχόλησης P_1 πελατών που αρχίζει με i P_1 πελάτες τότε ο LST της γεννήτριας πιθανοτήτων της απο κοινού κατανομής της $\Theta^{(i)}$ και του αριθμού $N(\Theta^{(i)})$ των νέων P_2 πελατών στο εν λόγω διάστημα δίνεται από

$$g^{(i)}(s, z_2) = x^i(s, z_2),$$

Επιπλέον για την περίπτωση του S_2 που ορίστηκε παραπάνω

$$a_2^*(z_1, z_2, s) = \beta_2^*(\sigma_2(z_2, s)) + \frac{1 - \beta_2^*(\sigma_2(z_2, s))}{\sigma_2(z_2, s)} [\nu + \mu z_2 \beta_3^*(\sigma_3(z_1, z_2, s))],$$

όπου

$$\sigma_2(z_2, s) = s + \lambda_{12}(1 - x(s, z_2)) + \lambda_{22}(1 - z_2) + \mu + \nu.$$

Αν V είναι το χρονικό διάστημα από τη στιγμή που ο εξυπηρετητής αναχωρεί για διακοπές μέχρι τη στιγμή που είναι άεργος για πρώτη φορά και $N(V)$ ο αριθμός των νέων P_2 στο εν λόγω διάστημα τότε κατ' αναλογία με τα παραπάνω

$$v^*(s, z_2) = \frac{\beta_4^*(\sigma_4(0, z_2, s))}{1 - \beta_4^*(\sigma_4(x(s, z_2), z_2, s)) + \beta_4^*(\sigma_4(0, z_2, s))}.$$

Ο χρόνος συμπλήρωσης εξυπηρέτησης C ενός P_2 πελάτη ορίζεται ως το χρονικό διάστημα από την στιγμή που αρχίζει να εξυπηρετείται έως τη στιγμή που ο εξυπηρετητής άεργος για πρώτη φορά. Τότε ο LST της πιθανογεννήτριας του αριθμού των νέων P_2 πελατών που φθάνουν στη διάρκεια το C είναι

$$c^*(s, z_2) = a_2^*(x(s, z_2), z_2, s). \quad (2)$$

Τέλος, ορίζουμε ως γενικευμένη περίοδος συνεχούς απασχόλησης P_1 πελατών τον χρόνο $W_1 = \Theta^{(1)} + V$ και γενικευμένο χρόνο συμπλήρωσης εξυπηρέτησης ενός P_2 πελάτη τον χρόνο $W_2 = C + V$. Σε αναλογία με τα παραπάνω

$$w_1^*(s, z_2) = x(s, z_2)v^*(s, z_2), \quad w_2^*(s, z_2) = c^*(s, z_2)v^*(s, z_2). \quad (3)$$

με

$$\begin{aligned} \tilde{\rho}_4 &= \frac{\partial}{\partial z_2} v^*(0, z_2)|_{z_2=1} = \frac{\rho_4}{1-\rho_1} \\ \tilde{\rho}_w &= \frac{\partial}{\partial z_2} w_2^*(0, z_2)|_{z_2=1} = \frac{\rho_2 + \rho_4}{1-\rho_1}, \end{aligned} \quad (4)$$

Οι παραπάνω ποσότητες προέκυψαν κατά την ανάλυση του μοντέλου και στην παρούσα ενότητα δώσαμε μια πιθανοθεωρητική ερμηνεία αυτών. Επιπλέον μας βοήθησαν στην απόδειξη του ακόλουθου βασικού θεωρήματος

Θεώρημα 1. Για (i) $\Re(s) > 0$, (ii) $\Re(s) \geq 0$, και $\rho = \rho_1 + \rho_2 + \rho_4 > 1$ η εξίσωση.

$$z_2 - w_2^*(s, z_2) = 0, \quad (5)$$

έχει μοναδική ρίζα, $z_2 = \phi(s)$ εντός της $|z_2| < 1$. Για $s = 0$, $\phi(0)$ είναι η μικρότερη θετική πραγματική ρίζα της (5) με $\phi(0) < 1$ αν $\rho > 1$ και $\phi(0) = 1$ για $\rho \leq 1$.

4. ΣΥΝΘΗΚΗ ΣΤΑΤΙΣΤΙΚΗΣ ΙΣΟΡΡΟΠΙΑΣ

Η συνθήκη στατιστικής ισορροπίας μας εξασφαλίζει την ύπαρξη των οριακών πιθανοτήτων της στοχαστικής διαδικασίας που περιγράφει το σύστημα μας. Για την εύρεση της εν λόγω συνθήκης θα χρησιμοποιήσουμε την θεωρία των Markov Renewal (Μαρκοβιανή Ανανεωτική Διαδικασία) και semi-regenerative διαδικασιών (Ημι-αναγεννητική Διαδικασία).

Ορίζοντας αρχικά μια μη διαχωρίσιμη, απεριοδική Μαρκοβιανή αλυσίδα $Y = \{N_{2i} = N_2(T_{i+}), i = 0, 1, \dots\}$ σε χρονικές στιγμές όπου ο εξυπηρετητής μένει άεργος για i -οστή φορά (T_i), μπορούμε να αποδείξουμε με χρήση του κριτηρίου Pakes (1969) ότι η Y είναι θετικώς επαναληπτική αν $\rho < 1$.

Επιπλέον η στοχαστική διαδικασία $\mathbf{X} = \{(N_1(t), N_2(t), \xi_t), t \geq 0\}$ που περιγράφει το σύστημα μας είναι μια ημι-αναγεννητική με υπεισερχόμενη Markov renewal διαδικασία την Μαρκοβιανή αλυσίδα που ορίσαμε παραπάνω. Τότε χρησιμοποιώντας το παραπάνω καθώς και αποτελέσματα από την semi-regenerative θεωρία (Cinlar (1975)), αποδεικνύεται ότι η συνθήκη $\rho < 1$ είναι ικανή για να είναι το σύστημα μας σε στατιστική ισορροπία.

5. ΑΝΑΛΥΣΗ ΣΕ ΣΤΑΤΙΣΤΙΚΗ ΙΣΟΡΡΟΠΙΑ

Έστω ότι το σύστημα μας είναι σε στατιστική ισορροπία, ώστε $\rho < 1$. Τότε αν $N_i = \lim_{t \rightarrow \infty} N_i(t)$, $i = 1, 2$, $\xi = \lim_{t \rightarrow \infty} \xi_t$, $u = \lim_{t \rightarrow \infty} u_t$, $\sigma_i(z_1, z_2) = \sigma_i(z_1, z_2, 0)$, $\sigma_2(z_2) = \sigma_2(z_2, 0)$, ορίζοντας τις γεννήτριες πιθανοτήτων

$$\begin{aligned} P_i(z_1, z_2, x) &= \sum_{k_1 \geq 0} \sum_{k_2 \geq 0} p_i(k_1, k_2, x) z_1^{k_1} z_2^{k_2} dx, \\ P_2(z_2, x) &= \sum_{k_2 \geq 0} p_2(k_2, x) z_2^{k_2} dx, \\ P_{i1}(z_1, z_2, x, y) &= \sum_{k_1 \geq 0} \sum_{k_2 \geq 0} p_{i1}(k_1, k_2, x, y) z_1^{k_1} z_2^{k_2} dx, \\ Q(z_2) &= \sum_{k_2 \geq 0} q(k_2) z_2^{k_2}. \end{aligned} \quad (6)$$

και εφαρμόζοντας την μέθοδο της συμπληρωματικής μεταβλητής συνδέοντας τις χρονικές στιγμές x , $x + dx$ μπορούμε να βρούμε εκφράσεις που συνδέουν τις παραπάνω γεννήτριες πιθανοτήτων. Με χρήση του Λήμματος 1 και καταλληλο συνδυασμό των εκφράσεων που προκύπτουν καταλήγουμε στην παρακάτω διαφορική εξίσωση

$$\alpha(z_2 - w_2^*(0, z_2)) \frac{\partial}{\partial z_2} Q(z_2) + [\lambda_0 - \lambda_{10} w_1^*(0, z_2) - \lambda_{20} w_2^*(0, z_2)] Q(z_2) = 0. \quad (7)$$

Η επίλυση της (7) θα μας δώσει όλες τις πιθανογεννήτριες που ορίστηκαν παραπάνω αφού αυτές δίνονται ως συναρτήσεις της $Q(z_2)$. Η επίλυση της (7) βασίζεται στο Θεώρημα 2, η σπουδαιότητα του οποίου, αναφέρεται στο ότι κάτω από τη συνθήκη $\rho < 1$ η $z_2 - w_2^*(0, z_2) = 0$ δεν μηδενίζεται ποτέ εντός του $|z_2| < 1$. Επιλύοντας την (7) προκύπτει

$$Q(z_2) = Q(1) \exp\left\{- \int_{z_2}^1 \frac{\lambda_0 - \lambda_{10} w_1^*(0, u) - \lambda_{20} w_2^*(0, u)}{\alpha[w_2^*(0, u) - u]} du\right\}. \quad (8)$$

Επιπλέον με χρήση του Θεωρήματος 2, μπορεί να αποδειχθεί ότι η συνθήκη $\rho < 1$ είναι επίσης αναγκαία για να είναι το σύστημα μας ευσταθές (βλέπε Dimitriou (2011)).

6. ΣΤΟΧΑΣΤΙΚΗ ΔΙΑΣΠΑΣΗ

Η έννοια της στοχαστικής διάσπασης στα συστήματα με επαναλαμβανόμεους πελάτες μας δίνει την δυνατότητα να διακρίνουμε την εξάρτηση των χαρακτηριστικών τους από τον ρυθμό επανάληψης α ενός πελάτη του retrial box.

Η μελέτη της στοχαστικής διάσπασης πραγματοποιείται θεωρώντας ένα βοηθητικό σύστημα, όμοιο με το αρχικό υποθέτοντας όμως ότι ο ρυθμός επανάληψης $\alpha \rightarrow \infty$. Σε αυτή τη περίπτωση οι επαναλαμβανόμενοι πελάτες συμπεριφέρονται σαν κανονικοί πελάτες. Τότε οι γεννήτριες πιθανοτήτων του αρχικού συστήματος και αυτές του βοηθητικού συνδέονται με τις ακόλουθες σχέσεις.

$$\begin{aligned} Q(z_2) &= q_0 Q_\alpha(z_2), & P_2(z_2) &= \hat{P}_i(z_2) Q_\alpha(z_2), \\ P_{i1}(z_1, z_2) &= \hat{P}_{i1}(z_1, z_2) Q_\alpha(z_2), & i &= 1, 3, \\ P_i(z_1, z_2) &= \hat{P}_i(z_1, z_2) Q_\alpha(z_2), & i &= 1, 3, 4, \end{aligned}$$

όπου $q_0 = Q(1) = \frac{1-\rho}{A}$ και

$$Q_\alpha(z_2) = \exp\left\{-\int_{z_2}^1 \frac{\lambda_0 - \lambda_{10}w_1^*(0, u) - \lambda_{20}w_2^*(0, u)}{\alpha[w_2^*(0, u) - u]} du\right\},$$

με A κατάλληλα ορισμένη ποσότητα (Dimitriou (2011)).

7. ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ ΣΥΣΤΗΜΑΤΟΣ

Με την βοήθεια των γεννητριών συναρτήσεων που ορίστηκαν παραπάνω μπορούμε να υπολογίσουμε σε κλειστή μορφή τις πιθανότητες των καταστάσεων του εξυπηρετητή καθώς και τον μέσο αριθμό πελατών στην ουρά και στο retrieval box αντίστοιχα. Για λόγους οικονομίας χώρου παραθέτουμε μόνο τις πιθανότητες των καταστάσεων του εξυπηρετητή.

$$P(\xi = 0) = Q(1) = \frac{1-\rho}{A},$$

$$P(\xi = 1, u = 0) = P_1(1, 1) = \frac{1-\beta_1^*(\mu+\nu)}{(\mu+\nu)A(1-\rho_1)} \left\{ \lambda_{10}(1-\rho) \left(1 + \frac{\lambda_{14}\bar{b}_4}{\beta_4^*(\lambda_{14})}\right) + (\lambda_{10}\rho_d + \lambda_{20}(1-\rho_1)) \left[\frac{\lambda_{13}\mu\bar{b}_3(1-\beta_2^*(\mu+\nu))}{\mu+\nu} + \frac{\lambda_{14}\bar{b}_4}{\beta_4^*(\lambda_{14})} \right] \right\}$$

$$P(\xi = 2) = P_2(1) = \frac{\lambda_{10}\rho_d + \lambda_{20}(1-\rho_1)}{A} \left(\frac{1-\beta_2^*(\mu+\nu)}{\mu+\nu} \right),$$

$$P(\xi = 4) = P_4(1, 1) = \left(\frac{\lambda_{10}(1-\rho+\rho_d) + \lambda_{20}(1-\rho_1)}{A} \right) \frac{\bar{b}_4}{\beta_4^*(\lambda_{14})},$$

$$P(\xi = 1, u = 1) = P_{11}(1, 1) = \frac{\lambda_{12}}{1-\rho_1} P_2(1) \frac{1-\beta_1^*(\mu+\nu)}{\mu+\nu},$$

$$P(\xi = 3, u = 1) = P_{31}(1, 1) = \mu\bar{b}_3 P_{11}(1, 1),$$

$$P(\xi = 3, u = 0) = P_3(1, 1) = \mu\bar{b}_3 [P_1(1, 1) + P_2(1)].$$

8. ΑΝΑΛΥΣΗ ΑΞΙΟΠΙΣΤΙΑΣ

Η ανάλυση αξιοπιστίας του εξυπηρετητή είναι υψίστης αναγκαιότητας στη μελέτη συστημάτων εξυπηρετήσης με βλάβες και επισκευές, εξαιτίας της ισχυρής επίδρασης των βλαβών στην λειτουργία του. Στην παρούσα ενότητα παρουσιάζουμε χαρακτηριστικά αξιοπιστίας όπως η πιθανότητα διαθεσιμότητας AV του εξυπηρετητή και ο ρυθμός εμφάνισης βλαβών F .

$$AV = 1 - \mu\bar{b}_3 \left[P_1(1, 1) + P_2(1) \left(1 + \frac{\lambda_{12}(1-\beta_1^*(\mu+\nu))}{(1-\rho_1)(\mu+\nu)} \right) \right],$$

$$F = \mu \left[P_1(1, 1) + P_2(1) \left(1 + \frac{\lambda_{12}(1-\beta_1^*(\mu+\nu))}{(1-\rho_1)(\mu+\nu)} \right) \right].$$

Επιπλέον μπορεί να υπολογιστεί ο LST της κατανομής του χρόνου μέχρι την πρώτη βλάβη $U(t) = P(\tau > t)$. Αυτός υπολογίζεται θεωρώντας ένα βοηθητικό σύστημα με λειτουργία όμοια με του αρχικού υποθέτοντας ότι οι στιγμές εμφάνισης βλαβών είναι απορροφητικές. Ακολουθως εφαρμόζουμε την μέθοδο της συμπληρωματικής μεταβλητής και χρησιμοποιώντας για το βοηθητικό σύστημα συμβολισμό ίδιο με αυτόν για το αρχικό καταλήγουμε στο ότι ο LST της κατανομής του χρόνου μέχρι την πρώτη βλάβη είναι

$$U^*(s) = Q^*(1, s) + P_1^*(1, 1, s) + P_2^*(1, s) + P_{11}^*(1, 1, s) + P_4^*(1, 1, s), \quad (9)$$

όπου

$$Q^*(z_2, s) = \int_{z_2}^{\omega_2(s)} \frac{1}{\alpha(\hat{f}_2(y, s) - y)} \exp \left\{ \int_y^{z_2} \frac{s + \lambda_0 - \lambda_{10} \hat{f}_1(x, s) - \lambda_{20} \hat{f}_2(x, s)}{\alpha(\hat{f}_2(x, s) - x)} dy \right\} dx, \quad z_2 \neq \omega_2(s),$$

και όλες οι ποσότητες που χρησιμοποιούνται είναι κατάλληλα ορισμένες. Επιπλέον, οι $P_{\bullet}^*(\bullet)$ δίνονται ως συναρτήσεις της $Q^*(1, s)$ και επομένως αντικαθιστώντας στην $U^*(s)$ έχουμε το ζητούμενο. Για περισσότερες λεπτομέρειες βλέπε Dimitriou (2011).

ABSTRACT

In this paper we consider an unreliable single server retrieval queue accepting two types of customers, with negative arrivals, preemptive resume priorities and vacations. A distinguishing feature of the model is that the rates of the Poisson arrival process depends on the server state. For this model we investigate the stability conditions and the joint queue length distribution in steady state. We also prove that our model satisfies the stochastic decomposition property. Transient, as well as steady state solutions for reliability measures are obtained.

ΒΙΒΛΙΟΓΡΑΦΙΑ

- Aissani, A. and Artalejo, J.R. (1998). On the single server retrieval queue subject to breakdowns, *Queueing Systems*, **30**, 309-321.
- Artalejo, J.R. and Gomez-Corral, A. (2008). *Retrial Queueing Systems, A Computational Approach*, Springer-Verlag Berlin Heidelberg.
- Artalejo, J.R. and Gomez-Corral A. (1997). Steady state solution of a single-server queue with linear requests repeated. *Journal of Applied Probability*, **34**, 223-233.
- Artalejo, J.R. and Gomez-Corral, A. (1998). Analysis of a stochastic clearing system with repeated attempts. *Stochastic Models*, **14**, 623-645.
- Boon, M., van Wijk, A., Adan, I. and Boxma, O. (2010). A polling model with smart customers, *Queueing Systems*, **66**, 239-274.
- Cinlar, E. (1975). *Introduction to Stochastic Processes*, Prentice-Hall.
- Dimitriou, I. (2011). A preemptive resume priority retrieval queue with state dependent arrivals, unreliable server and negative customers, *TOP*, Accepted, DOI 10.1007/s11750-011-0198-4.
- Dimitriou, I. and Langaris, C. (2010). A repairable queueing model with two-phase service, start-up times and retrial customers, *Computers and Operations Research*, **37**, 1181-1190.
- Falin, G.I., and Templeton, J.G.C. (1997). *Retrial Queues*, Chapman and Hall, London.
- Gelenbe, E. (1991). Product form networks with negative and positive customers. *Journal of Applied Probability*, **28**, 656-663.
- Pakes, A.G. (1969). Some conditions of ergodicity and recurrence of Markov chains, *Operation Research*, **17**, 1058-1061.
- Shogan, A.W. (1979). A single server queue with arrival rate dependent on server breakdowns, *Naval Research Logistics Quarterly*, **26**, 487-497.

Takacs, L. (1962). *Introduction to the Theory of Queues*, Oxford Univ. Press, New York.

Wang, J., Cao, J. and Li, Q.-L. (2001). Reliability analysis of the retrial queue with server breakdowns and repairs, *Queueing Systems*, **38**, 363-380.



ΙΣΟΤΙΜΙΑ ΑΓΟΡΑΣΤΙΚΗΣ ΔΥΝΑΜΗΣ ΚΑΙ ΕΛΕΓΧΟΙ ΜΟΝΑΔΙΑΙΑΣ ΡΙΖΑΣ ΣΕ ΠΑΝΕΛ ΔΕΔΟΜΕΝΑ

Νίκος Δριτσάκης¹

¹Τμήμα Εφαρμοσμένης Πληροφορικής, Πανεπιστήμιο Μακεδονίας
drits@uom.gr

ΠΕΡΙΛΗΨΗ

Στην εργασία αυτή ερευνούμε την ισχύ της ισοτιμίας της αγοραστικής δύναμης για τις δώδεκα νέες χώρες της Ε.Ε. με τη χρησιμοποίηση των ελέγχων της μοναδιαίας ρίζας σε διαχρονικά και διαστρωματικά στοιχεία (πάνελ δεδομένα). Ο σκοπός της εργασίας αυτής είναι να εξετάσουμε τη στασιμότητα της πραγματικής συναλλαγματικής ισοτιμίας. Χρησιμοποιώντας τους ελέγχους της μοναδιαίας ρίζας, του επαυξημένου Dickey-Fuller (ADF) και των Phillips-Peron (PP) σε απλές χρονοσειρές, διαπιστώνεται ότι η πραγματική συναλλαγματική ισοτιμία όλων των νέων χωρών της Ε.Ε. παρουσιάζει μοναδιαία ρίζα. Το αποτέλεσμα αυτό είναι πολύ πιθανόν να οφείλεται στη χαμηλή ισχύ που έχουν οι έλεγχοι αυτοί. Ο στόχος αυτής της εργασίας είναι να επανεξεταστεί αυτό το ζήτημα με την εκμετάλλευση των πρόσθετων πληροφοριών που παρέχονται από το συνδυασμό διαστρωματικών δεδομένων και στοιχείων χρονολογικών σειρών, καθώς και τα πλεονεκτήματα που παρέχουν οι σύγχρονοι έλεγχοι της μοναδιαίας ρίζας σε πάνελ δεδομένα. Τα αποτελέσματα από τους ελέγχους αυτούς έδειξαν ότι η πραγματική συναλλαγματική ισοτιμία στις νέες χώρες της Ε.Ε. δεν παρουσιάζει μοναδιαία ρίζα (είναι στάσιμη). Επομένως υπάρχει μακροχρόνια σχέση στην ισοτιμία της αγοραστικής δύναμης για τις δώδεκα νέες χώρες της Ε.Ε.

Λέξεις Κλειδιά: Ισοτιμία Αγοραστικής Δύναμης, Πάνελ Δεδομένα, Μοναδιαία ρίζα, Νέα Μέλη της Ε.Ε.

JEL classification: C12, C23, F15

1. ΕΙΣΑΓΩΓΗ

Την 1η Μαΐου του 2004, η Ε.Ε. γνώρισε τη μεγαλύτερη επέκτασή της με την προσχώρηση δέκα νέων χωρών μελών δηλαδή της Κύπρου, της Δημοκρατίας της Τσεχίας, της Εσθονίας, της Ουγγαρίας, της Λετονίας, της Λιθουανίας, της Μάλτας, της Πολωνίας, της Σλοβακίας και της Σλοβενίας. Την 1η Ιανουαρίου του 2007, η Βουλγαρία και η Ρουμανία προσχώρησαν επίσης στην Ε.Ε. Προκειμένου να προσχωρήσουν επιτυχώς στην Ε.Ε. οι χώρες αυτές έπρεπε να ικανοποιήσουν

ορισμένα οικονομικά και πολιτικά κριτήρια, τα οποία περιλαμβάνουν την ύπαρξη σταθερής δημοκρατίας, το σεβασμό στα ανθρώπινα δικαιώματα καθώς επίσης και τη λειτουργία της ελεύθερης αγοράς.

Εκτός από την Κύπρο και τη Μάλτα, όλες οι άλλες οικονομίες αντιμετώπισαν πολλά προβλήματα κατά τη διάρκεια της περιόδου πριν τις διαπραγματεύσεις προσχώρησή τους στην Ε.Ε. Προς το τέλος της δεκαετίας του '90, οι περισσότερες από τις χώρες αυτές προσάρμοσαν τα νομίσματά τους στο ευρώ προκειμένου να ενισχυθεί η περίπτωση τους για την προσθήκη τους στην Ε.Ε. Οι περισσότερες από αυτές τις χώρες έχουν γίνει ήδη μέλη του μηχανισμού συναλλαγματικής ισοτιμίας και προγραμματίζουν να υποβάλουν αίτηση για την ιδιότητα μέλους της ευρωζώνης. Επειδή, οι περισσότερες από αυτές τις χώρες αντιμετωπίζουν ακόμα σοβαρά προβλήματα πληθωρισμού, παρέχουν το ενδιαφέρον να ερευνηθεί η ισοτιμία της αγοραστικής δύναμης στις χώρες αυτές.

Η υπόθεση της ισοτιμίας της αγοραστικής δύναμης είναι μεταξύ των δημοφιλέστερων ερευνητικών θεμάτων στη διεθνή μακροοικονομική βιβλιογραφία. Η ισοτιμία της αγοραστικής δύναμης προτείνει ότι η αλλαγή στις συναλλαγματικές ισοτιμίες μεταξύ δύο νομισμάτων καθορίζεται από τις σχετικές τιμές δύο χωρών. Για παράδειγμα αν οι τιμές στην Ε.Ε. αυξάνονται κατά 2% κατά τη διάρκεια ενός έτους, ενώ στις Ηνωμένες Πολιτείες της Αμερικής η άνοδος του επιπέδου των τιμών αυξάνεται κατά 4%, και οι ειδικοί προβλέπουν ότι το ευρώ θα ανατιμηθεί έναντι του δολαρίου κατά 2% τότε η εκτίμηση του ευρώ έναντι του δολαρίου ακυρώνει το διαφορετικό ποσοστό του πληθωρισμού. Επομένως, οι σχετικές ξένες και εσωτερικές αγοραστικές δυνάμεις και των δύο νομισμάτων θα παραμείνουν αμετάβλητες.

Η βιβλιογραφία για την ισοτιμία της αγοραστικής δύναμης είναι πολύ μεγάλη. Πρόσφατα οι ερευνητές έχουν χρησιμοποιήσει τις τεχνικές της στασιμότητας και της συνολοκλήρωσης για τον έλεγχο της μακροχρόνιας σχέσης στην ισοτιμία της αγοραστικής δύναμης. Οι απόψεις για τη σχέση αυτή δίστανται. Μερικές από τις εργασίες που χρησιμοποίησαν τις τεχνικές της συνολοκλήρωσης, όπως οι εργασίες των Corbae and Ouliaris (1988), Kim (1990), Doganlar (1999) and Narayan (2007), έχουν δείξει ότι η μακροχρόνια σχέση στην ισοτιμία της αγοραστικής δύναμης δεν υφίσταται. Από την άλλη πλευρά υπάρχουν οι εργασίες των Kugler and Lenz (1993) and MacDonald (1993) που υποστηρίζουν ακριβώς την αντίθετη άποψη.

Χρησιμοποιώντας την τεχνική της στασιμότητας μερικοί ερευνητές όπως Hakkio (1986), Serletis and Zimonopoulos (1997), Bahmanee-Oskoae (1998), και Narayan (2005) έχουν δείξει ότι οι πραγματικές συναλλαγματικές ισοτιμίες περιέχουν μοναδιαίες ρίζες και επομένως η μακροχρόνια σχέση στην ισοτιμία της αγοραστικής δύναμης δεν υπάρχει. Σε κάποιες εργασίες που στα στοιχεία τους οι ερευνητές χρησιμοποίησαν τις διαρθρωτικές μεταβολές, όπως Marcela et al. (2003) για το Μεξικό, Narayan and Prasad (2005) για 11 χώρες της Μέσης Ανατολής, Sideris (2011) για 6 νέες χώρες της ΕΕ. βρήκαν ότι η πραγματική συναλλαγματική ισοτιμία είναι στάσιμη και επομένως η μακροχρόνια σχέση στην ισοτιμία της αγοραστικής δύναμης υπάρχει.

Οι πρόσφατοι έλεγχοι της μοναδιαίας ρίζας σε πάνελ δεδομένα έχουν εφαρμοστεί για να εξετάσουν τη στασιμότητα της πραγματικής συναλλαγματικής ισοτιμίας. Οι εργασίες των Frankel and Rose (1996), Papell and Theodoridis (1998), Wu and Wu (2001) διαπιστώνουν ότι οι πραγματικές συναλλαγματικές ισοτιμίες ακολουθούν μια στάσιμη διαδικασία.

Ο σκοπός αυτής της εργασίας είναι να εξεταστεί η ισχύς της ισοτιμίας της αγοραστικής δύναμης για δώδεκα νέες χώρες της Ε.Ε. Η υπόλοιπη εργασία οργανώνεται ως εξής. Η παράγραφος 2 καθορίζει τυπικά το αναλυτικό πλαίσιο και την οικονομετρική μεθοδολογία. Η παράγραφος 3 περιγράφει τα στοιχεία και παρουσιάζει το εμπειρικό αποτέλεσμα. Η παράγραφος 4 παρουσιάζει τα συμπεράσματα αυτής της εργασίας.

2. ΑΝΑΛΥΤΙΚΟ ΠΛΑΙΣΙΟ ΚΑΙ ΟΙΚΟΝΟΜΕΤΡΙΚΗ ΜΕΘΟΔΟΛΟΓΙΑ

Οι έλεγχοι της ισοτιμίας της αγοραστικής δύναμης αρχίζουν με τον υπολογισμό της πραγματικής συναλλαγματικής ισοτιμίας. Η συνάρτηση της πραγματικής συναλλαγματικής ισοτιμίας υπολογίζεται ως εξής:

$$RER = NER(P^*/P) \quad (1)$$

όπου RER είναι η πραγματική συναλλαγματική ισοτιμία, NER είναι η ονομαστική συναλλαγματική ισοτιμία, P είναι το γενικό επίπεδο τιμών αγαθών και υπηρεσιών εσωτερικού και P* είναι το γενικό επίπεδο τιμών αγαθών και υπηρεσιών εξωτερικού αντίστοιχα.

Λογαριθμώντας τη συνάρτηση (1) η συνάρτηση της πραγματικής συναλλαγματικής ισοτιμίας γίνεται ως εξής:

$$\log(RER) = \log(NER) + \log(P^*) - \log(P) \quad (2)$$

Οι περισσότερες έρευνες σχετικές με την ισοτιμία της αγοραστικής δύναμης ερευνούν το πρόβλημα με τη χρησιμοποίηση της τεχνικής της συνολοκλήρωσης. Εάν υπάρχει μια σχέση συνολοκλήρωσης μεταξύ της ονομαστικής συναλλαγματικής ισοτιμίας και του γενικού επιπέδου των τιμών (P*/P), τότε λέμε ότι υπάρχει ισοτιμία της αγοραστικής δύναμης μεταξύ των κρατών που μελετάμε. Ένας άλλος τρόπος να εξεταστεί η ισοτιμία της αγοραστικής δύναμης είναι να εφαρμοστούν οι έλεγχοι της μοναδιαίας ρίζας στην πραγματική συναλλαγματική ισοτιμία. Από το παραπάνω υπόδειγμα, η ισοτιμία της αγοραστικής δύναμης υπάρχει μόνο εάν υπάρχει μια τάση στην πραγματική συναλλαγματική ισοτιμία, [log(RER)], που θα επανέλθει στην θέση ισορροπίας του. Δηλαδή η διαδικασία της χρονικής σειράς της πραγματικής συναλλαγματικής ισοτιμίας είναι στάσιμη εάν η ισοτιμία της αγοραστικής δύναμης υπάρχει. Αυτή είναι και η προσέγγιση που χρησιμοποιούμε στην εργασία αυτή. Ακολουθώντας τα παραπάνω παρατηρούμε ότι το υπόδειγμα του μέσου όρου επανέρχεται στην πραγματική συναλλαγματική ισοτιμία ως ακολούθως:

$$\log(RER)_t = \alpha + b \log(RER)_{t-1} + e_t \quad (3)$$

όπου α είναι μια σταθερά, b είναι μία άγνωστη παράμετρος και e_t απεικονίζει μια διαδικασία λευκού θορύβου.

Εάν η πραγματική συναλλαγματική ισοτιμία είναι στάσιμη τότε οποιεσδήποτε αλλαγές του ποσοστού στο επίπεδο τιμών μεταξύ δύο χωρών θα αντισταθμίζεται από μια ίση υποτίμηση (υπερτίμησης) της ονομαστικής συναλλαγματικής ισοτιμίας. Εάν υπάρχει μοναδιαία ρίζα στην πραγματική συναλλαγματική ισοτιμία τότε οι διαρθρωτικές μεταβολές στην πραγματική συναλλαγματική ισοτιμία είναι μόνιμες και η ισοτιμία της αγοραστικής δύναμης μεταξύ δύο χωρών δεν υπάρχει.

2.1. ΕΛΕΓΧΟΣ ΜΟΝΑΔΙΑΙΑΣ ΡΙΖΑΣ

Στην αρχή η στασιμότητα της πραγματικής συναλλαγματικής ισοτιμίας των 12 χωρών της Ε.Ε. εξετάζονται με τη χρησιμοποίηση του επαυξημένου ελέγχου των Dickey-Fuller (1979, 1981) και του ελέγχου των Phillips-Peron (1988). Ο επαυξημένος έλεγχος των Dickey-Fuller (ADF) είναι βασισμένος στην εκτίμηση της παρακάτω εξίσωσης:

$$\Delta X_t = \delta_0 + \delta_1 t + \delta_2 X_{t-1} + \sum_{i=1}^k \alpha_i \Delta X_{t-i} + e_t \quad (4)$$

όπου X_t = η χρονική σειρά, t = στοχαστική τάση, e_t = λευκός θόρυβος.

Χρησιμοποιώντας την εξίσωση (4), εξετάζουμε τη μηδενική και την εναλλακτική υπόθεση για την μοναδιαία ρίζα ως εξής: $H_0 : \delta_2 = 0$ και $H_e : \delta_2 < 0$

Η στατιστική Z_t των Phillips (1987) και των Phillips και Perron (1988) είναι μια τροποποίηση του ελέγχου των Dickey-Fuller στη στατιστικής t που επιτρέπει την αυτοσυσχέτιση και την υπό όρους ετεροσκεδαστικότητα στον όρο σφάλματος της παλινδρόμησης των Dickey-Fuller και βασίζεται στην εκτίμηση της παρακάτω εξίσωσης:

$$\Delta X_t = \beta_0 + \beta_1 t + \beta_2 X_{t-1} + e_t \quad (5)$$

όπου X_t = η χρονική σειρά, t = στοχαστική τάση, e_t = λευκός θόρυβος.

2.2. ΕΛΕΓΧΟΣ ΜΟΝΑΔΙΑΙΑΣ ΡΙΖΑΣ ΣΕ ΠΑΝΕΛ ΔΕΔΟΜΕΝΑ

Στην περίπτωση όμως που υπάρχουν τόσο διαχρονικά, όσο και διαστρωματικά στοιχεία τότε για τον έλεγχο της μοναδιαίας ρίζας χρησιμοποιούνται πρόσφατες τεχνικές όπως είναι αυτές των Breitung (2000), Levin, Lin and Chu (2002) (LLC), Im, Pesaran and Shin (2003) W-test (IPS), ADF-Fisher Chi-square test (ADF-Fisher), PP Fisher Chi-Square test (PP-Fisher) Maddala and Wu (1999) and Hadri (2000).

Από τους ελέγχους αυτούς οι πιο δημοφιλέστεροι είναι αυτοί του Levin, Lin and Chu (2002) (LLC) που υποθέτουν την ομοιογένεια (homogeneity) των συντελεστών σε ένα αυτοπαλίνδρομο υπόδειγμα για όλα τα μέλη των ομάδων (στρωμάτων), καθώς και των Im, Pesaran and Shin (2003) που είναι γενικότερος από τον έλεγχο των LLC διότι επιτρέπει την ετερογένεια (heterogeneity) στα διαχρονικά και διαστρωματικά στοιχεία (panel data). Εδώ πρέπει να επισημάνουμε ότι και οι δύο έλεγχοι βασίζονται στον παραδοσιακό έλεγχο των Dickey-Fuller.

- Οι Levin, Lin και Chu (LLC) για τα διαχρονικά και διαστρωματικά δεδομένα (panel data) έχουν προτείνει ένα έλεγχο μοναδιαίας ρίζας που βασίζεται στον επαυξημένο έλεγχο των Dickey-Fuller (ADF) ακολουθώντας την παρακάτω εξίσωση:

$$\Delta y_{it} = c_i + \gamma_i y_{i,t-1} + \sum_{j=1}^k c_j \Delta y_{i,t-j} + e_{it} \quad (6)$$

όπου: $t = 1, \dots, T$ είναι οι χρονικές περιόδους, και $i = 1, \dots, N$ αριθμός ομάδων (στρωμάτων). Η μηδενική υπόθεση για τον έλεγχο των LLC είναι: $H_0: \gamma_1 = \gamma_2 = \gamma = 0$ για όλα τα στρώματα (i), έναντι της εναλλακτικής $H_a: \gamma_1 = \gamma_2 = \gamma < 0$ για όλα τα στρώματα, με τον έλεγχο να βασίζεται στο στατιστικό $t_\gamma = \hat{\gamma} / (s.e(\hat{\gamma}))$.

Όπως έχουμε επισημάνει και πιο πάνω, ο έλεγχος των (LLC) υποθέτει την ομοιογένεια (homogeneity) των συντελεστών σε ένα δυναμικό αυτοπαλίνδρομο υπόδειγμα Panel για όλα τα μέλη των ομάδων (στρωμάτων). Συγκεκριμένα ο παραπάνω έλεγχος υποθέτει ότι κάθε στρώμα έχει το ίδιο αυτοπαλίνδρομο συντελεστή πρώτης τάξης γ_i , αλλά επιτρέπει την επίδραση της χρονικής τάσης μεμονωμένα σε κάθε ομάδα (στρώμα). Οι χρονικές υστερήσεις της εξαρτημένης μεταβλητής μπορούν να προστεθούν στο υπόδειγμα για τον τμηματικό συσχετισμό στα λάθη.

- Οι Im Pesaran και Shin στο υπόδειγμά τους, επιτρέπουν μεμονωμένες επιδράσεις, χρονικές τάσεις, και κοινές χρονικές επιδράσεις για τα ετερογενή διαχρονικά και διαστρωματικά στοιχεία. Ο έλεγχος που προτείνεται από τους Im Pesaran και Shin (IPS) επιτρέπει την ετερογένεια μεταξύ των ομάδων (στρωμάτων) σε ένα δυναμικό πλαίσιο διαχρονικών και διαστρωματικών στοιχείων (panel data) και βασίζεται στον επαυξημένο έλεγχο των Dickey-Fuller (ADF):

$$\Delta Y_{it} = \rho_i Y_{it-1} + \sum_{k=1}^p \gamma_{ik} \Delta Y_{it-k} + Z_{it} \delta + \varepsilon_{it} \quad (7)$$

όπου Y_{it} κάθε μεταβλητή του υποδείγματος, p είναι ο αριθμός των χρονικών υστερήσεων για τα ανεξάρτητα συσχετιζόμενα κατάλοιπα, Z_{it} δείχνει το διάνυσμα των προσδιοριστικών μεταβλητών στο υπόδειγμα συμπεριλαμβανομένων των σταθερών επιδράσεων ή των μεμονωμένων τάσεων και δ είναι το αντίστοιχο διάνυσμα των συντελεστών.

Η μηδενική και η εναλλακτική υπόθεση ορίζονται ως εξής:

$$H_1 = \left\{ \begin{array}{l} \rho_i = 0 \text{ για } i = 1, \dots, N \\ \rho_i < 0 \text{ για } i = N + 1, N + 2, \dots, N \end{array} \right\}$$

όπου N είναι ο αριθμός των ομάδων (στρωμάτων).

Οι Im Pesaran και Shin χρησιμοποιούν χωριστούς ελέγχους μοναδιαίας ρίζας για την κάθε N διαστρωματική ομάδα. Οι Im Pesaran και Shin προτείνουν επίσης τη χρήση του μέσου της προσαρμοσμένης στατιστικής t (t -bar statistic) για κάθε ομάδα όπου οι στατιστικές από κάθε έλεγχο του επαυξημένου Dickey-Fuller υπολογίζονται κατά μέσο όρο σε όλα τα διαχρονικά και διαστρωματικά στοιχεία. Οι προσαρμοσμένοι παράγοντες χρειάζονται για να μεταφράσουν την κατανομή της προσαρμοσμένης στατιστικής t (t -bar statistic) σε μια τυποποιημένη κανονική

κατανομή κάτω από τη μηδενική υπόθεση H_0 . Ο μέσος όρος των μεμονωμένων στατιστικών του επαυξημένου Dickey-Fuller ADF και ορίζεται ως:

$$\bar{t} = (1/N) \sum_{i=1}^N (t_{pi}) \quad (8)$$

όπου t_{pi} είναι η μεμονωμένη t -στατιστική για τον έλεγχο της μηδενικής υπόθεσης. Σύμφωνα με τη μηδενική υπόθεση, όλες οι σειρές στα διαχρονικά και διαστρωματικά στοιχεία είναι μη στάσιμες διαδικασίες, ενώ σύμφωνα με την εναλλακτική υπόθεση, ένα μέρος της σειράς στα διαχρονικά και διαστρωματικά στοιχεία υποτίθεται ότι είναι στάσιμο.

• Ο Breitung (2000) πρότεινε ένα τυπικό στατιστικό έλεγχο (t -ratio) για τον έλεγχο της μοναδιαίας ρίζας σε πάνελ στοιχεία. Μέσω της αριθμητικής ανάλυσης, υποστήριξε ότι ο έλεγχός του έχει τις ιδιότητες μιας δύναμης σε γειτονικές ομάδες. Ο έλεγχος του Breitung (2000) διαφέρει από τον έλεγχο των Levin, Lin and Chu στα παρακάτω δύο σημεία: Το πρώτο σημείο είναι ότι, για την παραγωγή της τυποποιημένης διαδικασίας, το αυτοπαλίνδρομο μέρος του υποδείγματος παραλείπεται.

$$\Delta Y_{it} = \left(\Delta Y_{it} - \sum_{k=1}^{pt} \gamma_{ik} \Delta Y_{it-k} \right) / s_i \quad (9)$$

$$\Psi_{it-1} = \left(Y_{it-1} + \sum_{k=1}^{pt} \gamma_{ik} \Delta Y_{it-k} / s_i \right) \quad (10)$$

όπου s_i είναι τα εκτιμημένα τυποποιημένα λάθη.

Το δεύτερο σημείο είναι ότι οι μεταβλητές μετασχηματίζονται ως εξής:

$$\Delta Y_{it} = \sqrt{(T-t)/(T-t+1)} \left[\Delta Y_{it} (\Delta Y_{it+1} + \dots + \Delta Y_{it+T}) / (T-t) \right], \Delta Y_{it-1} = Y_{it-1} - c_{it} \quad (11)$$

όπου

$$c_{it} = \left\{ \begin{array}{ll} 0 & \text{χωρίς σταθερά και τάση} \\ Y_{it} & \text{με σταθερά χωρίς τάση} \\ Y_{it} - [T^{-1}(t-1)]Y_{iT} & \text{με σταθερά και τάση} \end{array} \right\}$$

• Οι Maddala και Wu (1999) προτείνουν έναν έλεγχο της μοναδιαίας ρίζας σε στοιχεία πάνελ, που στηρίζεται στην εργασία του Fisher (1932). Ο έλεγχός τους εξετάζει βασικά την πιθανότητα (p -value) του στατιστικού ελέγχου των μεμονωμένων ομάδων για μια μοναδιαία ρίζα, και την συνδυάζει με μια στατιστική με πάνελ στοιχεία. Ο έλεγχος της κατανομής X^2 με δύο βαθμούς ελευθερίας και έχει την ακόλουθη μορφή:

$$\lambda = -2 \sum_{i=1}^N \log_e \pi_i \quad (12)$$

όπου, π_i είναι η πιθανότητα (p -value) του στατιστικού ελέγχου για την ομάδα i .

Ένα σημαντικό πλεονέκτημα αυτού του ελέγχου είναι ότι μπορεί να χρησιμοποιηθεί ανεξάρτητα από το εάν υπάρχει ολοκλήρωση μηδενικής ή πρώτης τάξης ή υπάρχει στασιμότητα. Η πιθανότητα (p-value) υπολογίζεται από τον επαυξημένο έλεγχο των Dickey-Fuller (ADF) και τον έλεγχο των Phillips-Peron (PP). Η απλότητα αυτού του ελέγχου και η ευρωστία της στην επιλογή του μήκους των υστερήσεων και του μεγέθους δειγμάτων κάνουν τη χρήση της ελκυστική. Εντούτοις, η εμπειρία μας με τη δοκιμή Maddala και Wu είναι κάπως λιγότερο ενθαρρυντική.

- Ο Hadri (2000) υποστηρίζει ότι η μηδενική υπόθεση πρέπει να αντιστραφεί ώστε να γίνει στάσιμη προκειμένου να υπάρξει ένας έλεγχος με την ισχυρότερη δύναμη. Αυτό είναι μια γενίκευση του ελέγχου των Kwiatkowski, Phillips, Schmidt, and Shin (KPSS) από τη χρονική σειρά στα στοιχεία πάνελ. Ο έλεγχος του Hadri βασίζεται στα κατάλοιπα που προέρχονται από τις μεμονωμένες παλινδρομήσεις με τη μέθοδο των ελαχίστων τετραγώνων (OLS) από το παρακάτω υπόδειγμα της παλινδρόμησης:

$$y_{it} = \pi_i + \theta_i t + \mu_{it} = \pi_i + \theta_i t + \sum_{s=1}^t u_{is} + \varepsilon_{it} \quad (13)$$

όπου $\mu_{it} = \sum_{s=1}^t u_{is} + \varepsilon_{it}$

Η μηδενική υπόθεση της στασιμότητας είναι: $H_0 : \sigma_{it}^2 = 0$ οπότε σ' αυτή την περίπτωση $\mu_{it} = \varepsilon_{it}$. Λαμβάνοντας υπόψη τα κατάλοιπα $\hat{\mu}$ από τις μεμονωμένες παλινδρομήσεις, η στατιστική του πολλαπλασιαστή Lagrange (LM) είναι:

$$LM_I = \left[\left((1/N) \sum_{i=1}^N (1/T^2) \sum_{t=1}^T S_{it}^2 \right) / \hat{\sigma}_\varepsilon^2 \right], 1 = \mu, T \quad (14)$$

όπου S_{it}^l είναι το συσσωρευτικό άθροισμα των καταλοίπων $S_{it}^l = \sum_{j=1}^t \varepsilon_{ij}^l, 1 = \mu, T$.

Ο Hadri (2000) εξετάζει τις τυποποιημένες στατιστικές ως εξής:

$$Z_\mu = \sqrt{N}(LM_\mu - \xi_\mu) / \zeta_\mu \Rightarrow N(0,1) \quad (15)$$

και

$$Z_T = \sqrt{N}(LM_T - \xi_T) / \zeta_T \Rightarrow N(0,1) \quad (16)$$

Ο μέσος και η διακύμανση της τυχαίας μεταβλητής Z_μ είναι $\xi_\mu = 1/6$ και $\zeta_\mu^2 = 1/45$ αντίστοιχα. Ο μέσος και η διακύμανση της τυχαίας μεταβλητής Z_T είναι $\xi_T = 1/15$ και $\zeta_T^2 = 11/6300$ αντίστοιχα.

3. ΔΕΔΟΜΕΝΑ ΚΑΙ ΕΜΠΕΙΡΙΚΑ ΑΠΟΤΕΛΕΣΜΑΤΑ

3.1. ΔΕΔΟΜΕΝΑ

Όλα τα στοιχεία της παρούσας εργασίας λήφθηκαν από τη Ευρωπαϊκή Κεντρική Τράπεζα. Το δείγμα αποτελείται από μηνιαία στοιχεία με διαφορετικές χρονικές περιόδους για κάθε χώρα που καθορίζονται από τα διαθέσιμα στοιχεία. Η

περίοδος δειγμάτων είναι 1991:1 έως 2010:2 στη Βουλγαρία, τη Δημοκρατία της Τσεχίας, την Ουγγαρία, και τη Ρουμανία, 1993:1 έως 2010:2 στην Κύπρο, τη Λετονία, την Πολωνία και τη Σλοβακία, 1995:1 έως 2010:2 στην Εσθονία, τη Λιθουανία, τη Μάλτα και τη Σλοβενία. Για τις ονομαστικές συναλλαγματικές ισοτιμίες χρησιμοποιήθηκαν τα ποσοστά του ευρώ (δηλαδή το πηλίκο του εσωτερικού νομίσματος ανά ευρώ), και ο δείκτης τιμών αγαθών και υπηρεσιών ως μέτρο των τιμών. Όλες οι σειρές εκφράζονται στους λογαρίθμους. Οι ονομαστικές συναλλαγματικές ισοτιμίες μετατράπηκαν στις πραγματικές συναλλαγματικές ισοτιμίες με τη χρησιμοποίηση των δεικτών τιμών του καταναλωτή.

3.2. ΕΜΠΕΙΡΙΚΑ ΑΠΟΤΕΛΕΣΜΑΤΑ

Χρησιμοποιούμε στην αρχή τους κλασικούς ελέγχους της μοναδιαίας ρίζας όπως είναι ο επαυξημένος έλεγχος των Dickey-Fuller και των Phillips-Peron. Η μηδενική υπόθεση και στους δύο ελέγχους προϋποθέτει ότι υπάρχει μοναδιαία ρίζα στα επίπεδα των μεταβλητών. Στην αρχή υιοθετούμε το υπόδειγμα στην εμπειρική ανάλυση χωρίς χρονική τάση διότι η συμπερίληψη της γραμμικής χρονικής τάσης θα ήταν θεωρητικά ασυμβίβαστη με τη μακροχρόνια σχέση στην ισοτιμία της αγοραστικής δύναμης (βλέπε Zhang και Lowinger, 2006). Ένα επιχείρημα όμως για τον συνυπολογισμό της χρονικής τάσης είναι η μη στασιμότητα των πραγματικών συναλλαγματικών ισοτιμιών. Έτσι υπολογίσαμε το υπόδειγμα και στους δύο ελέγχους με σταθερά και με σταθερά και τάση. Το βέλτιστο μήκος των χρονικών υστερήσεων για το ADF επιλέχτηκαν από κριτήριο του Akaike. Το αποτέλεσμα των ελέγχων ADF και PP παρουσιάζεται στον πίνακα 1.

Πίνακας 1. Έλεγχοι μοναδιαίας ρίζας για συναλλαγματική ισοτιμία

ΧΩΡΕΣ	ADF		PP	
	Σταθερά	Σταθερά και Τάση	Σταθερά	Σταθερά και Τάση
ΒΟΥΛΓΑΡΙΑ	-1.328 (0)	-2.000 (0)	-1.359 [2]	-2.043 [2]
ΕΣΘΟΝΙΑ	-1.489 (1)	-2.395 (1)	-1.268 [3]	-2.286 [1]
ΚΥΠΡΟΣ	-1.439 (1)	-1.701 (1)	-1.134 [2]	-1.467 [2]
ΛΕΤΟΝΙΑ	-0.760 (1)	-2.069 (1)	-0.860 [6]	-1.877 [6]
ΛΙΘΟΥΑΝΙΑ	-1.932 (1)	-2.545 (1)	-1.736 [5]	-2.271 [5]
ΜΑΛΤΑ	-1.194 (1)	-2.259 (1)	-1.071 [4]	-2.076 [3]
ΟΥΓΓΑΡΙΑ	-2.473 (1)	-1.203 (1)	-2.495 [6]	-1.131 [5]
ΠΟΛΩΝΙΑ	-2.663 (1)*	-2.351 (1)	-2.788 [7]*	-2.251 [6]
ΡΟΥΜΑΝΙΑ	-2.468 (9)	-2.346 (9)	-2.815 [17]*	-2.591 [16]
ΣΛΟΒΑΚΙΑ	-0.264 (1)	-1.142 (1)	-0.149 [6]	-1.033 [5]
ΣΛΟΒΕΝΙΑ	-2.461 (2)	-1.898 (2)	-2.851 [2]*	-2.163 [3]
ΤΣΕΧΙΑ	-0.859 (1)	-1.791 (1)	-0.689 [5]	-1.610 [5]

Σημειώσεις: 1. ***, **, * για επίπεδα σημαντικότητας 1, 5 και 10 αντίστοιχα. 2. Οι μεταβλητές είναι στους φυσικούς λογαρίθμους. 3. Οι αριθμοί μέσα στις παρενθέσεις για τις στατιστικές του ADF, αντιπροσωπεύουν τον αριθμό των χρονικών υστερήσεων της εξαρτημένης μεταβλητής που χρησιμοποιείται για τα σφάλματα του λευκού θορύβου (διόρθωση της αυτοσυσχέτισης). 4. Ο αριθμός των χρονικών υστερήσεων για τις στατιστικές του ADF επιλέχτηκε

χρησιμοποιώντας το κριτήριο του Akaike (AIC). 5. Οι αριθμοί μέσα στις αγκύλες για τις στατιστικές των PP αναφέρονται στο εύρος ζώνης της αυτοσυσχέτισης βασισμένο στον εκτιμητή των Newey-West (1994) χρησιμοποιώντας το στατιστικό του Bartlett (διόρθωση της ετεροσκεδαστικότητας και της αυτοσυσχέτισης των καταλοίπων).

Ο πίνακας 1 δείχνει ότι τα αποτελέσματα και των δύο ελέγχων δεν μπορούν να απορρίψουν τη μηδενική υπόθεση της μοναδιαίας ρίζας σε επίπεδο 5% για το υπόδειγμα με σταθερά και με σταθερά και τάση. Για τις χώρες Πολωνία, Ρουμανία και Σλοβενία η στάσιμη υπόθεση θα μπορούσε να είναι αποδεκτή σε επίπεδο σημαντικότητας 10%. Το γενικό αποτέλεσμα υπονοεί ότι τόσο ο έλεγχος των Dickey-Fuller όσο και ο έλεγχος των Phillips-Peron δεν υποστηρίζουν την ισοτιμία της αγοραστικής δύναμης για τις 12 νέες χώρες της Ε.Ε. Το αποτέλεσμα αυτό συμφωνεί με την υπάρχουσα αρθρογραφία. Σύμφωνα με τα αποτελέσματα των δύο ελέγχων και οι 12 χώρες παρουσιάζουν μοναδιαία ρίζα στα επίπεδα των μεταβλητών στην πραγματική συναλλαγματική ισοτιμία. Ωστόσο, και οι δύο έλεγχοι έχουν χαμηλή ισχύ με το μέγεθος του δείγματος. Αν και υπάρχουν μέχρι και 228 μηνιαίες παρατηρήσεις εντούτοις αυτές εκτείνονται μόνο σε 19 έτη. Επομένως, ένας πιθανός λόγος για την αποτυχία των ελέγχων ADF και PP μπορεί να είναι το μικρό μέγεθος του δείγματος που εξετάζουμε για την κάθε χώρα. Τη πιθανότητα αυτής της αποτυχίας την ερευνούμε με τη χρησιμοποίηση των ελέγχων της μοναδιαίας ρίζας σε στοιχεία πάνελ για τις 12 χώρες της Ε.Ε. Τα αποτελέσματα των ελέγχων της μοναδιαίας ρίζας σε πάνελ στοιχεία παρουσιάζονται στον πίνακα 2.

Πίνακας 2. Ελεγχοι μοναδιαίας ρίζας για συναλλαγματική ισοτιμία σε πάνελ δεδομένα

Ελεγκοι		Στατιστικό	Πιθανότητα
LLC	Σταθερά	-4.12197	(0.000)***
	Σταθερά και Τάση	-3.879	(0.000)***
Breitung	Σταθερά		
	Σταθερά και Τάση	2.744	(0.997)
IPS	Σταθερά	-1.35353	(0.087)*
	Σταθερά και Τάση	0.157	(0.562)
Fisher - ADF	Σταθερά	45.473	(0.005)***
	Σταθερά και Τάση	23.620	(0.483)
Fisher - PP	Σταθερά	70.288	(0.000)***
	Σταθερά και Τάση	17.271	(0.836)
Hadri	Σταθερά	25.512	(0.000)***
	Σταθερά και Τάση	15.642	(0.000)***

Σημειώσεις: 1. Στα πάνελ δεδομένα συμπεριλαμβάνονται και οι 12 χώρες. 2. Οι μεταβλητές είναι στους φυσικούς λογαρίθμους. 3. Οι αριθμοί στις παρενθέσεις σημειώνουν την πιθανότητα. 4. ***, **, * σημειώνουν την απόρριψη της μηδενικής υπόθεσης σε επίπεδο σημαντικότητας 1%, 5% and 10% αντίστοιχα. 5. Η μηδενική υπόθεση για όλους τους ελέγχους για τα δεδομένα πάνελ είναι ότι υπάρχει μοναδιαία ρίζα (οι σειρές δεν είναι στάσιμες) εκτός από τον έλεγχο του Hadri που η μηδενική υπόθεση είναι ότι δεν υπάρχει μοναδιαία ρίζα.

Σε αντίθεση με τους ελέγχους του πίνακα 1 τα περισσότερα αποτελέσματα των ελέγχων στα δεδομένα πάνελ απορρίπτουν τη μηδενική υπόθεση για τη μοναδιαία ρίζα που δείχνει ότι οι πραγματικές σειρές συναλλαγματικής ισοτιμίας είναι

στάσιμες. Το αποτέλεσμα αυτό υποστηρίζει την ισοτιμία της αγοραστικής δύναμης για τις 12 νέες χώρες της Ε.Ε.

4. ΣΥΜΕΡΑΣΜΑΤΑ

Η εργασία αυτή εξετάζει την ισχύ της υπόθεσης για την ισοτιμία της αγοραστικής δύναμης των 12 νέων χωρών της Ε.Ε. Για το λόγο αυτό, εξετάζουμε τη στασιμότητα της πραγματικής συναλλαγματικής ισοτιμίας. Χρησιμοποιώντας τους κλασικούς ελέγχους της μοναδιαίας ρίζας των Dickey-Fuller και των Phillips-Peron, διαπιστώνουμε ότι οι πραγματικές σειρές συναλλαγματικής ισοτιμίας όλων των χωρών της Ε.Ε. παρουσιάζουν μοναδιαία ρίζα. Το αποτέλεσμα αυτό πιθανόν να ισχύει λόγω της χαμηλής δύναμης που έχουν οι έλεγχοι αυτοί. Ο στόχος αυτής της εργασίας είναι να επανεξεταστεί αυτό το ζήτημα με τις πρόσθετες πληροφορίες που παρέχονται από το συνδυασμό των διαχρονικών και διαστρωματικών στοιχείων, καθώς και τα πλεονεκτήματα που παρουσιάζουν οι έλεγχοι για τα δεδομένα αυτά. Τα εμπειρικά αποτελέσματα από τη χρήση των ελέγχων αυτών έδειξαν ότι οι πραγματικές συναλλαγματικές ισοτιμίες μεταξύ των 12 νέων χωρών της Ε.Ε. είναι στάσιμες. Επομένως υπάρχει μακροχρόνια σχέση στην ισοτιμία της αγοραστικής δύναμης των χωρών αυτών.

ABSTRACT

This paper investigates the validity of purchasing power parity (PPP) for the case of the 12 new EU countries by using a panel unit-root methodology. The aim of the paper is to examine the stationarity of real exchange rate. By applying the Augmented Dickey Fuller (ADF) and Phillips-Peron (PP) tests for stationarity on single time series, it is found that real exchange rate for the countries under examination have unit root. This outcome, however, might be due to the general low power of these tests. The current paper reconsiders this issue by exploiting the extra information provided by the combination of the time-series and cross-sectional data and the subsequent power advantages of panel data unit-root tests. According to estimation results real exchange rate in the 12 new countries are stationary. Therefore, there is long-run purchasing power parity in the 12 new countries.

ΑΝΑΦΟΡΕΣ

- Bahmani-Oskooee, M., (1998). Do exchange rates follow a random walk process in Middle Eastern countries? *Economics Letters*, 58, 339-344.
- Breitung, J. (2000). The local power of some unit root tests for panel data, in Baltagi, B.(ed). *Advances in Econometrics. Non-stationary panels, panel cointegration, and dynamic panels*, JAI Press, Amsterdam, pp. 161-178.
- Corbae, D. and Ouliaris, S., (1988). Cointegration and tests of purchasing power parity. *Review of Economics and Statistics*, 70, 508-511.
- Dickey D.A. and Fuller W.A, (1979). Distributions of the estimators for autoregressive time series with a unit root, *Journal of American Statistical Association*, 74, 427- 431.
- Dickey D.A and Fuller W.A, (1981). Likelihood ratio statistics for autoregressive time series with a unit root, *Econometrica*, 49, 1057-1072.

- Doganlar, M., (1999). Testing long-run validity of purchasing power parity for Asian countries. *Applied Economics Letters*, 6, 147-151.
- Frankel, J. and Rose, A.K., (1996). A Panel project on purchasing power parity: mean reversion within and between economies. *Journal of International Economics*, 40 (1-2), 209-224.
- Hadri, K. (2000). Testing for stationarity in heterogeneous panel data. *Econometric Journal*, H., & Shin, Y. (2003). Testing for unit roots in heterogeneous panels, *Journal of Econometrics*, Vol.115, pp.53-74.
- Hakkio, C.C., (1986). Does the exchange rate follow a random walk? A Monte Carlo study of four tests for a random walk. *Journal of International Money and Finance*, 5, 221-229.
- Im, K.S., Pesaran, M.H., and Shin, Y. (2003). Testing for unit roots in heterogeneous panels, *Journal of Econometrics*, Vol.115, pp.53-74.
- Kim, Y., (1990). Purchasing power parity in the long run: A cointegration approach. *Journal of Money Credit and Banking*, 4, 491-503.
- Kugler, P. and Lenz, C., (1993). Multivariate cointegration analysis and the long-run validity of PPP. *Review of Economics and Statistics*, 75, 180-184.
- Kwiatkowski, D., P. Phillips, P. Schmidt, and Y. Shin (1992). Testing the null hypothesis of stationarity against the alternative of a unit root. *Journal of Econometrics*, 54, 159-178.
- Levin, A., Lin, C.F., and Chu, J. (2002). Unit root test in panel data: Asymptotic and finite-sample properties, *Journal of econometrics*, Vol.108, pp.1-24.
- MacDonald, R., (1993). Long-run purchasing power parity: is it for real? *Review of Economics and Statistics*, 75, 690-695.
- Maddala, G. S., and Wu, S. (1999). A comparative study of unit root tests with panel data and a new simple test, *Oxford Bulletin of Economics and Statistics*, Vol.61, pp. 631-652.
- Marcela, S., Gadea, M.D. and Serrano, J.M., (2003). PPP and structural breaks. The peseta-sterling rate, 50 years of floating regime. *Journal of International Money and Finance*, 22, 613-627.
- Narayan, P.K., (2005). New evidence on purchasing power parity from 17 OECD countries. *Applied Economics*, 37, 1063-1071.
- Narayan, P.K. and Prasad, B.C., (2005). The validity of purchasing power parity hypothesis for eleven middle eastern countries. *Review of Middle East Economics and Finance*, 3 (2), 135-149.
- Narayan, P.K., (2007). Are nominal exchange rates and price levels co-integrated? New evidence from threshold autoregressive and momentum threshold autoregressive models. *Economic Record*, 83 (260), 74-85.
- Newey, W.K. and K.D. West (1994). Automatic lag selection in covariance matrix estimation. *Review of Economic Studies*, 61, 631-653.
- Papell, D. and Theodoridis, H., (1998). Increasing evidence of purchasing power over the current float. *Journal of International Money and Finance*, 17 (1), 41-50.
- Phillips, P. (1987). Time series regression with a unit root. *Econometrica*, 55, 277-301.

- Phillips, P.C. and Perron, P., (1988). Testing for a Unit Root in Time Series Regression, *Biometrika*, 75, 335 - 346.
- Serletis, A. and Zimonopoulos, G., (1997). Breaking trend functions in real exchange rates: evidence from seventeen OECD countries. *Journal of Macroeconomics*, 19, 781-802.
- Sideris, (2011). Optimum currency areas, structural changes and the endogeneity of the OCA criteria: Evidence from six new EU member states. *Applied Financial Economics*, 21, 195-206.
- Wu, J.L. and Wu, S., (2001). Are purchasing power parity overvalued. *Journal of Money, Credit and Banking*, 33, 804-812.
- Zhang, S. and Lowinger, T.C., (2006). An empirical test of purchasing power parity in selected developing countries: a panel data approach. *International Economic Journal*, 20 (1), 79-86.



ΑΝΑΛΥΣΗ ΑΝΤΙΣΤΟΙΧΙΩΝ ΜΕ ΑΣΑΦΗ ΔΕΔΟΜΕΝΑ: Η ΠΡΑΚΤΙΚΗ ΕΦΑΡΜΟΓΗ ΤΗΣ ΜΕΘΟΔΟΥ (ΑΝΤΙΠΡΟΣΩΠΕΥΤΙΚΟ ΠΑΡΑΔΕΙΓΜΑ ΜΕ ΑΠΟΑΣΑΦΟΠΟΙΗΜΕΝΗ ΠΑΡΑΓΟΝΤΙΚΗ ΓΕΩΜΕΤΡΙΚΗ ΠΑΡΑΣΤΑΣΗ)

Ιωάννης Θεοδώρου¹, Φίλιππος Αλεβίζος², Αριστείδης Κεχρινιώτης¹

¹Τμήμα Ηλεκτρονικής, Τ.Ε.Ι. Λαμίας

<mailto:{teo, kechrin}@teilam.gr>

²Τμήμα Μαθηματικών, Πανεπιστήμιο Πατρών

philippos@math.upatras.gr

ΠΕΡΙΛΗΨΗ

Η εργασία αυτή είναι άμεση συνέχεια των προηγούμενων εργασιών μας [6,7,8], όπου αναπτύχθηκε η μαθηματική θεμελίωση της επέκτασης της Ανάλυσης Αντιστοιχιών (*Correspondence Analysis-CA*) στην επεξεργασία ασαφών δεδομένων. Δεδομένου δε ότι αλγεβρικά η Ανάλυση Αντιστοιχιών είναι κυρίως ένα πρόβλημα ιδιοδομής, μία μέθοδος «δύο-βημάτων» αναπτύχθηκε για τη θεωρητική επεξεργασία του προβλήματος της ασαφούς ιδιοτιμής, όταν η CA εφαρμόζεται σε ένα πίνακα τυχαιότητας με ασαφή δεδομένα (*Correspondence Analysis with Fuzzy Data-CAFD*).

Στην εργασία αυτή ολοκληρώνουμε τα προηγούμενα θεωρητικά αποτελέσματα με την πρακτική εφαρμογή της CAFD, δηλ. παρουσιάζουμε τα γεωμετρικά (γραφικά) αποτελέσματα της CA όταν αυτή εφαρμόζεται σε έναν πίνακα τυχαιότητας που τα στοιχεία του είναι ασαφείς αριθμοί. Προς τούτο, θεωρούμε ένα απλό αλλά πραγματικό πρόβλημα της CAFD και όλες οι χρησιμοποιούμενες υπολογιστικές διαδικασίες (αλγόριθμοι) παρουσιάζονται αναλυτικά μέσω ενός στοιχειώδους αλλά αντιπροσωπευτικού αριθμητικού παραδείγματος. Για τον κύριο γεωμετρικό χαρακτήρα της CAFD και με σκοπό την προβολή των ασαφών δεδομένων (*fuzzy profiles*) σε ένα χώρο μικρότερης διάστασης, ένα νέο βήμα-αποασαφοποίησης (*defuzzification-step*) έχει αναπτυχθεί. Τέλος, σε μια ενιαία παραγοντική γραφική παράσταση, επιτυγχάνουμε να παρουσιάσουμε τα γεωμετρικά αποτελέσματα της CAFD ταυτόχρονα και σε σύγκριση με αυτά της κλασικής CA.

Ο κύριος στόχος αυτής της εργασίας είναι να δειχτεί ότι η CAFD μπορεί να λειτουργήσει όχι μόνο θεωρητικά, αλλά και πρακτικά-υπολογιστικά και κυρίως γεωμετρικά (γραφικά). Συνεπώς το τελικό βήμα που απομένει για την τελική πρακτική χρήση της CAFD είναι πλέον η δημιουργία του κατάλληλου λογισμικού (*software*).

Λέξεις Κλειδιά: Ανάλυση Αντιστοιχιών, Ασαφείς Αριθμοί, Ασαφείς Πίνακες, Ασαφής Ιδιοτιμή, Μέθοδοι Αποασαφοποίησης, Ασαφής Στατιστική και Ανάλυση Δεδομένων.

1. ΕΙΣΑΓΩΓΗ

Στις προηγούμενες εργασίες μας [6,7,8] αναπτύχθηκε η μαθηματική θεμελίωση της επέκτασης της κλασικής Ανάλυσης Αντιστοιχιών (*Correspondence Analysis-CA*), έτσι ώστε αυτή να καταστεί ικανή για την εφαρμογή της σε πολυδιάστατους πίνακες τυχαιότητας με Ασαφή Δεδομένα (*Correspondence Analysis with Fuzzy Data-CAFD*). Στην εργασία αυτή ολοκληρώνουμε τα προηγούμενα θεωρητικά αποτελέσματα με την πρακτική εφαρμογή της CAFD, δηλ. παρουσιάζουμε τα τελικά γεωμετρικά (γραφικά) αποτελέσματα της CA όταν αυτή εφαρμόζεται σε έναν πίνακα τυχαιότητας που τα στοιχεία του είναι ασαφείς αριθμοί.

2. ΠΡΑΚΤΙΚΗ ΕΦΑΡΜΟΓΗ ΚΑΙ ΓΕΩΜΕΤΡΙΚΗ-ΓΡΑΦΙΚΗ ΠΑΡΑΣΤΑΣΗ ΤΗΣ CAFD

2.1 Αντιπροσωπευτικό αριθμητικό παράδειγμα

Έστω 3 ομάδες υποψήφιων σπουδαστών που προετοιμάζονται για την εισαγωγή τους σε ένα πανεπιστήμιο, όπου κάθε ομάδα εκπροσωπεί μια διαφορετική γεωγραφική κοινότητα, που συμβολίζονται: *stud-1*, *stud-2*, *stud-3*. Με σκοπό την επιτυχία τους στο συγκεκριμένο πανεπιστήμιο, οι υποψήφιοι σπουδαστές μπορούν να επιλέξουν να συμμετέχουν σε ένα μόνο από τα 3 τμήματα-κατευθύνσεις του πανεπιστημίου, που είναι: Μαθηματικό (*MATH*), Φυσικό (*PHYS*), Φιλολογία (*LITR*).

Ειδικοί εκπαιδευτικοί αναλυτές κάνουν μια προσεγγιστική πρόβλεψη-εκτίμηση, χρησιμοποιώντας Τριγωνικούς Ασαφείς Αριθμούς (*Triangular Fuzzy Numbers-TFN*), σχετικά με το πόσοι υποψήφιοι από κάθε «σπουδαστική κοινότητα» θα επιτύχουν σε καθένα «πανεπιστημιακό τμήμα», που καταγράφεται στον ακόλουθο Πίνακα 1.

Πίνακας 1. Επιτυχία ως προς Σπουδαστικές Κοινότητες και Πανεπιστημιακό Τμήμα:

	<i>MATH</i>	<i>PHYS</i>	<i>LITR</i>
<i>stud-1</i>	$k_1^1 = (6,8,10)$	$k_1^2 = (20,23,26)$	$k_1^3 = (12,18,24)$
<i>stud-2</i>	$k_2^1 = (15,20,25)$	$k_2^2 = (30,40,50)$	$k_2^3 = (17,17,17)$
<i>stud-3</i>	$k_3^1 = (1,2,3)$	$k_3^2 = (5,5,5)$	$k_3^3 = (7,8,9)$

Συνεπώς παίρνουμε έτσι, έναν (3×3) ασαφή μη-αρνητικό πίνακα $K = (k_{ij}^j)$, $(i, j = 1, 2, 3)$, (*ασαφή πίνακα τυχαιότητας-fuzzy contingency table*), που δίνει «την επιτυχία των υποψηφίων» ως προς «τις σπουδαστικές κοινότητες» και τα «πανεπιστημιακά τμήματα», δηλ. κατανέμει τις συμπτώσεις δύο ποιοτικών μεταβλητών I, J, όπου οι γραμμές (I) είναι «οι σπουδαστικές ομάδες-γεωγραφικές κοινότητες» *stud-1*, *stud-2*, *stud-3*, και οι στήλες (J) είναι «τα πανεπιστημιακά τμήματα» *MATH*, *PHYS* και *LITR*.

$$S_{3,3} = X^T \otimes D \otimes X \otimes Q = (s_i^j) = (l_i^j, c_i^j, r_i^j) \cong$$

$$\begin{bmatrix} s_1^1=(0.04,0.23,1.36) & s_1^2=(0.04,0.22,1.15) & s_1^3=(0.04,0.20,0.95) \\ s_2^1=(0.08,0.50,2.88) & s_2^2=(0.10,0.49,2.46) & s_2^3=(0.11,0.47,2.11) \\ s_3^1=(0.05,0.28,1.56) & s_3^2=(0.06,0.29,1.38) & s_3^3=(0.08,0.34,1.62) \end{bmatrix}$$

Ακολουθώντας υπολογίζουμε μέσω της μεθόδου «δύο βημάτων (*two-step*)», στο 1^ο βήμα τον πίνακα διαστημάτων (α -cut) S^α του ασαφούς πίνακα $S_{3,3}$:

$$S^\alpha = [S_\ell^\alpha, S_r^\alpha] = [(s_i^j)_\ell^\alpha, (s_i^j)_r^\alpha] = [l_i^j + \alpha(c_i^j - l_i^j), r_i^j - \alpha(r_i^j - c_i^j)] =$$

$$(X^T \otimes D \otimes X \otimes Q)^\alpha = [(X^T \otimes D \otimes X \otimes Q)_\ell^\alpha, (X^T \otimes D \otimes X \otimes Q)_r^\alpha] =$$

$$\begin{bmatrix} [0.04+0.19\alpha, 1.36-1.13\alpha] & [0.04+0.18\alpha, 1.15-0.93\alpha] & [0.04+0.16\alpha, 0.95-0.95\alpha] \\ [0.08+0.42\alpha, 2.88-2.38\alpha] & [0.10+0.39\alpha, 2.46-1.97\alpha] & [0.11+0.36\alpha, 2.11-1.64\alpha] \\ [0.05+0.23\alpha, 1.56-1.28\alpha] & [0.06+0.23\alpha, 1.38-1.09\alpha] & [0.08+0.26\alpha, 1.62-1.28\alpha] \end{bmatrix},$$

όπου,

$$X^\alpha = [X_\ell^\alpha, X_r^\alpha] = [(x_i^j)_\ell^\alpha, (x_i^j)_r^\alpha] =$$

$$\begin{bmatrix} [0.1+0.06\alpha, 0.26-0.1\alpha] & [0.33+0.14\alpha, 0.68-0.21\alpha] & [0.2+0.17\alpha, 0.63-0.26\alpha] \\ [0.16+0.1\alpha, 0.40-0.14\alpha] & [0.33+0.19\alpha, 0.81-0.29\alpha] & [0.18+0.04\alpha, 0.27-0.05\alpha] \\ [0.06+0.07\alpha, 0.23-0.1\alpha] & [0.29+0.04\alpha, 0.38-0.05\alpha] & [0.41+0.12\alpha, 0.69-0.16\alpha] \end{bmatrix}$$

$$D^\alpha = [D_\ell^\alpha, D_r^\alpha] = [(d_i^j)_\ell^\alpha, (d_i^j)_r^\alpha] =$$

$$\begin{bmatrix} [0.23+0.12\alpha, 0.53-0.18\alpha] & 0 & 0 \\ 0 & [0.37+0.18\alpha, 0.82-0.27\alpha] & 0 \\ 0 & 0 & [0.08+0.03\alpha, 0.15-0.04\alpha] \end{bmatrix}$$

$$Q^\alpha = [Q_\ell^\alpha, Q_r^\alpha] = [(q_i^j)_\ell^\alpha, (q_i^j)_r^\alpha] =$$

$$\begin{bmatrix} [2.97+1.73\alpha, 7.69-2.99\alpha] & 0 & 0 \\ 0 & [1.39+0.68\alpha, 2.75-0.68\alpha] & 0 \\ 0 & 0 & [2.26+1.02\alpha, 4.7-1.42\alpha] \end{bmatrix}$$

και

$$(Q^\alpha)^{1/2} = \begin{bmatrix} [\sqrt{2.97+1.73\alpha}, \sqrt{7.69-2.99\alpha}] & 0 & 0 \\ 0 & [\sqrt{1.39+0.68\alpha}, \sqrt{2.75-0.68\alpha}] & 0 \\ 0 & 0 & [\sqrt{2.26+1.02\alpha}, \sqrt{4.7-1.42\alpha}] \end{bmatrix}$$

Επιπλέον στο 2^ο βήμα της “two-step” μεθόδου, από τον πίνακα διαστημάτων S^α μπορούμε να υπολογίσουμε τους συνήθεις αριθμητικούς τετραγωνικούς πίνακες:

$$S_{\ell}^{\alpha} = (X_{\ell}^{\alpha})^T D_{\ell}^{\alpha} X_{\ell}^{\alpha} Q_{\ell}^{\alpha} = \begin{bmatrix} 0.04+0.19\alpha & 0.04+0.18\alpha & 0.04+0.16\alpha \\ 0.08+0.42\alpha & 0.10+0.39\alpha & 0.11+0.36\alpha \\ 0.05+0.23\alpha & 0.06+0.23\alpha & 0.08+0.26\alpha \end{bmatrix}$$

$$S_{r}^{\alpha} = (X_{r}^{\alpha})^T D_{r}^{\alpha} X_{r}^{\alpha} Q_{r}^{\alpha} = \begin{bmatrix} 1.36-1.13\alpha & 1.15-0.93\alpha & 0.95-0.95\alpha \\ 2.88-2.38\alpha & 2.46-1.97\alpha & 2.11-1.64\alpha \\ 1.56-1.28\alpha & 1.38-1.09\alpha & 1.62-1.28\alpha \end{bmatrix}.$$

Επίσης μπορούμε να υπολογίσουμε τον ισοδύναμο με τον S , ασαφή συμμετρικό

πίνακα A , δηλ.: $A = Q_{3,3}^{\frac{1}{2}} \otimes X_{3,3}^T \otimes D_{3,3} \otimes X_{3,3} \otimes Q_{3,3}^{\frac{1}{2}} \cong$

$$\cong \begin{bmatrix} a_1^1=(0.04,0.22,1.34) & a_1^2=(0.06,0.32,1.71) & a_1^3=(0.04,0.23,1.2) \\ a_2^1=(0.06,0.32,1.71) & a_2^2=(0.1,0.49,2.22) & a_2^3=(0.08,0.37,1.6) \\ a_3^1=(0.04,0.23,1.2) & a_3^2=(0.08,0.37,1.6) & a_3^3=(0.08,0.34,1.6) \end{bmatrix}$$

με αντίστοιχο πίνακα διαστημάτων, σύμφωνα με την δύο βημάτων (*two-step*) μέθοδο,

$$A^{\alpha} = [A_{\ell}^{\alpha}, A_{r}^{\alpha}] = [(a_i^j)_{\ell}^{\alpha}, (a_i^j)_{r}^{\alpha}] =$$

$$= [(Q_{\ell}^{\alpha})^{\frac{1}{2}} (X_{\ell}^{\alpha})^T D_{\ell}^{\alpha} X_{\ell}^{\alpha} (Q_{\ell}^{\alpha})^{\frac{1}{2}}, (Q_{r}^{\alpha})^{\frac{1}{2}} (X_{r}^{\alpha})^T D_{r}^{\alpha} X_{r}^{\alpha} (Q_{r}^{\alpha})^{\frac{1}{2}}]$$

$$= \begin{bmatrix} [0.04+0.18\alpha, 1.34-1.12\alpha] & [0.06+0.26\alpha, 1.71-1.39\alpha] & [0.04+0.19\alpha, 1.2-0.97\alpha] \\ [0.06+0.26\alpha, 1.71-1.39\alpha] & [0.1+0.39\alpha, 2.22-1.73\alpha] & [0.08+0.29\alpha, 1.6-1.23\alpha] \\ [0.04+0.19\alpha, 1.2-0.97\alpha] & [0.08+0.29\alpha, 1.6-1.23\alpha] & [0.08+0.26\alpha, 1.6-1.26\alpha] \end{bmatrix}$$

$$= \{(1-\theta)A_{\ell}^{\alpha} + \theta A_{r}^{\alpha}\} = A(\alpha, \theta), \quad \text{για κάθε } \alpha, \theta \in [0, 1],$$

από όπου μπορούμε προφανώς να υπολογίσουμε τους αντίστοιχους συνήθεις αριθμητικούς πίνακες $A(\alpha, \theta)$, για οποιαδήποτε τιμή των $\alpha, \theta \in [0, 1]$.

2.3 Υπολογισμός των παραγοντικών συντεταγμένων της CAFD

Σύμφωνα με τη μέθοδο «δύο βημάτων», στο 1^ο βήμα, μετατρέπουμε τον ασαφή πίνακα των προφίλ-γραμμών X στην οικογένεια των (α -cut) διαστημάτων πινάκων,

$$X^{\alpha} = [X_{\ell}^{\alpha}, X_{r}^{\alpha}] = [(x_i^j)_{\ell}^{\alpha}, (x_i^j)_{r}^{\alpha}], \quad \text{για } \alpha \in [0, 1].$$

Στο 2^ο βήμα, κάθε πίνακας διαστημάτων X^{α} μπορεί να γραφεί, $\forall \alpha \in [0, 1]$:

$$X^{\alpha} = [X_{\ell}^{\alpha}, X_{r}^{\alpha}] := \{(1-\theta)X_{\ell}^{\alpha} + \theta X_{r}^{\alpha}, \quad \theta \in [0, 1]\} = X(\alpha, \theta).$$

Έτσι, ο συνήθης πίνακας των προφίλ-γραμμών, που υπολογίζεται για $\alpha=1$ και $\theta=0$ (είτε $\theta=1$), είναι

$$X^{\alpha=1} = [X_{\ell}^1, X_{r}^1] = [(x_i^j)_{\ell}^1, (x_i^j)_{r}^1] = \{(1-\theta)(x_i^j)_{\ell}^1 + \theta(x_i^j)_{r}^1\} = X_{(\alpha=1, \theta=0)} = \begin{bmatrix} 0.16 & 0.47 & 0.37 \\ 0.26 & 0.52 & 0.22 \\ 0.13 & 0.33 & 0.53 \end{bmatrix},$$

και έχει προφανώς ως προς τους κλασικούς παραγοντικούς άξονες των κλασικών ιδιοδιανυσμάτων w_s ($s=2,3$), τις ίδιες παραγοντικές συντεταγμένες (factorial coordinates) όπως στην κλασική CA, που είναι οι ακόλουθες:

$$\Psi_{w_2}(X)_{(\alpha=1, \theta=0)} = \text{proj}_{w_2}(X)_{(\alpha=1, \theta=0)} = (XQ^{1/2}w_2)_{(\alpha=1, \theta=0)} = \begin{pmatrix} 0.15 \\ -0.19 \\ 0.49 \end{pmatrix},$$

$$\Psi_{w_3}(X)_{(\alpha=1, \theta=0)} = \text{proj}_{w_3}(X)_{(\alpha=1, \theta=0)} = (XQ^{1/2}w_3)_{(\alpha=1, \theta=0)} = \begin{pmatrix} -0.05 \\ 0.02 \\ 0.08 \end{pmatrix}.$$

Όμοια, μπορούμε να υπολογίσουμε τις παραγοντικές συντεταγμένες κάθε συνήθους πίνακα $X(\alpha, \theta)$, για κάθε $\alpha, \theta \in [0, 1]$, από τη γνωστή σχέση:

$$\Psi_{w_s}(X)_{(\alpha, \theta \in [0, 1])} = \text{proj}_{w_s}(X)_{(\alpha, \theta \in [0, 1])} = (XQ^{1/2}w_s)_{(\alpha, \theta \in [0, 1])}.$$

3. ΑΠΟΑΣΑΦΟΠΟΙΗΣΗ ΤΩΝ ΑΣΑΦΩΝ ΠΡΟΒΟΛΩΝ-ΠΑΡΑΓΟΝΤΙΚΩΝ ΣΥΝΤΕΤΑΓΜΕΝΩΝ (ΕΞΟΔΩΝ) ΤΗΣ CAFD

Τελικά, όπως και στον ασαφή έλεγχο (fuzzy control), μπορούμε να χρησιμοποιήσουμε διάφορες τεχνικές αποασαφοποίησης (defuzzification) για μια αντιπροσωπευτική κλασική γεωμετρική έξοδο της CAFD, που θα εκπροσωπεί το σύνολο των παραγοντικών συντεταγμένων που παράγονται από τη «δύο βημάτων» μέθοδο. Υπολογίζουμε ακολούθως μερικές παραλλαγές των συνήθων τεχνικών αποασαφοποίησης, προκειμένου να βρεθεί μια αντιπροσωπευτική τιμή-έξοδος όλων των παραγοντικών συντεταγμένων για τις διάφορες τιμές $\alpha, \theta \in [0, 1]$, επί των κλασικών παραγοντικών αξόνων.

1) Κέντρο Βάρους έξοδος (Centroid output):

Προβάλλουμε το Κέντρο Βάρους των στοιχείων του πίνακα των προφίλ-γραμμών X (που εδώ είναι τρίγωνα-TFN, δηλ. για $\alpha = \frac{1}{3}, \theta = \frac{1}{2}$), και υπολογίζουμε τις συντεταγμένες τους ως προς τους κλασικούς παραγοντικούς άξονες w_2, w_3 , σύμφωνα με τους επόμενους εναλλακτικούς τρόπους:

$$\text{a) } \Psi_{w_2(\alpha=1, \theta=0)}(X_{(\alpha=\frac{1}{3}, \theta=\frac{1}{2})}) = \text{proj}_{w_2(\alpha=1, \theta=0)}(X_{(\alpha=\frac{1}{3}, \theta=\frac{1}{2})}) = X_{(\alpha=\frac{1}{3}, \theta=\frac{1}{2})}(Q^{1/2}w_2)_{(\alpha=1, \theta=0)} = \begin{pmatrix} 0.22 \\ -0.22 \\ 0.49 \end{pmatrix} \quad (1)$$

$$\Psi_{w_3(\alpha=1, \theta=0)}(X_{(\alpha=\frac{1}{3}, \theta=\frac{1}{2})}) = \text{proj}_{w_3(\alpha=1, \theta=0)}(X_{(\alpha=\frac{1}{3}, \theta=\frac{1}{2})}) = X_{(\alpha=\frac{1}{3}, \theta=\frac{1}{2})}(Q^{1/2}w_3)_{(\alpha=1, \theta=0)} = \begin{pmatrix} 0.05 \\ 0.004 \\ 0.1 \end{pmatrix} \quad (2)$$

και

$$\text{b) } \Psi_{w_2}(X)_{(\alpha=\frac{1}{3}, \theta=\frac{1}{2})} = \text{proj}_{w_2}(X)_{(\alpha=\frac{1}{3}, \theta=\frac{1}{2})} = (XQ^{1/2}w_2)_{(\alpha=\frac{1}{3}, \theta=\frac{1}{2})} = \begin{pmatrix} 0.25 \\ -0.20 \\ 0.52 \end{pmatrix} \quad (3)$$

$$\Psi_{w_3}(X)_{(\alpha=\frac{1}{3}, \theta=\frac{1}{2})} = \text{proj}_{w_3}(X)_{(\alpha=\frac{1}{3}, \theta=\frac{1}{2})} = (XQ^{1/2}w_3)_{(\alpha=\frac{1}{3}, \theta=\frac{1}{2})} = \begin{pmatrix} -0.004 \\ 0.005 \\ -0.03 \end{pmatrix} \quad (4)$$

2) Αριθμητικού Μέσου έξοδος (Arithmetic Mean output):

$$\Psi_{w_s}(X)_{(\alpha,\theta \in [0,1])} = \text{proj}_{w_s}(X)_{(\alpha,\theta \in [0,1])} = \frac{1}{(\nu+1)(\mu+1)} \sum_{\alpha,\theta \in [0,1]} \{X(\alpha,\theta)Q^{1/2}(\alpha,\theta)w_s(\alpha,\theta)\}, \quad (5)$$

όπου, $\alpha = \frac{0}{\nu}, \frac{1}{\nu}, \dots, \frac{\nu}{\nu}$, $\theta = \frac{0}{\mu}, \frac{1}{\mu}, \dots, \frac{\mu}{\mu}$, και $\nu, \mu \in \mathbb{N}$.

Έτσι, π.χ. για $\alpha, \theta=0, 1$, (δηλ., $\nu, \mu=1$ και $(\nu+1)(\mu+1)=4$), έχουμε:

$$\Psi_{w_2}(X)_{(\alpha,\theta \in [0,1])} = \text{proj}_{w_2}(X)_{(\alpha,\theta \in [0,1])} = \frac{1}{(1+1)(1+1)} \sum_{\alpha,\theta=0,1} \{X(\alpha,\theta)Q^{1/2}(\alpha,\theta)w_2(\alpha,\theta)\} = \quad (5a)$$

$$\frac{1}{4} \{X(0,0)Q^{1/2}(0,0)w_2(0,0) + X(0,1)Q^{1/2}(0,1)w_2(0,1) + X(1,0)Q^{1/2}(1,0)w_2(1,0) + X(1,1)Q^{1/2}(1,1)w_2(1,1)\} \cong \begin{bmatrix} 0.16 \\ -0.25 \\ 0.48 \end{bmatrix}.$$

$$\Psi_{w_3}(X)_{(\alpha,\theta \in [0,1])} = \text{proj}_{w_3}(X)_{(\alpha,\theta \in [0,1])} = \frac{1}{(1+1)(1+1)} \sum_{\alpha,\theta=0,1} \{X(\alpha,\theta)Q^{1/2}(\alpha,\theta)w_3(\alpha,\theta)\} = \quad (5b)$$

$$\frac{1}{4} \{X(0,0)Q^{1/2}(0,0)w_3(0,0) + X(0,1)Q^{1/2}(0,1)w_3(0,1) + X(1,0)Q^{1/2}(1,0)w_3(1,0) + X(1,1)Q^{1/2}(1,1)w_3(1,1)\} = \begin{bmatrix} -0.029851 \\ 0.049425 \\ 0.11566 \end{bmatrix} \cong \begin{bmatrix} -0.03 \\ 0.05 \\ 0.11 \end{bmatrix}.$$

Γενικά, η μέθοδος αυτή είναι πιο αντιπροσωπευτική αλλά απαιτεί προφανώς στη πράξη πολλές τιμές των $\alpha, \theta \in [0, 1]$, για μια ικανοποιητική προσέγγιση των παραγοντικών συντεταγμένων (π.χ., $\nu, \mu \geq 100$).

3.1 Ταυτόχρονη γραφική παράσταση των προβολών της CAFD και CA

Στην επόμενη Εικόνα 1, δίνεται η ταυτόχρονη γραφική παράσταση των παραγοντικών συντεταγμένων της CAFD και CA (με χρήση του SPAD, βλ. [5,9]).

Εικόνα 1: Ταυτόχρονη παραγοντική γραφική παράσταση της CAFD και CA, όπου:

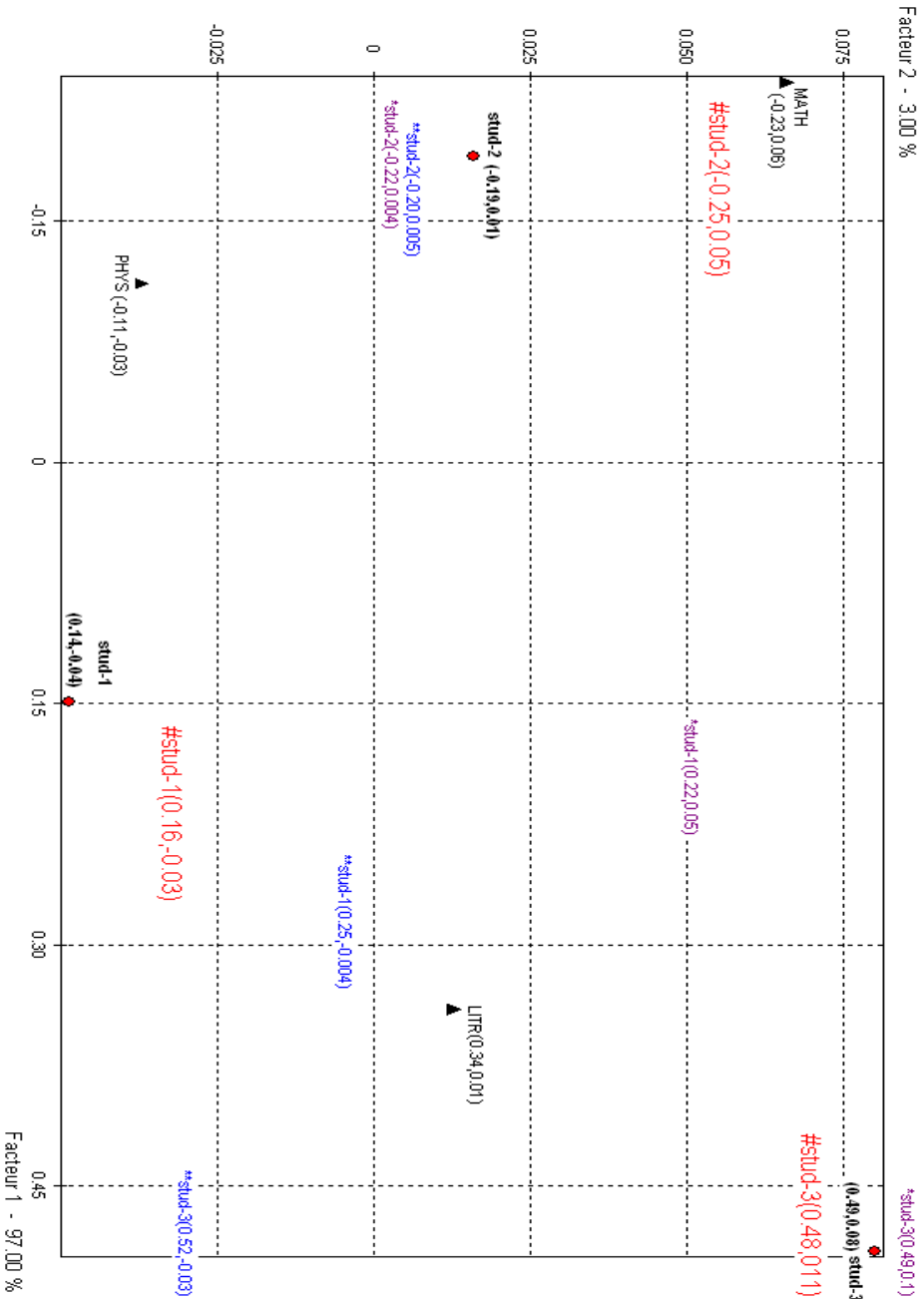
(i) Τα έντονα μαύρα σύμβολα (**•stud-1**, **•stud-2**, **•stud-3**) εκφράζουν τις παραγοντικές συντεταγμένες των προφίλ-γραμμών της κλασικής CA, δηλ. τις κεντρικές τιμές των TFN-στοιχείων των προφίλ-γραμμών του ασαφούς πίνακα X , δηλ. του συνήθους αριθμητικού πίνακα $X_{\ell}^{\alpha=1} = X(\alpha=1, \theta=0) = X(\alpha=1, \theta=1) = X_r^{\alpha=1}$.

(ii) Τα κεφαλαία γράμματα ή σύμβολα (**▲MATH**, **▲PHYS**, **▲LITR**) εκφράζουν τις παραγοντικές συντεταγμένες των προφίλ-στηλών της κλασικής CA.

(iii) Τα μωβ γράμματα ή σύμβολα (***stud-1**, ***stud-2**, ***stud-3**) εκφράζουν τις παραγοντικές συντεταγμένες που παράγονται από τη CAFD μέσω της αποασαφοποιητικής διαδικασίας «Κέντρου Βάρους-1a», βλ. σχέσεις (1), (2).

(iv) Τα μπλε γράμματα ή σύμβολα (****stud-1**, ****stud-2**, ****stud-3**) εκφράζουν τις παραγοντικές συντεταγμένες που παράγονται από τη CAFD μέσω της αποασαφοποιητικής διαδικασίας «Κέντρου Βάρους-1b», βλ. σχέσεις (3), (4).

(v) Τα κόκκινα γράμματα ή σύμβολα (#stud-1, #stud-2, #stud-3) εκφράζουν τις παραγοντικές συντεταγμένες που παράγονται από τη CAFD μέσω της αποασαφοποιητικής διαδικασίας «Αριθμητικού Μέσου-2», βλ. σχέσεις (5), (5a), (5b).



Τέλος, είναι αξιοσημείωτα στην παραπάνω Εικόνα 1, τα εξής:

- a) Αυτή η πρώτη αλλά αντιπροσωπευτική παραγοντική γραφική παράσταση δείχνει ότι η CAFD μπορεί να λειτουργήσει τόσο αυτοτελώς όσο και σε σύγκριση με τη CA.
- b) Εμφανίζεται ήδη μια αρμονική ομοιότητα μεταξύ των αντίστοιχων παραγοντικών συντεταγμένων (βλ. π.χ. την όμοια θέση των: • stud-3, #stud-3, *stud-3).
- c) Όσο πιο ασαφή είναι τα στοιχεία του X, τόσο πιο μεγάλη είναι η απόκλιση μεταξύ των αντίστοιχων κλασικών (της CA) και μη-κλασικών (της CAFD) προβολών-συντεταγμένων (βλ. π.χ. τη θέση των: • stud-1, #stud-1, *stud-1, **stud-1).
- d) Ως προς τη φυσική ερμηνεία του ανωτέρω γραφήματος, παρατηρούμε π.χ. κατά την CAFD (#), ότι: η γεωγρ. περιοχή (#stud-2) συνδέεται ισχυρά με το Μαθηματικό Τμήμα (▲ MATH) δηλ. οι υποψήφιοι από την περιοχή αυτή εκτιμάται ότι θα εισαχθούν κυρίως στο Μαθηματικό Τμήμα, η περιοχή (#stud-3) βρίσκεται μακριά από το Μαθηματικό και πιο κοντά στο Φιλολογικό (▲ LITR), ενώ η περιοχή (#stud-1) φαίνεται διεσπαρμένη (πιο κοντά στα Τμήματα (▲ LITR), (▲ PHYS)), κ.λπ.

ABSTRACT

This paper is the continuation of our previous works [6,7,8] where Correspondence Analysis (CA) for fuzzy data was introduced, and the mathematical foundation for Correspondence Analysis with Fuzzy Data (CAFD) was investigated.

In this work, we complete these theoretical results with the practical approach of implementing CAFD. A real problem is considered and the employed computational processes are explained in detail through a simple numerical but representative example. Finally, on a first simultaneous geometrical display, we are presenting and illustrating the defuzzified-factorial projections of CAFD in comparison with those of standard CA.

ΑΝΑΦΟΡΕΣ

- [1] Benzécri J.P. & al. (1984). *L'analyse des données*, Dunod, Paris.
- [2] Bertier P., Bouroche J.M. (1977). *Analyse des données multidimensionnelles*. PUF, Paris.
- [3] Kaufmann A., Gupta M. (1985). *Introduction to Fuzzy Arithmetic*. Van Nostrand Reinhold Company.
- [4] Moore R. (1979), *Interval Analysis*, Prentice Hall-Englewood Cliffs.
- [5] SPAD: *Système Portable pour l'Analyse des Données* (2001), V.5.0, CISIA-CERESTA, France.
- [6] Θεοδώρου Ι., Δρόσος Κ., Αλεβίζος Φ. (2005). *Ασαφής Ανάλυση Αντιστοιχιών*. ΕΣΙ, Πρακτικά 18^{ου} Πανελληνίου Συνεδρίου Στατιστικής-Ρόδος, σελ. 137-148.
- [7] Theodorou Y., Alevizos Ph. (2006). The Fuzzy Eigenvalue problem of Fuzzy Correspondence Analysis, *Journal of Interdisciplinary Mathematics*, **9**, No.1, pp.115-137.
- [8] Theodorou Y., Drossos C., Alevizos Ph. (2007). Correspondence Analysis with Fuzzy Data: The Fuzzy Eigenvalue Problem. *Fuzzy Sets and Systems*, **158**, pp.704-721.
- [9] Wolkenhauer O. (1998). *Possibility theory with applications to Data Analysis*. Research Studies Press Ltd., England.

ΑΝΑΓΝΩΡΙΣΗ ΣΤΟΧΑΣΤΙΚΩΝ ΣΥΣΤΗΜΑΤΩΝ ΣΤΑΣΙΜΩΝ ΣΗΜΕΙΑΚΩΝ ΑΝΕΛΙΞΕΩΝ

Γ. Καραβασίλης, Α. Ρήγας

Δημοκρίτειο Πανεπιστήμιο Θράκης

Πολυτεχνική Σχολή

Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών

{gkaravas, rigas}@ee.duth.gr

ΠΕΡΙΛΗΨΗ

Στην εργασία αυτή μελετάται ένα στοχαστικό σύστημα στάσιμων σημειακών ανελίξεων με δύο εισόδους και μία έξοδο. Το σύστημα περιγράφεται επαρκώς από τη μέση τιμή των γεγονότων της εξόδου όταν είναι γνωστές οι χρονικές στιγμές εμφάνισης των εισερχομένων γεγονότων. Επιπλέον υποθέτουμε ότι δεν υπάρχει αλληλεπίδραση μεταξύ των γεγονότων των δύο εισόδων. Η μέση τιμή της εξόδου στηρίζεται στις χρουστικές αποκρίσεις του συστήματος, οι οποίες εκφράζουν την επίδραση των δύο εισόδων στην απόκριση του συστήματος. Αποδεικνύεται με τεχνικές φασματικής ανάλυσης, ότι οι εκτιμητές των χρουστικών συναρτήσεων απόκρισης ακολουθούν ασυμπτωτικά Κανονική κατανομή, με σταθερή διακύμανση. Επεξηγηματικά παραδείγματα της μεθόδου παρατίθενται με τη χρήση βιοϊατρικών δεδομένων.

Λέξεις Κλειδιά: Φασματική ανάλυση, κανονική κατανομή, μυϊκή άτρακτος.

1. ΕΙΣΑΓΩΓΗ

Έστω $\mathcal{P}(t) = \{\mathcal{P}_1(t), \mathcal{P}_2(t), \mathcal{P}_3(t)\}$, $-\infty < t < \infty$, είναι μια τρισδιάστατη στοχαστική σημειακή ανελίξη, όπου $\mathcal{P}_a(t)$ ($a = 1, 2, 3$) απαριθμεί το πλήθος των γεγονότων τύπου a , τα οποία παρατηρήθηκαν στο χρονικό διάστημα $(0, t]$. Οποιαδήποτε συνιστώσα του $\mathcal{P}(t)$ αποτελεί μια τυχαία, ακεραίων τιμών και μη-αρνητική τυχαία μεταβλητή, με προσauζήσεις $d\mathcal{P}_a(t) = \mathcal{P}_a(t, t + dt)$, που ορίζονται ως εξής:

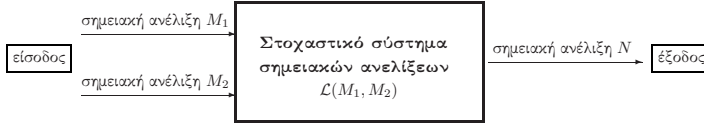
$$d\mathcal{P}_a(t) = \begin{cases} 1, & \text{αν έχουμε γεγονός τύπου } a \text{ στο διάστημα } (t, t + dt] \\ 0, & \text{διαφορετικά} \end{cases}.$$

Επιπλέον, υποθέτουμε ότι οι σημειακές ανελίξεις είναι στάσιμες, δηλαδή τα πιθανοκρατικά χαρακτηριστικά τους δεν επηρεάζονται από το χρόνο t .

Ένα στοχαστικό σύστημα σημειακών ανελίξεων \mathcal{L} , είναι ένας στοχαστικός μηχανισμός, ο οποίος περιγράφει την επίδραση δύο σημειακών ανελίξεων M_1, M_2 (είσοδοι του συστήματος) στην απόκριση του συστήματος, που εκφράζεται μέσω της σημειακής ανελίξης N , δηλαδή

$$N = \mathcal{L}(M_1, M_2).$$

Σχήμα 1: Διάγραμμα ενός στοχαστικού συστήματος στάσιμων σημειακών ανελίξεων



Στην παρούσα εργασία βρίσκουμε τις ασυμπτωτικές κατανομές των εκτιμητών των κρουστικών συναρτήσεων απόκρισης του συστήματος, που επηρεάζουν τη μέση τιμή της απόκρισής του.

2. ΟΡΙΣΜΟΣ ΣΤΑΣΙΜΩΝ ΣΗΜΕΙΑΚΩΝ ΑΝΕΛΙΞΕΩΝ

Στις επόμενες υποπαραγράφους θα ορίσουμε σημαντικές παραμέτρους των σημειακών ανελίξεων στο πεδίο του χρόνου και της συχνότητας.

2.1 Πεδίο του χρόνου

Η συνάρτηση γινομένου (product density) πρώτης τάξης των γεγονότων τύπου ($a = 1, 2, 3$), p_a , παριστάνει τη μέση τιμή των γεγονότων της διεργασίας στη μονάδα του χρόνου:

$$p_a = \lim_{dt \rightarrow 0} \frac{Pr\{d\mathcal{P}_a(t) = 1\}}{dt}.$$

Η δεύτερης τάξης συνάρτηση γινομένου μεταξύ των γεγονότων τύπου a και b ($a, b = 1, 2, 3$), ορίζεται ως εξής:

$$p_{ab}(u) = \lim_{dt, du \rightarrow 0} \frac{Pr\{d\mathcal{P}_a(t+u) = 1 \text{ και } d\mathcal{P}_b(t) = 1\}}{dt du}, \quad u \neq 0.$$

Η σπουδαιότητα των συναρτήσεων γινομένου έγκειται στη σχέση τους με τη μέση τιμή των προσαυξήσεων της στοχαστικής διεργασίας, όπως φαίνεται στους παρακάτω τύπους:

$$\mathbf{E}\{d\mathcal{P}_a(t)\} = Pr\{d\mathcal{P}_a(t) = 1\} \approx p_a dt,$$

$$\mathbf{E}\{d\mathcal{P}_a(t+u)d\mathcal{P}_b(t)\} = Pr\left\{ \begin{array}{l} d\mathcal{P}_a(t+u) = 1 \text{ και} \\ d\mathcal{P}_b(t) = 1 \end{array} \right\} \approx p_{ab}(u) dt du,$$

για πολύ μικρές ποσότητες dt, du . Επιπλέον, μέσω των συναρτήσεων γινομένου και της συνάρτησης Dirac $\delta(u)$, ορίζουμε τις πυκνότητες των συναρτήσεων συσσώρευσης (cumulant densities) $q_{ab}(u)$ και $q_{aa}(u)$, ως εξής:

$$\begin{aligned} \mathbf{cum}\{d\mathcal{P}_a(t+u), d\mathcal{P}_b(t)\} &= q_{ab}(u) dt du = (p_{ab}(u) - p_a p_b) dt du, \\ \mathbf{cum}\{d\mathcal{P}_a(t+u), d\mathcal{P}_a(t)\} &= (q_{aa}(u) + p_a \delta(u)) dt du \\ &= (p_{aa}(u) - p_a^2 + p_a \delta(u)) dt du, \quad \text{όπου } a \neq b. \end{aligned}$$

Με παρόμοιο τρόπο μπορούμε να ορίσουμε τις πυκνότητες των συναρτήσεων συσώρευσης μεγαλύτερης τάξης $q_{a_1 \dots a_\ell}(u_1, \dots, u_{\ell-1})$, όπου $\ell = 2, 3, 4, \dots$ και $a_1, \dots, a_\ell = 1, 2, 3$, για τις οποίες ο αναγνώστης μπορεί να ανατρέξει στον Brillinger (1972).

2.2 Πεδίο της συχνότητας

Για να εξασφαλίσουμε την ύπαρξη των φασματικών πυκνοτήτων (spectral densities) είναι αναγκαία η παρακάτω υπόθεση, η οποία ισχύει για στοχαστικές διεργασίες στις οποίες γεγονότα με μεγάλη χρονική απόσταση θεωρούνται ανεξάρτητα.

Υπόθεση 1. Η στάσιμη σημειακή ανέλιξη $\mathcal{P}(t) = \{\mathcal{P}_1(t), \mathcal{P}_2(t), \mathcal{P}_3(t)\}$, ικανοποιεί τις σχέσεις

$$\int \dots \int (1 + |u_j|) |q_{a_1 \dots a_\ell}(u_1, \dots, u_{\ell-1})| du_1 \dots du_{\ell-1} < \infty,$$

για $\ell \in \{2, 3, 4, \dots\}$, $j \in \{1, 2, \dots, \ell - 1\}$ και $a_1, \dots, a_\ell = 1, 2, 3$.

Έστω ότι η στάσιμη σημειακή ανέλιξη $\mathcal{P}(t) = \{\mathcal{P}_1(t), \mathcal{P}_2(t), \mathcal{P}_3(t)\}$, πληροί την Υπόθεση 1, τότε η διαφασματική συνάρτηση πυκνότητας $f_{ab}(\lambda)$, ορίζεται ως ο μετασχηματισμός Fourier του $q_{ab}(u)$, όπου $a, b = 1, 2, 3$:

$$f_{ab}(\lambda) = (2\pi)^{-1} \int q_{ab}(u) e^{-i\lambda u} du, \quad -\infty < \lambda < \infty.$$

Αντιστοίχως, το φάσμα ισχύος της διεργασίας $\mathcal{P}_a(t)$, δίνεται από τη σχέση:

$$f_{aa}(\lambda) = (2\pi)^{-1} p_a + (2\pi)^{-1} \int q_{aa}(u) e^{-i\lambda u} du, \quad -\infty < \lambda < \infty.$$

Αντιστρέφοντας τις παραπάνω σχέσεις προκύπτουν τα $q_{ab}(u)$ και $q_{aa}(u)$:

$$q_{ab}(u) = \int f_{ab}(\lambda) e^{i\lambda u} d\lambda, \quad -\infty < u < \infty,$$

$$q_{aa}(u) = \int [f_{aa}(\lambda) - (2\pi)^{-1} p_a] e^{i\lambda u} d\lambda, \quad -\infty < u < \infty.$$

3. ΑΝΑΓΝΩΡΙΣΗ ΣΤΟΧΑΣΤΙΚΟΥ ΣΥΣΤΗΜΑΤΟΣ

Έστω $\mathcal{P}(t) = \{\mathcal{P}_1(t), \mathcal{P}_2(t), \mathcal{P}_3(t)\} = \{N(t), M_1(t), M_2(t)\}$, $-\infty < t < \infty$ η διεργασία που περιγράφει το στοχαστικό σύστημα που μελετούμε. Το σύστημα περιγράφεται επαρκώς μέσω της δεσμευμένης μέσης τιμής της απόκρισης του συστήματος:

$$\begin{aligned} \mathbf{E}\{dN(t) | M_1, M_2\} &= Pr \left\{ \text{γεγονός της διεργασίας } N \text{ στο } (t, t + dt) \mid \mathcal{H}_t^{M_1, M_2} \right\} \\ &= \lambda \left(t \mid \mathcal{H}_t^{M_1, M_2} \right) dt + o(dt), \end{aligned}$$

όπου $\mathcal{H}_t^{M_1, M_2}$ εκφράζει τα γεγονότα των διεργασιών εισόδου $\{M_1, M_2\}$ έως τη χρονική στιγμή t . Επομένως, η μέση δεσμευμένη ένταση (conditional mean intensity) $\lambda(t | \mathcal{H}_t^{M_1, M_2})$ δίνεται από τη σχέση

$$\lambda(t | \mathcal{H}_t^{M_1, M_2}) = \lim_{dt \rightarrow 0} \frac{\text{Pr} \left\{ \text{γεγονός της } N \text{ στο } (t, t + dt) | \mathcal{H}_t^{M_1, M_2} \right\}}{dt}.$$

Έστω ότι τα γεγονότα των σημειακών ανελίξεων M_1 και M_2 έχουν παρατηρηθεί στις χρονικές στιγμές u_1, u_2, \dots, u_{k_1} και v_1, v_2, \dots, v_{k_2} αντίστοιχα. Αν υποθέσουμε ότι δεν υπάρχει αλληλεπίδραση μεταξύ των γεγονότων των M_1, M_2 , τότε μια κατάλληλη μορφή για τη $\lambda(t | \mathcal{H}_t)$, θα είναι η:

$$\lambda(t | \mathcal{H}_t) = a_0 + [a_1(t - u_1) + \dots + a_1(t - u_{k_1})] + [a_2(t - v_1) + \dots + a_2(t - v_{k_2})],$$

όπου η σταθερά a_0 αντιστοιχεί στην έξοδο του συστήματος όταν δεν έχουμε καμία είσοδο (δηλαδή $M_1(u) = M_2(v) = 0$, για $-\infty < u, v \leq t$) και $a_1(\cdot), a_2(\cdot)$ στις επιδράσεις των M_1, M_2 στη $\lambda(t | \mathcal{H}_t)$. Οι συναρτήσεις $a_1(\cdot), a_2(\cdot)$ ονομάζονται χρουστικές αποκρίσεις του συστήματος (impulse responses). Τελικά, το ακόλουθο μοντέλο περιγράφει σε πιο συμπαγή μορφή τη γραμμική σχέση μεταξύ των N και M_1, M_2 , χρησιμοποιώντας ολοκληρώματα Riemann-Stieltjes, ως εξής:

$$\mathbf{E}\{dN(t) | M_1, M_2\} \cong \underbrace{\left[a_0 + \int_{-\infty}^t a_1(t-u) dM_1(u) + \int_{-\infty}^t a_2(t-v) dM_2(v) \right]}_{\lambda(t | \mathcal{H}_t)} dt.$$

Εφαρμόζοντας τις ιδιότητες της μέσης τιμής για δεσμευμένες τυχαίες μεταβλητές και το μετασχηματισμό Fourier στην παραπάνω σχέση, προκύπτει ότι οι εξισώσεις που περιγράφουν το σύστημα είναι:

$$p_N = a_0 + p_{M_1} \int_0^\infty a_1(u) du + p_{M_2} \int_0^\infty a_2(v) dv, \quad (1)$$

$$q_{NM_k}(u) = p_{M_k} a_k(u) + \int_0^\infty a_k(u-v) q_{M_k M_k}(v) dv, \quad \text{για } k = 1, 2,$$

$$f_{NM_k}(\lambda) = A_k(\lambda) f_{M_k M_k}(\lambda), \quad \text{όπου } A_k(\lambda) = \int_0^\infty a_k(u) e^{-i\lambda u} du. \quad (2)$$

Με βάση τη (2) η σχέση (1) γίνεται:

$$p_N = a_0 + p_{M_1} A_1(0) + p_{M_2} A_2(0). \quad (3)$$

4. ΣΤΑΤΙΣΤΙΚΗ ΣΥΜΠΕΡΑΣΜΑΤΟΛΟΓΙΑ

Έστω ότι έχουμε καταγράψει τα γεγονότα της διεργασίας $\{N(t), M_1(t), M_2(t)\}$ στο χρονικό διάστημα $(0, T]$. Για την εκτίμηση των a_0 , $a_1(u)$ και $a_2(u)$, ορίζουμε αρχικά τους προφανείς εκτιμητές των συναρτήσεων γινομένου πρώτης τάξης:

$$\hat{p}_N = \frac{\text{πλήθος γεγονότων της } N}{T} = \frac{N(T)}{T}$$

$$\hat{p}_{M_k} = \frac{\text{πλήθος γεγονότων της } M_k}{T} = \frac{M_k(T)}{T}, \text{ όπου } k = 1, 2.$$

Για να ορίσουμε τους εκτιμητές του φάσματος ισχύος $I_{aa}(\lambda)$ και της διαφασματικής πυκνότητας $I_{ab}(\lambda)$, ορίζουμε:

$$d_a(\lambda) = \int_0^T e^{-i\lambda t} [dP_a(t) - p_a dt]$$

$$I_{aa}(\lambda) = \frac{1}{2\pi T} d_a(\lambda) \overline{d_a(\lambda)} = \frac{1}{2\pi T} |d_a(\lambda)|^2$$

$$I_{ab}(\lambda) = \frac{1}{2\pi T} d_a(\lambda) \overline{d_b(\lambda)}.$$

Για να βελτιώσουμε τις ιδιότητες των παραπάνω εκτιμητών εισάγουμε στον υπολογισμό τους ένα παράθυρο εξομάλυνσης:

$$\hat{f}_{aa}(\lambda) = \frac{2\pi}{T} \sum_j W_T(\lambda - z_j) I_{aa}(z_j)$$

$$\hat{f}_{ab}(\lambda) = \frac{2\pi}{T} \sum_j W_T(\lambda - z_j) I_{ab}(z_j),$$

όπου $z_j = \frac{2\pi j}{T}$, $j \in \mathbb{Z}$ και $W_T(\lambda) = b_T^{-1} \widetilde{W}(b_T^{-1} \lambda)$ με \widetilde{W} ένα φασματικό παράθυρο (spectral kernel), με εύρος b_T .

Από τη σχέση (2) προκύπτει

$$\hat{A}_k(\lambda) = \frac{\hat{f}_{NM_k}(\lambda)}{\hat{f}_{M_k M_k}(\lambda)}, \text{ δοθέντος ότι } \hat{f}_{M_k M_k}(\lambda) \neq 0.$$

Οπότε, από τις σχέσεις (3) και (2) εξάγουμε τους τελικούς φασματικούς εκτιμητές:

$$\hat{a}_0 = \hat{p}_N - \hat{p}_{M_1} \hat{A}_1(0) - \hat{p}_{M_2} \hat{A}_2(0)$$

$$\hat{a}_k(u) = \frac{1}{T} \sum_j K_T(\lambda_j) \hat{A}_k(\lambda_j) e^{i\lambda_j u},$$

όπου $\lambda_j = \frac{2\pi j}{T}$, $j \in \mathbb{Z}$ και $K_T(\lambda) = K(h_T \lambda)$, είναι ένας παράγοντας σύγκλισης (convergence factor) με εύρος h_T . Περισσότερα για τα φασματικά παράθυρα και τους παράγοντες σύγκλισης αναφέρονται από τον Brillinger (2001, pp. 52–57).

Μπορούμε τώρα να βρούμε την ασυμπτωτική κατανομή των εκτιμητών $\hat{a}_1(u)$ και $\hat{a}_2(u)$, που είναι το βασικό αποτέλεσμα της εργασίας μας.

Θεώρημα 1. Έστω $\mathcal{P}(t) = \{N(t), M_1(t), M_2(t)\}$ η σημειακή ανέλιξη που περιγράφει το στοχαστικό μας σύστημα σημειακών ανελίξεων, η οποία ικανοποιεί την Υπόθεση 1 και τη σχέση $\int |u| |a_k(u)| du < \infty$, όπου $k = 1, 2$. Υποθέτουμε επίσης ότι W και K είναι δύο παράγοντες σύγκλισης. Τότε, αν $b_T T \rightarrow \infty$, $h_T T \rightarrow \infty$, $\frac{b_T}{h_T} \rightarrow 0$ καθώς $T \rightarrow \infty$, $b_T \rightarrow 0$, $h_T \rightarrow 0$, ισχύει ότι

$$\sqrt{b_T h_T T^2} [\hat{a}_k(u) - a_k(u)] \xrightarrow{D} \text{Normal} \left(0, \frac{p_N}{2\pi p_{M_k}} \int W^2(z) dz \int K^2(\lambda) d\lambda \right).$$

Απόδειξη. Έστω $g : \mathbb{R}^2 \mapsto \mathbb{R}$, μια διαφορίσιμη συνάρτηση με $g(x, y) = \frac{x}{y}$, όπου $y \neq 0$. Αν απαλείψουμε το δείκτη k για λόγους απλότητας και εφαρμόσουμε για την g , με $x = f_{NM}(\lambda)$ και $y = f_{MM}(\lambda)$, την πολυδιάστατη μέθοδο δέλτα (Brockwell and Davis, 1991, p. 211), προκύπτει ότι η ποσότητα $\sqrt{b_T h_T T^2} [g(\hat{x}, \hat{y}) - g(x, y)] = \sqrt{b_T h_T T^2} [\hat{A}(\lambda) - A(\lambda)]$ έχει την ίδια ασυμπτωτική κατανομή με την παράσταση

$$\frac{1}{f_{MM}(\lambda)} \sqrt{b_T h_T T^2} [\hat{f}_{NM}(\lambda) - f_{NM}(\lambda)] - \frac{A(\lambda)}{f_{MM}(\lambda)} \sqrt{b_T h_T T^2} [\hat{f}_{MM}(\lambda) - f_{MM}(\lambda)]. \quad (4)$$

Λαμβάνοντας επιπλέον υπόψη την ασυμπτωτική κατανομή των φασματικών πυκνοτήτων (Brillinger, 1974) μπορούμε να αποδείξουμε το ζητούμενο. Πιο συγκεκριμένα,

$$\begin{aligned} \lim_{T \rightarrow \infty} \mathbf{E} \left\{ \sqrt{b_T h_T T^2} [\hat{a}(u) - a(u)] \right\} = \\ \lim_{T \rightarrow \infty} \frac{1}{T} \sum_j K_T(\lambda_j) \sqrt{b_T h_T T^2} e^{i\lambda_j u} \mathbf{E} \left[\hat{A}(\lambda_j) - A(\lambda_j) \right] = 0. \end{aligned}$$

Για τη συνδιακύμανση των $\hat{a}(u)$ και $\hat{a}(v)$ έχουμε

$$\begin{aligned} b_T h_T T^2 \mathbf{cov}[\hat{a}(u) - a(u), \hat{a}(v) - a(v)] = \frac{1}{T^2} \sum_j \sum_k K_T(\lambda_j) K_T(\lambda_k) \times \\ \left\{ b_T h_T T^2 \mathbf{cov} \left[\hat{A}(\lambda_j) - A(\lambda_j), \hat{A}(\lambda_k) - A(\lambda_k) \right] \right\} e^{i(\lambda_j u - \lambda_k v)}. \quad (5) \end{aligned}$$

Όμως,

$$\begin{aligned} b_T h_T T^2 \mathbf{cov} \left[\hat{A}(\lambda_j) - A(\lambda_j), \hat{A}(\lambda_k) - A(\lambda_k) \right] = \\ \delta\{\lambda_j - \lambda_k\} \left[\frac{f_{NN}(\lambda_j)}{f_{MM}(\lambda_k)} - \frac{f_{NM}(\lambda_j) f_{NM}(-\lambda_k)}{f_{MM}^2(\lambda_k)} \right] 2\pi h_T T \int \widetilde{W}^2(z) dz. \end{aligned}$$

Κατόπιν, θέτουμε $u = v$ στη σχέση (5), προσεγγίσουμε τα αθροίσματα με ολοκληρώματα στα οποία αλλάζουμε τη μεταβλητή ολοκλήρωσης και τέλος χρησιμοποιούμε

τις ιδιότητες $\lim_{|\lambda| \rightarrow \infty} f_{ab}(\lambda) = \begin{cases} \frac{p_a}{2\pi}, & \text{αν } a = b \\ 0, & \text{αν } a \neq b \end{cases}$ και $\int \widetilde{W}^2(z) dz = \frac{1}{2\pi} \int W^2(z) dz$,
για να πάρουμε το αποτέλεσμα

$$\begin{aligned} \lim_{h_T \rightarrow 0} b_T h_T T^2 \mathbf{Var}[\hat{a}(u) - a(u)] &= \frac{p_N}{p_M} \int \widetilde{W}^2(z) dz \int K^2(\lambda) \\ &= \frac{1}{2\pi} \frac{p_N}{p_M} \int W^2(z) dz \int K^2(\lambda) d\lambda. \end{aligned}$$

Για να αποδείξουμε την κανονικότητα θεωρούμε την κανονικοποιημένη συνάρτηση συσσώρευσης R τάξης ($R > 2$),

$$\mathbf{cum} \left\{ \sqrt{b_T h_T T^2} [\hat{a}(u_1) - a(u_1)], \dots, \sqrt{b_T h_T T^2} [\hat{a}(u_R) - a(u_R)] \right\},$$

η οποία σύμφωνα με τη σχέση (4) είναι ένα άθροισμα όρων της μορφής (Brillinger, 1974)

$$\mathbf{cum} \left[\hat{f}_{ab}(\lambda_1), \dots, \hat{f}_{ab}(\lambda_R) \right] = O \left(b_T^{-R+1} T^{-R+1} \right),$$

όπου a, b δεν είναι απαραίτητως διαφορετικά μεταξύ τους. Επομένως,

$$\begin{aligned} \mathbf{cum} \left\{ \sqrt{b_T h_T T^2} [\hat{a}(u_1) - a(u_1)], \dots, \sqrt{b_T h_T T^2} [\hat{a}(u_R) - a(u_R)] \right\} = \\ O \left((b_T T)^{-\frac{R}{2}+1} \left(\frac{h_T}{T} \right)^{\frac{R}{2}} \right), \end{aligned}$$

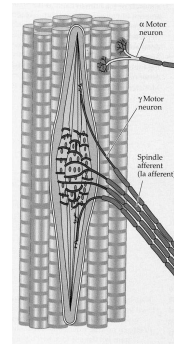
που συγκλίνει στο μηδέν κάτω από τις υποθέσεις του θεωρήματος. □

5. ΕΦΑΡΜΟΓΗ ΣΤΗ ΝΕΥΡΟΦΥΣΙΟΛΟΓΙΑ

Η παραπάνω μέθοδος μπορεί να εφαρμοστεί σε ένα πρόβλημα από το χώρο της νευροφυσιολογίας. Συγκεκριμένα, το νευρομυϊκό σύστημα της μυϊκής ατράκτου μπορεί να προσομοιωθεί από ένα στοχαστικό σύστημα σημειακών ανελιξίων.

Η μυϊκή ατράκτος είναι ένα αισθητήριο όργανο που συναντάται στους μύες όλων των θηλαστικών και ανιχνεύει το μήκος του μυ μέσα στον οποίο βρίσκεται καθώς και το ρυθμό μεταβολής αυτού του μήκους. Είναι ένας αισθητήριος μεταφορέας πληροφοριών στον εγκέφαλο (μέσω του Ia afferent neuron), που διεγείρεται από τα γ και α κινητονεύρα, τα οποία επηρεάζουν τη λειτουργία της μυϊκής ατράκτου, όπως δείχνει το Σχήμα 2. Επομένως,

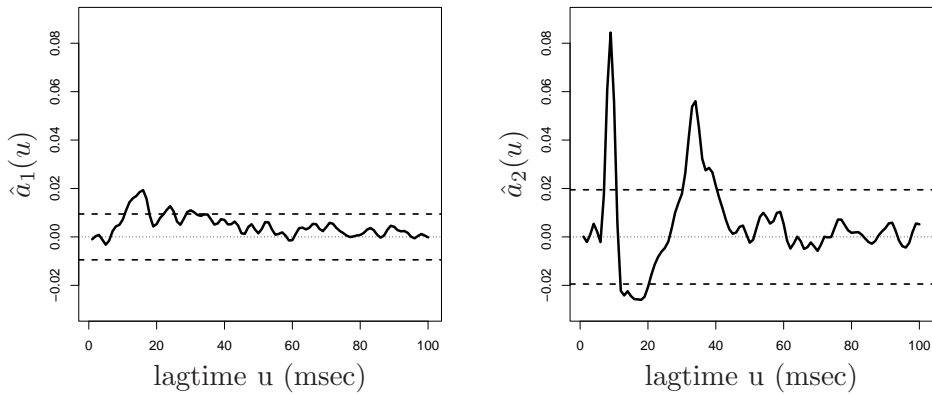
Σχήμα 2: Γραφική αναπαράσταση της μυϊκής ατράκτου



οι νευρικοί παλμοί του Ia afferent neuron μπορούν να θεωρηθούν ως η απόκριση του συστήματος στις επιδράσεις των γ και α κινητονεύρων, που αντιστοιχούν στις δύο εισόδους του συστήματος.

Στο Σχήμα 3 απεικονίζονται οι εκτιμητές των κρουστικών συναρτήσεων απόκρισης, ενός συνόλου δεδομένων που περιλαμβάνουν $N(T) = 358$, $M_1(T) = 691$ και $M_2(T) = 163$ γεγονότα των Ia, γ και α νευρώνων αντίστοιχα, που έχουν καταγραφεί σε χρονικό διάστημα $T = 11360$ ms. Οι οριζόντιες διακεκομμένες γραμμές παριστάνουν τα όρια ενός 95% διαστήματος εμπιστοσύνης για τους εκτιμητές μας, το οποίο κατασκευάστηκε με βάση τα αποτελέσματα του Θεωρήματος 1. Για τους υπολογισμούς χρησιμοποιήθηκε το φασματικό παράθυρο του Fejer με $b_T = 0.005$ και ο παράγοντας σύγκλισης του Bartlett με $h_T = 0.5$ (Brillinger, 2001, p. 55). Η επιλογή των b_T και h_T γίνεται με τέτοιο τρόπο ώστε αφενός να ικανοποιούνται οι συνθήκες του Θεωρήματος 1 και αφετέρου να μειώνεται η διακύμανση των εκτιμητών $\hat{a}_k(u)$.

Σχήμα 3: Οι εκτιμητές των κρουστικών συναρτήσεων απόκρισης του συστήματος της μυϊκής ατράκτου



(α') Υπό την επίδραση του γ κινητονεύρου

(β') Υπό την επίδραση του α κινητονεύρου

ABSTRACT

In the present paper a point process stochastic system is considered with two independent inputs and one output. The system is adequately described by the average number of events of the response which is based on the impulse responses of the incoming point processes. Using spectral analysis techniques it is proved that the estimates of the impulse responses are asymptotically normally distributed with constant variance. An illustrative example is given where the theoretical results are applied on biomedical data.

ΑΝΑΦΟΡΕΣ

- Brillinger, D.R. (1972). The spectral analysis of stationary interval functions. In: Le Cam, L.M., Neyman, J. and Scott E.L. (eds) *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, Berkeley, California, 483–513.
- Brillinger, D.R. (1974). Cross-spectral analysis of processes with stationary increments including the stationary G/G/∞ queue. *Ann Probab*, **2**, 815–827.
- Brillinger, D.R. (2001). *Time Series: Data Analysis and Theory*, Philadelphia: SIAM.
- Brockwell, P.J. and Davis, R.A. (1991). *Time Series: Theory and Methods*, second edition. New York: Springer-Verlag.



ΒΑΣΙΚΑ ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ ΚΑΙ ΙΔΙΑΙΤΕΡΟΤΗΤΕΣ ΤΗΣ ΣΤΑΤΙΣΤΙΚΗΣ ΓΛΩΣΣΑΣ ΠΡΟΓΡΑΜΜΑΤΙΣΜΟΥ R ΚΑΤΑ ΤΟΝ ΥΠΟΛΟΓΙΣΜΟ ΤΩΝ ΣΥΝΑΡΤΗΣΕΩΝ ΠΥΚΝΟΤΗΤΑΣ ΓΙΝΟΜΕΝΟΥ ΣΤΑΣΙΜΩΝ ΣΗΜΕΙΑΚΩΝ ΑΝΕΛΙΞΕΩΝ

Γ. Καραβασίλης¹, Α. Ρήγας¹, Δ. Κύδρος²

¹Δημοκρίτειο Πανεπιστήμιο Θράκης

Πολυτεχνική Σχολή

Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών

{gkaravas, rigas}@ee.duth.gr

²Τεχνολογικό Εκπαιδευτικό Ίδρυμα Σερρών

Σχολή Διοίκησης και Οικονομίας

Τμήμα Λογιστικής

dkydros@teiser.gr

ΠΕΡΙΛΗΨΗ

Στην εργασία αυτή παρουσιάζεται ο κώδικας για τον υπολογισμό εκτιμητών των συναρτήσεων πυκνότητας γινομένου (product densities) δεύτερης και τρίτης τάξης στάσιμων σημειακών ανελίξεων (point processes). Οι συναρτήσεις αυτές αποτελούν θεμελιώδη εργαλεία στη διερεύνηση της αλληλεξάρτησης δύο ή τριών σημειακών ανελίξεων. Ο υπολογισμός των συναρτήσεων αυτών γίνεται πολύ απαιτητικός σε υπολογιστική ισχύ καθώς μεταβαίνουμε από τη δεύτερη στην τρίτη τάξη. Χρησιμοποιώντας τη γλώσσα προγραμματισμού R εφαρμόζονται δύο βασικές μέθοδοι υπολογισμού. Η πρώτη χρησιμοποιεί τις υπάρχουσες εντολές της R, ενώ η δεύτερη χρησιμοποιεί τις δυνατότητες γλωσσών προγραμματισμού όπως η FORTRAN και η C, οι οποίες καλούνται μέσα από το περιβάλλον της R. Με αφορμή αυτές τις διαφορετικές προσεγγίσεις τονίζονται ορισμένες ιδιαιτερότητες της γλώσσας R, οι οποίες είναι αξιοσημείωτες στον προγραμματισμό και στη διδασκαλία της υπολογιστικής στατιστικής.

Λέξεις Κλειδιά: Γλώσσα προγραμματισμού R, σημειακές ανελίξεις, πυκνότητα γινομένου, C, FORTRAN.

1. ΕΙΣΑΓΩΓΗ

Η R είναι πλέον μία ευρέως διαδεδομένη γλώσσα προγραμματισμού με ενσωματωμένο περιβάλλον, εμπλουτισμένο με συναρτήσεις (built-in functions), για στατιστικούς υπολογισμούς και γραφήματα (<http://www.R-project.org>). Η R συνδυάζει δύο επίπεδα προγραμματισμού. Στο πρώτο, δίνεται η δυνατότητα αντικειμενοστραφούς

προγραμματισμού (object-oriented programming) με τις ενσωματωμένες συναρτήσεις και τις νέες που μπορούν να δημιουργηθούν από το χρήστη. Στο δεύτερο επίπεδο οι κρίσιμες αλγοριθμικές διαδικασίες εκτελούνται από γλώσσες προγραμματισμού, όπως η FORTRAN και η C, οι οποίες καλούνται μέσα από το περιβάλλον της R. Αυτά τα δύο επίπεδα μπορούν να συγκεραστούν για να επιτευχθεί η αποτελεσματικότητα και η βέλτιστη απόδοση στην ταχύτητα του κώδικα προγραμματισμού. Οι σχετικές βιβλιογραφικές αναφορές είτε επικεντρώνονται στη στατιστική επεξεργασία δεδομένων (Venables and Ripley 2002 και Dalgaard 2002), η οποία συνάδει περισσότερο με το πρώτο επίπεδο προγραμματισμού, είτε σε πιο προχωρημένες προγραμματιστικές διαδικασίες (Venables and Ripley 2000, Braun and Murdoch 2007 και Chambers 2008), όπως αυτές του δευτέρου επιπέδου.

Το βασικό κίνητρο για την παρούσα εργασία στάθηκε ο υπολογισμός εκτιμητών για τις συναρτήσεις πυκνότητας γινομένου στάσιμων σημειακών ανελιξεων. Οι συναρτήσεις αυτές είναι πολύ σημαντικές διότι εμπλέκονται στην εύρεση της μέσης τιμής τυχαίων μεταβλητών που σχετίζονται με την εμφάνιση γεγονότων στοχαστικών διεργασιών, όπως θα αναλυθεί στην ενότητα 2. Οι υπολογισμοί που απαιτούνται υλοποιήθηκαν με την R, κάνοντας χρήση και των δύο επιπέδων προγραμματισμού (ενότητα 3). Τέλος, στην ενότητα 4, παρατίθενται τα συμπεράσματα της εργασίας που βασίζονται σε μια συγκριτική παρουσίαση αυτών των δύο μεθόδων.

2. ΔΙΔΙΑΣΤΑΤΗ ΣΗΜΕΙΑΚΗ ΑΝΕΛΙΞΗ

Μια σημειακή ανέλιξη περιγράφει τις μεμονωμένες εμφανίσεις ενός φαινομένου σε μια ακολουθία χρονικών στιγμών. Ως παραδείγματα σημειακών ανελιξεων μπορούμε να αναφέρουμε, τις αφίξεις πελατών σε ένα σύστημα εξυπηρέτησης, τη χρονική ακολουθία των σεισμών σε μια περιοχή και τους νευρικούς παλμούς που καταγράφονται σε ένα νευρώνα. Η μαθηματική αναπαράσταση μιας σημειακής ανέλιξης $\mathcal{P}_a(t)$ ως προς το χρόνο t δίνεται μέσω των προσαυξήσεων της $d\mathcal{P}_a(t)$, από τη σχέση

$$d\mathcal{P}_a(t) = \begin{cases} 1, & \text{αν έχουμε γεγονός τύπου } a \text{ στο διάστημα } (t, t + dt] \\ 0, & \text{διαφορετικά} \end{cases} . \quad (1)$$

2.1 Συναρτήσεις πυκνότητας γινομένου

Για τη μελέτη των σημειακών ανελιξεων ορίζουμε τις συναρτήσεις πυκνότητας γινομένου (product densities). Η συνάρτηση πυκνότητας γινομένου δεύτερης τάξης $p_{ab}(u)$ δύο στάσιμων σημειακών ανελιξεων a και b , ορίζεται ως εξής:

$$\begin{aligned} p_{ab}(u) &= \lim_{dt, du \rightarrow 0} \frac{Pr \left\{ \begin{array}{l} \text{γεγονός τύπου } a \text{ στο διάστημα } (t + u, t + u + du] \text{ και} \\ \text{γεγονός τύπου } b \text{ στο διάστημα } (t, t + dt] \end{array} \right\}}{dt du} \\ &= \lim_{dt, du \rightarrow 0} \frac{Pr \{ d\mathcal{P}_a(t + u) = 1 \text{ και } d\mathcal{P}_b(t) = 1 \}}{dt du} . \end{aligned} \quad (2)$$

Η παράμετρος $p_{ab}(u)$ δεν εξαρτάται από το χρόνο t όταν οι σημειακές ανελίξεις είναι στάσιμες. Η παραπάνω σχέση μπορεί να πάρει την προσεγγιστική μορφή, η οποία εκφράζει την πιθανότητα εμφάνισης ενός γεγονότος από κάθε σημειακή ανελίξη a και b με χρονική απόσταση u μεταξύ τους:

$$Pr \left\{ \begin{array}{l} \text{γεγονός τύπου } a \text{ στο διάστημα } (t+u, t+u+du] \text{ και} \\ \text{γεγονός τύπου } b \text{ στο διάστημα } (t, t+dt] \end{array} \right\} \approx p_{ab}(u)du dt, \quad (3)$$

όπου du, dt είναι μικρές θετικές ποσότητες. Επιπλέον, μπορούμε να συνάγουμε ότι

$$\mathbf{E}\{d\mathcal{P}_a(t+u)d\mathcal{P}_b(t)\} = Pr \left\{ \begin{array}{l} d\mathcal{P}_a(t+u) = 1 \text{ και} \\ d\mathcal{P}_b(t) = 1 \end{array} \right\} \approx p_{ab}(u)dt du. \quad (4)$$

Ομοίως, για τη συνάρτηση πυκνότητας γινομένου τρίτης τάξης $p_{abc}(u, v)$ τριών στάσιμων σημειακών ανελίξεων a, b και c ισχύει ότι:

$$\mathbf{E}\{d\mathcal{P}_a(t+u)d\mathcal{P}_b(t+v)d\mathcal{P}_c(t)\} \approx p_{abc}(u, v)dt du dv. \quad (5)$$

Περισσότερα για τις συναρτήσεις αυτές δίνονται από τους Karavasilis and Rigas (2009).

2.2 Εκτιμητική

Ας υποθέσουμε ότι έχουμε τις σημειακές ανελίξεις a και b οι οποίες πραγματοποιούνται στο χρονικό διάστημα $(0, T]$. Έστω s_j οι χρονικές στιγμές εμφάνισης των γεγονότων της σημειακής ανελίξης a και αντίστοιχα t_k της b , τότε ένας εκτιμητής της συνάρτησης πυκνότητας γινομένου δεύτερης τάξης είναι:

$$\hat{p}_{ab}(u) = \frac{\# \{(s_j, t_k) \text{ με } |(s_j - t_k) - u| < h_T, \forall s_j, t_k \in (0, T]\}}{h_T T}, \quad (6)$$

όπου το σύμβολο $\#$ σημαίνει «το πλήθος των» και η παράμετρος h_T καθορίζει το εύρος κάθε κλάσης με κεντρική τιμή u μέσα στην οποία βρίσκεται η διαφορά $s_j - t_k$. Ομοίως, αν r_i, s_j και t_k παριστάνουν τις χρονικές στιγμές εμφάνισης των γεγονότων των σημειακών ανελίξεων a, b και c αντίστοιχα στο χρονικό διάστημα $(0, T]$, τότε ένας εκτιμητής της συνάρτησης πυκνότητας γινομένου τρίτης τάξης δίνεται από τη σχέση:

$$\hat{p}_{abc}(u, v) = \frac{\# \left\{ (r_i, s_j, t_k) \text{ με } \begin{array}{l} |(r_i - t_k) - u| < h_T \\ |(s_j - t_k) - v| < h_T \end{array}, \forall r_i, s_j, t_k \in (0, T] \right\}}{h_T^2 T}. \quad (7)$$

Οι παραπάνω στατιστικές συναρτήσεις είναι ασυμπτωτικά αμερόληπτοι εκτιμητές των συναρτήσεων πυκνότητας γινομένου $p_{ab}(u)$ και $p_{abc}(u, v)$ (Karavasilis et al. 2005, Karavasilis and Rigas 2009).

3. ΠΡΟΓΡΑΜΜΑΤΙΣΜΟΣ ΜΕ ΤΗΝ R

Ο υπολογισμός των εκτιμητών των συναρτήσεων πυκνότητας γινομένου, πραγματοποιήθηκε με τη στατιστική γλώσσα προγραμματισμού της R, με δύο τρόπους:

- εξ ολοκλήρου με τις υπάρχουσες εντολές και συναρτήσεις της R
- με κλήση των γλωσσών προγραμματισμού FORTRAN και C μέσα από την R

Στην πρώτη προσέγγιση, οι εκτιμητές δεύτερης τάξης υπολογίζονται αρχικά με την εντολή `outer`, η οποία υπολογίζει όλες τις δυνατές τιμές των διαφορών για τα στοιχεία των διανυσμάτων (vectors) `a`, `b` και της εντολής `sapply` η οποία εφαρμόζει μια συνάρτηση στα αποτελέσματα της εντολής `outer`. Ο κώδικας, με $h_T = 1$, είναι ο ακόλουθος:

```
pab <- sapply(1:n,  
             function(u) sum(abs( outer(a, b, "-")-u ) < 1 )) / Time
```

Στη δεύτερη προσέγγιση, οι εκτιμητές δεύτερης τάξης υπολογίζονται με την κλήση υποπρογραμμάτων του τύπου `subroutine` για τη FORTRAN και `void` για τη C. Τα υποπρογράμματα αυτά περιέχουν τις κλασσικές ανακυκλώσεις μέσω των εντολών `do` (για τη FORTRAN) και `for` (για τη C), μαζί με την εντολή ελέγχου `if`. Ο κώδικας για τη FORTRAN είναι:

```
subroutine jabu(a, b, nat, nbt, n, jab)  
*DEC$ ATTRIBUTES DLLEXPORT::jabu  
*DEC$ ATTRIBUTES C, REFERENCE, ALIAS:'jabu_'::jabu  
integer n, nat, nbt, a(nat), b(nbt), jab(n)  
do 10 u=1,n  
  do 20 i=1,nbt  
    do 30 j=1,nat  
      if (abs(a(j)-b(i)-u).lt.1) then  
        jab(u)=jab(u)+1  
      end if  
    30 continue  
  20 continue  
10 continue  
return  
end
```

Ο αντίστοιχος κώδικας για τη C είναι:

```
void jabu(int *a, int *b, int *nat, int *nbt, int *n, int *jab)  
{  
  int i, j, u;  
  for(u=0;u<*n;u++)  
    for(i=0;i<*nbt;i++)  
      for(j=0;j<*nat;j++)  
        if(abs(a[j]-b[i]-u) < 1)  
          jab[u]++;  
}
```

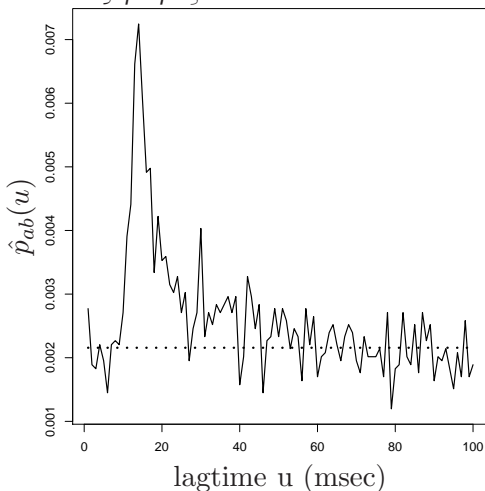
Η διεπαφή (interface) της R με τις γλώσσες αυτές επιτυγχάνεται με τις εντολές `.C` και `.Fortran`, αφού πρώτα μετατρέψουμε τον κώδικα σε αρχείο DLL (Dynamic Link Library) και το «φορτώσουμε» μέσω της εντολής `dyn.load`. Σε κάθε μία από αυτές τις δύο εντολές το πρώτο όρισμα είναι οι χαρακτήρες του ονόματος του υποπρογράμματος. Τα υπόλοιπα ορίσματα αντιστοιχούν στις μεταβλητές με τα απαραίτητα δεδομένα καθώς και τα εξαγόμενα που θα προκύψουν. Η εντολή `as.integer` έχει ως αποτέλεσμα την αποθήκευση του ορίσματος της ως ακέραιο διάνυσμα. Αντίστοιχες μετατροπές μπορούν αν γίνουν και σε άλλους τύπους μεταβλητών, όπως `logical`, `double` και άλλα. Είναι σημαντικό να δώσουμε σε όλα τα ορίσματα την κατάλληλη μέθοδο αποθήκευσης για την R, ώστε να αποφύγουμε λάθος αποτελέσματα ή ανεπιτυχή εκτέλεση του κώδικα. Τα υποπρογράμματα αυτά καλούνται μέσα από περιβάλλον της R με τις εντολές (για τη FORTRAN):

```
NaT <- length(a);  NbT <- length(b)
dyn.load("jab.dll")
  Jabu.fortran <- .Fortran( "jabu",
                           a=as.integer(a), b=as.integer(b),
                           nat=as.integer(NaT), nbt=as.integer(NbT),
                           n=as.integer(n),
                           Jab=as.integer(rep(0, n)) )
  pab <- Jabu.fortran$Jab/Time
```

Για τους εκτιμητές τρίτης τάξης χρησιμοποιείται αρχικά η εντολή `outer` και μετά η εντολή `mapply` που είναι “a multivariate version of `sapply`”, σύμφωνα με τη βοήθεια του λογισμικού.

```
outer.ac <- outer(a, c, "-");  outer.bc <- outer(b, c, "-")
pabc <- mapply( function(u,v) {
  sum ( sapply(1:length(c), function(i) {
    any(abs(outer.ac[,i]-u) < 1) &
```

Σχήμα 1: Η συνάρτηση πυκνότητας γινόμενου μεταξύ των σημειακών ανεξάρτητων a και b , με πλήθος γεγονότων 538 και 1010 αντίστοιχα, σε ένα χρονικό διάστημα $T = 15,872$ msec. Το συμπέρασμα που προκύπτει είναι ότι η b συσχετίζεται θετικά με την a κυρίως όταν η χρονική απόσταση u μεταξύ των γεγονότων τους είναι περίπου 10 – 30 msec. Η οριζόντια διακεκομμένη γραμμή παριστάνει την τιμή του εκτιμητή όταν οι διεργασίες a και b είναι ανεξάρτητες.




```

    any(abs(outer.bc[,i]-v) < 1) } ) ) },
    rep(1:n, each=n), rep(1:n, times=n) )
pabc <- matrix(pabc, nrow=n, ncol=n, byrow=TRUE) / Time

```

Εναλλακτικά, μπορούμε να επεκτείνουμε τα υποπρογράμματα των FORTRAN και C που χρησιμοποιήθηκαν στους εκτιμητές δεύτερης τάξης. Ειδικότερα, ο κώδικας της FORTRAN που χρησιμοποιήθηκε είναι γραμμένος σε μια πιο σύγχρονη μορφή συμβατή με τη FORTRAN 95. Αυτό όμως μπορεί σε ορισμένες περιπτώσεις να προκαλέσει προβλήματα συμβατότητας διότι όπως αναφέρει το εγχειρίδιο Writing R Extensions (2011, p. 86) η εντολή .Fortran είναι “primarily intended for compiled FORTRAN 77 code”.

```

subroutine jabcuv(a, b, c, nat, nbt, nct, n, jabc)
!DEC$ ATTRIBUTES DLLEXPORT::jabcuv
  integer :: n, nat, nbt, nct
  integer, dimension(nat) :: a
  integer, dimension(nbt) :: b
  integer, dimension(nct) :: c
  integer, dimension(n, n) :: jabc
  do u=1, n
  do v=1, n
    do r=1, nat
    do j=1, nct
      if (abs(a(r)-c(j)-u) < 1) then
        do s=1, nbt
          if (abs(b(s)-c(j)-v) < 1) then
            jabc(u,v) = jabc(u,v)+1
          end if
        end do
      end if
    end do
  end do
  end do
end do
end do
return
end

```

ο οποίος καλείται μέσα από την R με τις εντολές:

```

NaT <- length(a); NbT <- length(b); NcT <- length(c)
dyn.load("jabc.dll")
Jabcuv.fortran <- .Fortran("jabcuv",
  a=as.integer(a), b=as.integer(b), c=as.integer(c),
  nat=as.integer(NaT), nbt=as.integer(NbT),
  nct=as.integer(NcT), n=as.integer(n),
  Jabc=matrix(data=as.integer(0), nrow=n, ncol=n))

```

```
pabc <- Jabcuv.fortran$Jabc/Time
```

4. ΣΥΜΠΕΡΑΣΜΑΤΑ

Με βάση την προηγούμενη παράγραφο μπορούμε να συγκρίνουμε τις δύο προσεγγίσεις. Ακολουθώντας το σκεπτικό που έχει αναπτυχθεί από τον Chambers, (2008, pp. 412–413) τα μειονεκτήματα της εισαγωγής κώδικα των γλωσσών C και FORTRAN σε μια συνάρτηση της R, είναι:

- Απαιτείται πολύ περισσότερη δουλειά. Είναι χαρακτηριστική η περίπτωση των εκτιμητών δεύτερης τάξης, οι οποίοι υπολογίζονται εξ ολοκλήρου μέσω των εντολών της R μόνο σε δύο γραμμές, σε αντίθεση με τη δεύτερη προσέγγιση που χρειάζεται ένα σαφώς εκτενέστερο υποπρόγραμμα. Σε αυτό συντελούν οι εντολές `sapply()` και `mapply()` της R, οι οποίες δέχονται ως ορίσματα αντικείμενα δεδομένων, ενώ αποφεύγεται η συνήθης πρακτική των μεταγλωττισμένων γλωσσών προγραμματισμού (compiled languages) όπως η FORTRAN και η C, που είναι η χρησιμοποίηση εντολών ανακύκλωσης όπως η `do` και η `for`, για να εκτελεστούν επαναληπτικές διαδικασίες.
- Τα προγραμματιστικά σφάλματα (software bugs) είναι σαφώς περισσότερα και πιο δυσεπίλυτα. Ένα τέτοιο σφάλμα μπορεί να οφείλεται όχι μόνο σε κάποια αβλεψία που αφορά την αλγοριθμική διαδικασία αλλά και σε ένα πρόβλημα αναντιστοιχίας στον τρόπο αποθήκευσης των διανυσμάτων της R και τον τύπο των ορισμάτων στα υποπρογράμματα των C και FORTRAN. Η κατάλληλη αντιστοίχιση περιγράφεται στο σχετικό πίνακα του εγχειριδίου της R, *Writing R Extensions* (2011, p. 86). Επίσης, η χρήση των αρχείων DLL, δυσκολεύει την εξουδετέρωση τέτοιων σφαλμάτων, διότι η επίλυσή τους απαιτεί τη μετάβαση στον αρχικό κώδικα, μετατροπή εκ νέου σε DLL, εκτέλεση της συνάρτησης της R και ούτω καθεξής. Οι δυσκολίες αυτές περιορίζονται με τη χρήση ορισμένων πακέτων που έχουν στόχο τη βελτίωση της διεπαφής μεταξύ της R και άλλων γλωσσών. Ενδεικτικά αναφέρουμε τα πακέτα RFortran (<http://www.rfortran.org>) και Rcpp (<http://dirk.eddelbuettel.com/code/rcpp.html>).

Τα αντίστοιχα πλεονεκτήματα συνοψίζονται στα εξής:

- Αύξηση της ταχύτητας εκτέλεσης. Μια επαναληπτική διαδικασία που εκτελείται μέσα στην R μπορεί να επιφέρει σημαντικές καθυστερήσεις στην εκτέλεση του προγράμματος. Στην περίπτωσή μας, η δεύτερη προσέγγιση εκτελείται πολύ πιο γρήγορα σε σχέση με την πρώτη. Οι χρόνοι εκτέλεσης (σε sec) των δύο μεθόδων για διαφορετικό αριθμό στοιχείων, σε υπολογιστή με επεξεργαστή Intel Core 2, 2.13 GHz και 3 MB RAM, φαίνονται στον παρακάτω πίνακα:

T	$\mathcal{P}_a(T)$	$\mathcal{P}_b(T)$	$\mathcal{P}_c(T)$	FORTRAN	C	R
1,000	31	45		0.03	0.02	0.05
7,000	238	440		0.08	0.10	1.61
15,872	538	1,010		0.40	0.39	9.80
1,000	32	57	14	0.11		1.11
7,000	228	430	105	2.25		16.66
11,361	358	691	163	5.42		40.39

- Διαχείριση της μνήμης. Όταν αντικείμενα δεδομένων τροποποιούνται στη διάρκεια επαναληπτικών διαδικασιών στην R, τότε αποθηκεύονται στη μνήμη διάφορα αντίγραφα αυτών των αντικειμένων, με αποτέλεσμα να μειώνεται σημαντικά η ταχύτητα εκτέλεσης ή να έχουμε ακόμα και ανώμαλη διακοπή της διαδικασίας, όταν οι διαστάσεις των αντικειμένων των δεδομένων είναι πολύ μεγάλες.
- Η δεύτερη προσέγγιση επιτρέπει τη χρησιμοποίηση πολλών προγραμμάτων που είναι ήδη γραμμένα σε κάποια από αυτές τις δύο γλώσσες χωρίς σημαντικές μετατροπές, μέσα στο περιβάλλον της R.

Οι παραπάνω επισημάνσεις μπορούν να βοηθήσουν στον υπολογισμό αντίστοιχων εκτιμητών όταν απαιτούνται επαναληπτικές διαδικασίες και έχουμε μεγάλα σύνολα δεδομένων, αλλά και στη διδασκαλία φοιτητών που είναι εξοικειωμένοι με τις μεταγλωττισμένες γλώσσες προγραμματισμού και έρχονται σε επαφή με την R.

ABSTRACT

In the present paper computational procedures for the estimation of the product densities of stationary point processes are provided. The computations are implemented both with built-in functions of R and incorporating additional FORTRAN and C code. The concluding remarks contain the advantages and disadvantages between these two approaches.

ΑΝΑΦΟΡΕΣ

- Braun, W.J. and Murdoch, D.J. (2007). *A First Course in Statistical Programming with R*. New York: Cambridge University Press.
- Chambers, J. (2008). *Software for Data Analysis. Programming with R*. New York: Springer.
- Dalgaard, P. (2002). *Introductory Statistics with R*. New York: Springer.
- Karavasilis, G.J., Kotti, V.K., Tsitsis, D.S., Vassiliadis, V.G. and Rigas, A.G. (2005). Statistical methods and software for risk assessment: applications to a neurophysiological data set. *Computational Statistics and Data Analysis*, **49**, 243–263.
- Karavasilis, G.J. and Rigas, A.G. (2009). The use of nonparametric methods of stationary point processes in the study of complex interactions in the neuromuscular system. *Journal of Biological Systems*, **17**(4), 577–595.

R Development Core Team (2011). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org/>.

R Development Core Team (2011). *Writing R Extensions*.

Venables, W.N. and Ripley, B.D. (2000). *S Programming*. New York: Springer.

Venables, W.N. and Ripley, B.D. (2002). *Modern Applied Statistics with S*. New York: Springer.



ΠΡΟΒΛΕΠΤΙΚΑ ΜΟΝΤΕΛΑ ΧΡΟΝΟΣΕΙΡΩΝ ΜΕ ΕΞΑΡΤΗΣΗ ΜΑΚΡΑΣ ΔΙΑΡΚΕΙΑΣ: ΕΦΑΡΜΟΓΗ ΣΤΟΝ ΓΕΝΙΚΟ ΔΕΙΚΤΗ ΤΟΥ ΧΡΗΜΑΤΙΣΤΗΡΙΟΥ ΑΞΙΩΝ ΑΘΗΝΩΝ

Χ. Κάτρης¹, Σ. Δασκαλάκη¹

¹Γενικό Τμήμα, Πανεπιστήμιο Πατρών
{chriskatris, sdask}@upatras.gr

ΠΕΡΙΛΗΨΗ

Σε αυτήν την εργασία μελετάμε την ύπαρξη μακροχρόνιας μνήμης (long-memory) στο Γενικό Δείκτη του Χρηματιστηρίου Αξιών Αθηνών με απώτερο στόχο τη δημιουργία κατάλληλου μοντέλου για την πρόβλεψη του. Θεωρούμε τις ημερήσιες αποδόσεις μιας πενταετίας από το 2004 μέχρι το 2009 και υπολογίζουμε την παράμετρο Hurst με την μέθοδο ανακλιμακούμενου εύρους (Rescaled Range) R/S . Επιπλέον με το τροποποιημένο στατιστικό ανακλιμακούμενου εύρους (modified R/S statistic) ελέγχουμε για ύπαρξη στατιστικά σημαντικής μακροχρόνιας μνήμης. Στη συνέχεια, για την πρόβλεψη μελλοντικών τιμών της ημερήσιας απόδοσης προσαρμόζουμε απ' ενός το βέλτιστο ARIMA μοντέλο με τη μεθοδολογία Box & Jenkins, απ' ετέρου ένα κατάλληλο FARIMA μοντέλο που σύμφωνα με τη βιβλιογραφία προτείνονται για τη μοντελοποίηση χρονοσειρών με μακροχρόνια μνήμη. Τέλος, εφαρμόζουμε μια εναλλακτική προσέγγιση που διαχωρίζει τις τιμές της χρονοσειράς σε τάση και κύκλους με το φίλτρο Hodrick-Prescott, εξειδικεύει τα κατάλληλα μοντέλα για προσέγγιση της τάσης και των κύκλων ξεχωριστά και στη συνέχεια δημιουργεί μια ενιαία πρόβλεψη για τη σειρά. Η σύγκριση της προβλεπτικής ικανότητας των μοντέλων γίνεται με κριτήρια την τετραγωνική ρίζα του μέσου τετραγωνικού σφάλματος (Root Mean Squared Error (RMSE)) και του μέσου απόλυτου σφάλματος (Mean Absolute Error (MAE)) και με δεδομένα ελέγχου τις ημερήσιες αποδόσεις του δείκτη από το 2009 έως σήμερα.

Keywords: Μακροπρόθεσμη μνήμη (long-memory), Εξάρτηση μακράς διάρκειας (Long Range Dependence (LRD)), παράμετρος Hurst, Rescaled Range (R/S), modified R/S statistic, ARIMA, FARIMA, Γενικός δείκτης ΧΑΑ, φίλτρο Hodrick-Prescott.

1. ΕΙΣΑΓΩΓΗ

Η εξάρτηση μακράς διάρκειας (Long Range Dependence) ή μακροπρόθεσμη μνήμη (long-memory) σε χρονοσειρές, μπορεί να οριστεί ως η αυτοσυσχέτιση σε μεγάλες χρονικές υστερήσεις (ακόμα και σε εκατοντάδες χρονικές περιόδους). Στη περίπτωση αυτή μακρινές παρελθοντικές παρατηρήσεις μπορούν να χρησιμοποιηθούν για την πρόβλεψη μελλοντικών τιμών της χρονοσειράς. Η παρουσία μακροπρόθεσμης μνήμης

στην απόδοση χρηματιστηριακών αγορών είναι ένα θέμα το οποίο έχει απασχολήσει πολλούς ερευνητές. Σε κάποιες μελέτες, όπως στην (Hiemstra and Jones, 1997) δεν υπήρξε σημαντική παρουσία μακροπρόθεσμης μνήμης. Αντίθετα, σε άλλες όπως στην (Cheung and Lai, 1995) και στην (Tolvi, 2003) παρατηρήθηκε σε μικρότερο ή μεγαλύτερο βαθμό στις αποδόσεις συγκεκριμένων μετοχών.

Στην εργασία αυτή μελετάται η ύπαρξη μακροπρόθεσμης μνήμης στις ημερήσιες αποδόσεις του Γενικού Δείκτη τιμών του Χρηματιστηρίου Αξιών Αθηνών. Σε προηγούμενη μελέτη (Panas, 2001) μελετήθηκαν οι καθημερινές αποδόσεις Ελληνικών μετοχών και βρέθηκε στατιστικά σημαντική μακροπρόθεσμη μνήμη στην πλειοψηφία των μετοχών. Αντίστοιχα, στην (Barkoulas et al. 2000) μελετήθηκαν οι εβδομαδιαίες αποδόσεις του δείκτη τιμών μετοχών κατά την διάρκεια της δεκαετίας του '80 όπου και πάλι παρατηρήθηκε στατιστικά σημαντική μακροπρόθεσμη μνήμη.

Για τη μελέτη μας χρησιμοποιήσαμε δεδομένα της περιόδου 2004-2011. Προσαρμόσαμε ένα μοντέλο *ARIMA* με τη μεθοδολογία Box-Jenkins καθώς και ένα μοντέλο *FARIMA*(p, d, q), που είναι καταλληλότερο για μοντελοποίηση χρονοσειρών με μακροπρόθεσμη μνήμη. Στη συνέχεια παρουσιάζουμε μια υπολογιστική διαδικασία διάσπασης της σειράς με χρήση του φίλτρου Hodrick-Prescott και τέλος συγκρίνουμε τα μοντέλα ως προς την πρόβλεψη τιμών του γενικού δείκτη.

2. ΕΚΤΙΜΗΣΗ ΤΗΣ ΠΑΡΑΜΕΤΡΟΥ HURST

Η παράμετρος *Hurst* (H), η οποία προτάθηκε από τον (*Hurst*, 1951) στη συνέχεια χρησιμοποιήθηκε σε πολλά ερευνητικά πεδία. Χρησιμοποιείται σαν ένα μέτρο μακροπρόθεσμης μνήμης στις χρονολογικές σειρές.

Τιμές της παραμέτρου $0 < H < 0.5$ δηλώνουν μια χρονοσειρά χωρίς εξάρτηση μακράς διάρκειας (με χαρακτηριστικό την αλλαγή κατεύθυνσης της σειράς συχνότερα από μια τυχαία διαδικασία - αντιεμμονή). Όσο πιο κοντά στο 0 η παράμετρος, τόσο ισχυρότερο το χαρακτηριστικό. Τιμές της παραμέτρου $0.5 < H < 1$ δηλώνουν μια χρονολογική σειρά με εξάρτηση μακράς διάρκειας. Όσο πιο κοντά στο 1 η παράμετρος, τόσο μεγαλύτερος ο βαθμός εξάρτησης μακράς διάρκειας. Μια τιμή $H = 0.5$ δηλώνει έναν πραγματικό τυχαίο περίπατο.

Η πρώτη μέθοδος εκτίμησης της παραμέτρου πάντως αναπτύχθηκε από τον ίδιο τον *Hurst* αλλά χρησιμοποιήθηκε ευρύτατα και από τον *Mandelbrot* (1972) και τους συνεργάτες του. Αυτή την προσέγγιση θα χρησιμοποιήσουμε κι εδώ για την εκτίμηση της παραμέτρου H από τα δεδομένα μας.

2.1 Μέθοδος Ανακλιμακούμενου Εύρους (R/S Statistic)

Το βασικό χαρακτηριστικό που κάνει ιδιαίτερα προτιμητέα την *R/S* ανάλυση σαν μέθοδο εκτίμησης της H είναι η ανθεκτικότητά της στις μεταβολές των περιθωρίων κατανομών όταν αυτές είναι ασύμμετρες ή long-tailed. Αντίθετα, η μέθοδος είναι λιγότερο αποτελεσματική όταν οι περιθωρίες κατανομές είναι κοντά στην κανονική.

Το *R/S* στατιστικό δίνεται από τον τύπο:

$$(R/S) = \frac{1}{s(n)} \left[\text{Max}_{1 \leq k \leq n} \sum_{j=1}^k (X_j - \bar{X}_n) - \text{Min}_{1 \leq k \leq n} \sum_{j=1}^k (X_j - \bar{X}_n) \right] \quad (1)$$

όπου

$$s(n) = \left[\frac{1}{n} \sum_j (X_j - \bar{X}_n)^2 \right]^{1/2} \quad (2)$$

είναι ο συνήθης εκτιμητής (μέγιστης πιθανοφάνειας) για την τυπική απόκλιση, το k είναι ο αριθμός των στοιχείων κάθε συνόλου για το οποίο υπολογίζουμε το R/S , ενώ \bar{X}_n είναι η μέση τιμή του συνόλου των δεδομένων. Η κλίση του γραφήματος της $\log(R(n)/s(n))$ με το $\log(n)$ είναι η παράμετρος *Hurst*.

Οι εργασίες (Mandelbrot and Wallis, 1969· Mandelbrot, 1972· Mandelbrot and Taqqu, 1979) έδειξαν την υπεροχή της μεθόδου αυτής έναντι άλλων πιο βολικών μεθόδων στην εκτίμηση Εξάρτησης Μακράς Διάρκειας, όπως για παράδειγμα την ανάλυση αυτοσυσχετίσεων και τις φασματικές διασπάσεις.

3. ΕΛΕΓΧΟΣ ΕΞΑΡΤΗΣΗΣ ΜΑΚΡΑΣ ΔΙΑΡΚΕΙΑΣ ΜΕ ΤΟ ΤΡΟΠΟΠΟΙΗΜΕΝΟ R/S ΣΤΑΤΙΣΤΙΚΟ

Ο έλεγχος υπόθεσης για ύπαρξη ή μη εξάρτησης μακράς διάρκειας με το R/S στατιστικό ολοκληρώνει τη διαδικασία που στο σύνολό της ονομάζεται R/S Analysis.

Ελέγχουμε H_0 : Τα δεδομένα δεν παρουσιάζουν Εξάρτηση Μακράς Διάρκειας

Το σημαντικότερο μειονέκτημα του R/S στατιστικού είναι η ευαισθησία του ως προς την βραχυπρόθεσμη μνήμη (η σύνδεση μεταξύ τιμών διαφορετικών χρονικών περιόδων, μειώνεται ραγδαία όσο η διαφορά μεταξύ των χρονικών περιόδων αυξάνεται).

Σύμφωνα με τον (Lo, 1991) για να διακρίνουμε μεταξύ βραχυπρόθεσμης μνήμης και μακροπρόθεσμης μνήμης το R/S στατιστικό πρέπει να τροποποιηθεί έτσι ώστε η στατιστική του συμπεριφορά να είναι αμετάβλητη για διαδικασίες βραχυπρόθεσμης μνήμης (η επίδρασή της ενσωματώνεται στον παρανομαστή), αλλά να αποκλίνει για διαδικασίες μακροπρόθεσμης μνήμης.

Αυτό επιτυγχάνεται με το τροποποιημένο R/S στατιστικό:

$$(R/S)_{mod} = \frac{1}{\hat{\sigma}_n(q)} \left[\text{Max}_{1 \leq k \leq n} \sum_{j=1}^k (X_j - \bar{X}_n) - \text{Min}_{1 \leq k \leq n} \sum_{j=1}^k (X_j - \bar{X}_n) \right] \quad (3)$$

όπου

$$\begin{aligned}\hat{\sigma}_n^2(q) &= \frac{1}{n} \sum_{j=1}^n (X_j - \bar{X}_n)^2 + \frac{2}{n} \sum_{j=1}^q \omega_j(q) \left\{ \sum_{i=j+1}^n (X_i - \bar{X}_n)(X_{i-j} - \bar{X}_n) \right\} = \\ &= \hat{\sigma}_x^2 + 2 \sum_{j=1}^q \omega_j(q) \hat{\gamma}_j\end{aligned}\quad (4)$$

Στην εξίσωση (4) ισχύει ότι $\omega_j(q) = 1 - \frac{j}{q+1}$, $q < n$. Επίσης, $\hat{\sigma}_x^2$ και $\hat{\gamma}_j$ είναι οι δειγματικοί εκτιμητές της διακύμανσης και της αυτοσυσχέτισης αντίστοιχα, ενώ το q είναι ο αριθμός των υστερήσεων (lags).

Όταν ισχύει η μηδενική υπόθεση η ποσότητα $(V_n)_{mod} = \frac{1}{\sqrt{n}} \left(\frac{R}{S} \right)_{mod}$ ακολουθεί την κατανομή F_v , της οποίας οι κρίσιμες τιμές ανά επίπεδο σημαντικότητας δίνονται στην εργασία (Lo, 1991). Για το 95% επίπεδο σημαντικότητας δεν απορρίπτουμε τη μηδενική υπόθεση όταν το $(V_n)_{mod}$ βρίσκεται στο διάστημα $[0.809, 1.862]$. Αντίστοιχα, για το 99% επίπεδο σημαντικότητας το διάστημα μη απόρριψης της μηδενικής υπόθεσης γίνεται $[0.721, 2.098]$.

4. ΠΡΟΒΛΕΠΤΙΚΑ ΜΟΝΤΕΛΑ ΧΡΟΝΟΣΕΙΡΩΝ

Ακολουθεί μια σύντομη περιγραφή ορισμένων προβλεπτικών μοντέλων χρονοσειρών τα οποία χρησιμοποιούνται στη συνέχεια της εργασίας. Τα μοντέλα *ARIMA* θεωρούνται κατάλληλα για την μοντελοποίηση διαδικασιών με βραχυπρόθεσμη μνήμη, τα *FARIMA* θεωρούνται κατάλληλα για την μοντελοποίηση διαδικασιών με μακροπρόθεσμη μνήμη, ενώ αναπτύσσεται με βάση τα μοντέλα αυτά μια υπολογιστική διαδικασία που βασίζεται στη διάσπαση της σειράς σε τάση και κύκλους.

4.1 Προβλεπτικά μοντέλα ARIMA

Η μοντελοποίηση χρονοσειρών με στοχαστικά μοντέλα *ARMA* ή *ARIMA* (Autoregressive Integrated Moving Average) μπορεί να πραγματοποιηθεί με την κλασική προσέγγιση των (Box-Jenkins, 1976).

4.2 Προβλεπτικά μοντέλα FARIMA

Τα Autoregressive Fractionally Integrated Moving Average (FARIMA) μοντέλα αναπτύχθηκαν από τους (Mandelbrot, 1977), (Hosking, 1981), (Granger, C. W. J. and R. Joyeux, 1980). Στην ουσία αποτελούν μια επέκταση των μοντέλων *ARIMA*, άν η παράμετρος d μπορεί να πάρει πραγματικές (όχι μόνο ακέραιες) τιμές.

Μια χρονολογική σειρά X_t ακολουθεί μια *FARIMA* διαδικασία, εάν:

$$\Phi_p(B)(1-B)^d X_t = \Theta_q(B)\epsilon_t, \text{ όπου}$$

$$\Phi_p(B) = 1 - \phi_1 B - \dots - \phi_p B^p$$

$$\Theta_q(B) = 1 + \theta_1 B + \dots + \theta_q B^q$$

$$(1-B)^d = 1 - dB - \frac{d}{2}(1-d)B^2 - \dots$$

και τα ϵ_t είναι ανεξάρτητες τυχαίες μεταβλητές με μέση τιμή μηδέν και πεπερασμένη διακύμανση ($\sigma_\epsilon^2 < \infty$).

Όταν $-0.5 < d < 0.5$ η χρονοσειρά X_t είναι στάσιμη και αντιστρέψιμη. Ειδικότερα, αν $0 < d < 0.5$ τότε η χρονοσειρά είναι στάσιμη και παρουσιάζει μακροπρόθεσμη μνήμη, ενώ όταν $-0.5 < d < 0$ τότε η χρονοσειρά έχει ενδιάμεση μνήμη (intermediate memory) και παρουσιάζει αντιστοιχία, δηλαδή αλλάζει κατεύθυνση συχνότερα από τον τυχαίο περίπατο. Τέλος, αν $d \geq 0.5$ η διαδικασία είναι μη-στάσιμη, ενώ όταν $d = 0$ τότε η διαδικασία είναι μια *ARMA* διαδικασία με βραχυπρόθεσμη μνήμη.

4.3 Μέθοδος πρόβλεψης με διάσπαση της σειράς με το φίλτρο Hodrick-Prescott

Η ποσοτική οικονομική ανάλυση υποστηρίζει ότι οι αποδόσεις στο χρηματιστήριο είναι ένας τυχαίος περίπατος, ενώ από την άλλη η τεχνική ανάλυση υποστηρίζει ότι υπάρχουν κύκλοι (Συριόπουλος - Σιρλαντζής (1993)). Για τον έλεγχο αυτών των απόψεων αναπτύσσουμε μια μέθοδο με διάσπαση της σειράς σε τάση και κύκλους.

Σκοπός της μεθόδου είναι να αφαιρεθούν από τη σειρά οι δευτερογενείς κινήσεις (κύκλοι), καθώς μεταβάλλονται πιο γρήγορα από την τάση, δεν παρουσιάζουν τα ίδια χαρακτηριστικά, και μοντελοποιούνται διαφορετικά. Η διάσπαση πραγματοποιείται με το ευρέως χρησιμοποιούμενο σε οικονομικές σειρές φίλτρο (Hodrick-Prescott, 1997). Η μέθοδος ακολουθεί τα κάτωθι βήματα:

- **Βήμα 1:** Χρησιμοποιώντας το φίλτρο Hodrick-Prescott χωρίζουμε τη σειρά σε τάση και κύκλους επιλέγοντας μια τιμή για την παράμετρο λ .
- **Βήμα 2:** Εκτιμάμε την παράμετρο *Hurst* για τη χρονοσειρά της τάσης και προσαρμόζουμε το κατάλληλο *FARIMA* μοντέλο, με τον αλγόριθμο των (Hyndman-Khandakar, 2008), δηλαδή επιλέγεται ο αριθμός των παραμέτρων p και q του μοντέλου και εκτιμώνται οι παράμετροι του μοντέλου (p, d, q) μέσω του αλγόριθμου των (Haslett-Raftery, 1989).
- **Βήμα 3:** Επαναλαμβάνουμε τη διαδικασία για τους κύκλους της σειράς.
- **Βήμα 4:** Ολοκληρώνουμε με πρόσθεση των επιμέρους προβλέψεων.

Σημείωση: Σε περίπτωση μακροπρόθεσμης μνήμης ($0.5 < H < 1$) επιλέγεται ένα *FARIMA* μοντέλο με $0 < d < 0.5$, ενώ σε περίπτωση αντιστοιχίας επιλέγεται ένα *FARIMA* με $-0.5 < d < 0$.

5. ΜΕΛΕΤΗ ΠΕΡΙΠΤΩΣΗΣ ΣΤΟΝ ΓΕΝΙΚΟ ΔΕΙΚΤΗ ΧΑΑ

Τα δεδομένα μας για την εκτίμηση των μοντέλων είναι οι λογαριθμικές αποδόσεις $Z_i = \ln(P_i/P_{i-1})$, όπου P_i είναι η τιμή του ΓΔ του ΧΑΑ την ημέρα i , από τον Σεπτέμβριο του 2004 μέχρι τον Μάιο του 2009. Για έλεγχο των προβλέψεων χρησιμοποιούνται δεδομένα από τον Ιούνιο 2009 μέχρι Μάρτιο 2011 και η ανάλυση γίνεται με χρήση των λογισμικών *R* και *STATA*. Οι ημερήσιες αποδόσεις που χρησιμοποιήθηκαν στην εργασία, προέρχονται από την ιστοσελίδα www.stockwatch.com.cy.

Η μέση τιμή είναι -0.0000342 (δεν διαφέρει στατιστικά σημαντικά από το μηδέν - $t - test$), υπάρχει ελαφριά αριστερή ασυμμετρία (-0.49) και η κατανομή είναι έντονα λεπτόκυρτη (6.476). Επίσης, από τον έλεγχο *Shapiro–Wilk* φαίνεται ότι τα δεδομένα δεν ακολουθούν κανονική κατανομή. Η απόκλιση από την κανονική κατανομή και ο χαρακτηρισμός της ως έντονα λεπτόκυρτης είναι μια ένδειξη μη γραμμικής δυναμικής στις αποδόσεις. Επίσης, από τον επαυξημένο έλεγχο *Dickey – Fuller* βλέπουμε ότι η υπόθεση μοναδιαίας ρίζας απορρίπτεται ($p - value \ll 0.01$) και συνεπάγεται μη-στασιμότητα.

Η τιμή της παραμέτρου *Hurst* για την χρονοσειρά είναι 0.6569453 , κάτι το οποίο δίνει μια ένδειξη εξάρτησης μακράς διάρκειας στις παρατηρήσεις. Το τροποποιημένο στατιστικό του *Lo*, παίρνει την τιμή 2.19 και ως εκ τούτου απορρίπτεται η μη ύπαρξη μακροπρόθεσμης μνήμης σε επίπεδο σημαντικότητας 1% . Χρησιμοποιώντας τον αλγόριθμο των *Hyndman – Khandakar* επιλέχθηκαν οι τιμές των παραμέτρων p και q του μοντέλου και στη συνέχεια εκτιμήθηκαν οι παράμετροι του μοντέλου (p, d, q) μέσω του αλγορίθμου των *Haslett – Raftery*. Επιλέχθηκε το *FARIMA*($0, d, 0$) μοντέλο, με $d = 0.06673517$. Η τιμή του κριτηρίου του *Akaike* ήταν -7346.435 . Τέλος, με επιπλέον ελέγχους που διενεργήθηκαν, οι υποθέσεις της ομοσκεδαστικότητας και της ανεξαρτησίας των καταλοίπων δεν απορρίπτονται.

Πίνακας 1: Αξιολόγηση των προβλέψεων με κριτήρια % RMSE (% MAE)

	0-21	22-65	66-100
<i>FARIMA</i> ($0, d, 0$)	2.3896(1.8618)	1.5432(1.2342)	1.6158(1.2796)
<i>ARIMA</i> ($2, 0, 5$)	2.3786(1.8366)	1.5448(1.2367)	1.6254(1.2795)
$\lambda=10$	2.4936(1.897)	1.5351(1.2339)	1.6072(1.2714)
$\lambda=25$	2.4206(1.8864)	1.5416(1.2333)	1.6145(1.2786)
$\lambda=75$	2.4527(1.8457)	1.5379(1.2316)	1.6081(1.2730)
$\lambda=100$	2.506(1.8598)	1.5409(1.2405)	1.6049(1.2701)
$\lambda=1000$	2.4882(1.8341)	1.5463(1.248)	1.6157(1.2795)
$\lambda=1600$	2.4613(1.8197)	1.5472(1.2492)	1.6149(1.2786)

	101-150	151-200	201-463
<i>FARIMA</i> ($0, d, 0$)	2.4958(1.8967)	2.2204(1.7818)	2.1024(1.5953)
<i>ARIMA</i> ($2, 0, 5$)	2.4967(1.8975)	2.2206(1.7820)	2.1028(1.5956)
$\lambda=10$	2.499(1.8995)	2.2207(1.7821)	2.1026(1.5955)
$\lambda=25$	2.4966(1.8974)	2.2205(1.7820)	2.1025(1.5954)
$\lambda=75$	2.5001(1.9006)	2.2221(1.7824)	2.1029(1.5957)
$\lambda=100$	2.5022(1.9024)	2.2214(1.7827)	2.1031(1.596)
$\lambda=1000$	2.4960(1.8969)	2.2204(1.7819)	2.1026(1.5954)
$\lambda=1600$	2.4963(1.897)	2.2205(1.782)	2.1026(1.5954)

Βάσει του κριτηρίου του *Akaike* επιλέχθηκε το *ARIMA*($2, 0, 5$) μοντέλο με τιμή κριτηρίου -7358.19 , η οποία είναι χαμηλότερη από εκείνη του αντίστοιχου *FARIMA* μοντέλου. Με επιπλέον ελέγχους δεν απορρίπτουμε την υπόθεση της ομοσκεδαστικότητας και την υπόθεση της ανεξαρτησίας των καταλοίπων.

Στον Πίνακα 1, οι προβλέψεις με τα εκτιμηθέντα υποδείγματα αξιολογούνται για διαφορετικές χρονικές περιόδους χρησιμοποιώντας τα κριτήρια % Root Mean Squared Error (RMSE) και % Mean Absolute Error (MAE). Το *FARIMA* μοντέλο, αν και στις πρώτες χρονικές περιόδους φαίνεται ελάχιστα πιο αναποτελεσματικό, εντούτοις στη συνέχεια φαίνεται να υπερτερεί και να δίνει ακριβέστερες προβλέψεις από το *ARIMA* μοντέλο. Παρατηρούμε ότι σε ορισμένα χρονικά φάσματα (όπως 22-65 και 66-100), ορισμένες διασπάσεις κυρίως για μικρά λ οδηγούν σε μια βελτίωση της προβλεπτικής ικανότητας χωρίς να υπερέχουν έναντι του ενιαίου *FARIMA* μοντέλου σε όλες τις χρονικές περιόδους.

6. ΣΥΜΠΕΡΑΣΜΑΤΑ ΚΑΙ ΜΕΛΛΟΝΤΙΚΕΣ ΚΑΤΕΥΘΥΝΣΕΙΣ

Σύμφωνα με τη μελέτη που παρουσιάστηκε οι αποδόσεις του Γενικού Δείκτη του ΧΑΑ παρουσιάζουν εξάρτηση μακράς διάρκειας. Το συμπέρασμα αυτό προκύπτει τόσο από την εκτίμηση του συντελεστή *Hurst* όσο και από τον έλεγχο υπόθεσης για εξάρτηση μακράς διάρκειας με το τροποποιημένο *R/S* στατιστικό. Για την πρόβλεψη του ΓΔ, σύμφωνα με τη μελέτη, τα *FARIMA* μοντέλα υπερέχουν έναντι των *ARIMA* στις μεσοπρόθεσμες και μακροπρόθεσμες προβλέψεις. Βάσει της τεχνικής ανάλυσης, υπάρχουν κύκλοι στις αποδόσεις. Η μέθοδος πρόβλεψης που βασίζεται στην διάσπαση της σειράς σε τάση και κύκλους με το φίλτρο Hodrick-Prescott φαίνεται να οδηγεί σε βελτιωμένες προβλέψεις μόνο σε ορισμένες περιπτώσεις. Μελλοντική έρευνα θα πρέπει να αφορά στον προσδιορισμό της καταλληλότερης τιμής για την παράμετρο λ , καθώς και στον καθορισμό καλύτερων και αποτελεσματικότερων φίλτρων ώστε να οδηγηθούμε σε ακόμα καλύτερες προβλέψεις.

ABSTRACT

In this paper we study long memory and its existence to General Index of Athens Stock Exchange, in order to create a suitable model for predictions. We consider daily returns from 2004 to 2009 and calculate Hurst parameter with Rescaled Range (*R/S*) method. Then, with modified (*R/S*) statistic we are testing for statistical significance of long memory. For prediction of daily returns we fit an optimal ARIMA model with Box-Jenkins methodology, and a suitable FARIMA model which is according to bibliography appropriate for time series with long memory. Finally we apply an approximation with separation of time series to trend and cycles with Hodrick-Prescott filter, then consider appropriate models for trend and cycles separately and finally construct predictions for the whole time series. Prediction ability of models done with criteria % Root Mean Square Error (RMSE) and % Mean Absolute Error (MAE) with test dataset daily returns from 2009 until 2011.

ΑΝΑΦΟΡΕΣ

- Barkoulas J., Baum C., Travlos N., (2000) Long Memory In the Greek Stock Market *Applied Financial Economics*, **10**, 177-184.
- Box G.E.P., Jenkins G.M., (1976) Time Series Analysis, Forecasting and Control *3rd edition, San Francisco:Holden-Bay*
- Cheung, Y.W. and K. S. Lai, (1995) A search for long memory in international stock market returns *J. of International Money and Finance* **14**, 597-615.
- Granger, C. J. and R. Joyeux, (1980) An introduction to long-memory time series models and fractional differencing *J. of Time Series Analysis*, **1**, 15-39.
- Haslett, J. and Raftery, A.E (1989) Space-time modelling with long-memory dependence: Assessing Ireland's wind power resource (with Discussion) *Journal of the Royal Statistical Society, series C - Applied Statistics*, **38**, 1-50.
- Hiemstra, C. and J. D. Jones (1997) Another look at long memory in common stock returns *Journal of Empirical Finance*, **4**, 373-401.
- Hodrick, R.J and E.C. Prescott (1997) Post-war US Business cycles:An Empirical Investigation *Journal of Money, Credit and Banking*, **29**, 1-16.
- Hosking, J.R.M. (1981) Fractal differencing , *Biometrika*, **68**, 165-176.
- Hurst, H.E., (1951) Long-term storage capacity of reservoirs *Transactions of the American Society of Civil Engineers*, **116**, 770-779.
- Hyndman, R.J. and Khandakar, Y. (2008) Automatic time series forecasting: The forecast package for R, *Journal of Statistical Software*, **26**, (3).
- Lo A.W., (1991) Long Term Memory in Stock Market Prices *Econometrica*, **59**, **5**, 1279-1313.
- Mandelbrot, B.B. and M.Taqqu (1979) Robust R/S Analysis of Long Run Serial Correlations *Bulletin of the Int. Statistical Institute*, **48**, book 2: 59-104.
- Mandelbrot, B.B. (1972) Statistical methodology for non-periodic cycles: From the covariance to R/S analysis, *Ann. of Economics and Soc. Measurement*, **1**, 259-90.
- Mandelbrot, B.B. and James R. Wallis, (1969) Robustness of the rescaled range R/S in the measurement of noncyclic long-run statistical dependence , *Water Resources Research*, **5**, 967-988.
- Panas, E., (2001) Estimating fractal dimension using stable distributions and exploring long memory through ARFIMA models in Athens Stock Exchange *Applied Financial Economics*, **11**, 395-402.
- Tolvi, J., (2003) Long memory in a small stock market *Economics Bulletin*, **7**, **3**, 1-13.
- Συριόπουλος Κώστας και Σιρλαντζής Κώστας, (1993) Χρηματιστηριακές Διακυμάνσεις και Προσδιοριστικό Χάος, *Σπουδαί, Τόμος 43*, Τεύχος 3-4 .



ΣΥΓΚΡΙΣΗ ΤΩΝ ΔΕΙΚΤΩΝ ΠΟΙΚΙΛΟΤΗΤΑΣ ΧΡΗΣΙΜΟΠΟΙΩΝΤΑΣ ΤΗΝ ΜΕΘΟΔΟ JACKKNIFE

Κετζάκη Ελένη¹, Κετζάκη Δήμητρα², Φαρμάκης Νικόλαος¹

¹Τμήμα Μαθηματικών, Α.Π.Θ.

eketzaki@yahoo.gr, farmakis@math.auth.gr

²Τμήμα Ηλεκτρολόγων Μηχανικών & Μηχανικών Υπολογιστών, Α.Π.Θ.

dketzaki@auth.gr

ΠΕΡΙΛΗΨΗ

Οι δείκτες ποικιλότητας αποτελούν ένα στατιστικό μέτρο υπολογισμού της ποικιλομορφίας ενός πληθυσμού, ο οποίος αποτελείται από διαφορετικά είδη. Η μελέτη των δεικτών ποικιλότητας παρουσιάζει ιδιαίτερο ενδιαφέρον, διότι είναι ευρύ το φάσμα των επιστημονικών πεδίων στα οποία μπορούν να εφαρμοστούν. Αρχικά στην παρούσα εργασία θα περιγραφεί η έννοια και τα είδη της ποικιλότητας. Στην συνέχεια θα παρουσιαστούν ορισμένοι δείκτες ποικιλότητας και έπειτα θα γίνει περιγραφή της μεθόδου Jackknife. Σκοπός αυτής της εργασίας είναι να εφαρμοστεί η μέθοδος Jackknife, για να συγκριθούν οι δείκτες ποικιλότητας. Ολοκληρώνοντας την εργασία περιγράφεται μία εφαρμογή που πραγματοποιήθηκε για να συγκριθούν ορισμένοι δείκτες ποικιλότητας και θα επιλεγεί αυτός που αποδίδει μικρότερο σφάλμα στον υπολογισμό της ποικιλότητας.

Λέξεις κλειδιά: δείκτες ποικιλότητας, μέθοδος Jackknife.

1. ΕΙΣΑΓΩΓΗ

Η μελέτη των δεικτών ποικιλότητας (diversity indices) παρουσιάζει ιδιαίτερο ενδιαφέρον όχι μόνο γιατί αποτελούν ένα μέτρο για την μέτρηση της ποικιλότητας, αλλά διότι είναι ευρύ το φάσμα των επιστημονικών πεδίων στα οποία μπορούν να εφαρμοστούν. Συναντώνται συνήθως στην επιστήμη της οικολογίας, με σκοπό να μετρήσουν την ποικιλότητα σ' ένα οικοσύστημα. Επιπλέον εφαρμόζονται σε δημογραφικά και οικονομικά δεδομένα, αλλά ακόμα και στην επιστήμη της πληροφορίας (information science) για να περιγράψουν την πολυπλοκότητα ενός συνόλου πληροφοριών.

Για να πραγματοποιηθεί η μέτρηση της ποικιλότητας των ειδών ενός πληθυσμού, αρχικά θα πρέπει να προσδιοριστεί το είδος της ποικιλότητας και στην συνέχεια να γίνει επιλογή των κατάλληλων δεικτών, οι οποίοι σχετίζονται με το συγκεκριμένο είδος ποικιλότητας. Τις περισσότερες φορές δεν είναι μόνο ένας ο δείκτης ο οποίος μπορεί να χρησιμοποιηθεί αλλά δύο ή περισσότεροι δείκτες. Στην εργασία αυτή

αρχικά θα περιγραφούν τα είδη ποικιλότητας και οι αντίστοιχοι δείκτες που τα αφορούν και στη συνέχεια θα παρουσιαστεί μια εφαρμογή για δεδομένα τα οποία αφορούν συγκεκριμένο είδος ποικιλότητας, με σκοπό να επιλεγεί εκείνος ο δείκτης που είναι καταλληλότερος για τον υπολογισμό της βιο-ποικιλότητας.

2. ΠΟΙΚΙΛΟΤΗΤΑ

Ποικιλότητα ονομάζεται η ιδιότητα που έχουν τα στοιχεία μίας ομάδας να παρουσιάζουν ποικιλία στην μορφή. Για παράδειγμα στην περίπτωση που μελετηθούν οι μικροοργανισμοί που κατοικούν σε μία λίμνη και διαπιστώσουμε ότι υπάρχουν πολλά και διαφορετικά είδη μικροοργανισμών, αυτό είναι γεγονός το οποίο φανερώνει την ύπαρξη ποικιλομορφίας στην συγκεκριμένη περιοχή.

Η μελέτη της ποικιλότητας, είναι θέμα που απασχολεί πολλές επιστήμες, όπως την βιολογία, την γεωπονία, την οικολογία, την οικονομία, αλλά και τις κοινωνικές επιστήμες.

Η σημαντικότητα της ποικιλότητας στην βιολογία εύκολα μπορεί να διαπιστωθεί από το γεγονός, ότι η βιοποικιλότητα χρειάστηκε να οριστεί στα πλαίσια ενός νομικού πλαισίου. Σύμφωνα με τον νόμο:

« Βιοποικιλότητα είναι η ποικιλία των ζώντων οργανισμών πάσης προελεύσεως, περιλαμβανομένων, μεταξύ άλλων, των χερσαίων, θαλασσίων και άλλων υδατικών οικοσυστημάτων και οικολογικών συμπλεγμάτων, των οποίων αποτελούν μέρος. Επίσης, περιλαμβάνεται η ποικιλότητα εντός των ειδών, μεταξύ ειδών και οικοσυστημάτων (αρ. 2, ν.2204/1994, 59 Α')»

Η ποικιλότητα διακρίνεται στα ακόλουθα είδη, ορίζεται ως α-ποικιλότητα (α-diversity) η ποικιλότητα των ειδών που συναντώνται σε μία τοποθεσία, β-ποικιλότητα (β-diversity) η διαφορά στην ποικιλότητα ανάμεσα σε δύο διαφορετικές τοποθεσίες. Η γ-ποικιλότητα (γ-diversity) αναφέρεται στον συνολικό πλούτο των ειδών που υπάρχουν σε μία μεγάλη περιοχή. Ε-ποικιλότητα (E-diversity) είναι η συνολική ποικιλότητα περιοχών γ- ποικιλότητας και τέλος η point diversity είναι η ποικιλότητα των ειδών ενός δείγματος.

3. ΔΕΙΚΤΕΣ ΠΟΙΚΙΛΟΤΗΤΑΣ

Οι δείκτες ποικιλότητας είναι το μέτρο με το οποίο μετράται η ποικιλότητα. Δείκτες ποικιλότητας υπάρχουν πολλοί, θεωρώντας ως κριτήριο το είδος της ποικιλότητας που πρέπει να εξεταστεί, γίνεται και η επιλογή του σωστού δείκτη.

Στην ενότητα αυτή θα περιγραφούν ορισμένοι δείκτες ποικιλότητας. Ο δείκτης του Simpson, επινοήθηκε και μελετήθηκε από τον Simpson το 1949. Στην βιβλιογραφία συναντάται σε τρεις διαφορετικές μορφές. Έστω n το πλήθος του δείγματος από μία περιοχή, στο οποίο διακρίνονται διαφορετικά είδη. Ας είναι p_i η αναλογία του πλήθους των στοιχείων του i -είδους ως προς το συνολικό πλήθος. Ο δείκτης Simpson

$$D = \sum_{i=1}^n p_i^2, \quad D \in \left[\frac{1}{n}, 1 \right] \quad (1)$$

σχετίζεται με την πιθανότητα δύο στοιχεία τα οποία επιλέγονται τυχαία από ένα δείγμα να ανήκουν στο ίδιο είδος. Όταν το $n \rightarrow \infty$ παρατηρείτε ότι το πεδίο τιμών του D συγκλίνει στο $[0,1]$. Όσο οι τιμές του δείκτη D πλησιάζουν το 1 συμπεραίνουμε πως η ποικιλότητα στο δείγμα δεν είναι πολύ μεγάλη, ενώ για τιμές κοντά στο 0 διαπιστώνουμε ότι υπάρχει πολύ μεγάλη ποικιλότητα. Ο δείκτης ποικιλότητας του Simpson ο οποίος υπολογίζεται από τον τύπο

$$S_D = 1 - D, \quad S_D \in \left[0, 1 - \frac{1}{n} \right] \quad (2)$$

σχετίζεται με την πιθανότητα δύο ξεχωριστά αντικείμενα τα οποία έχουν επιλεγθεί τυχαία να μην ανήκουν στο ίδιο είδος. Σε περίπτωση που το $n \rightarrow \infty$ παρατηρείτε ότι το πεδίο τιμών $S_D \in [0,1]$. Σε αυτή την περίπτωση όμως τα αποτελέσματα ερμηνεύονται διαφορετικά, όταν οι τιμές του δείκτη είναι κοντά στο 0, αυτό συνεπάγεται ότι δεν υπάρχει ποικιλομορφία, ενώ τιμές του δείκτη κοντά στο 1 φανερώνουν μεγάλη ποικιλομορφία.

Επιπλέον ο αντίστροφος δείκτης ποικιλότητας του Simpson υπολογίζεται από τον ακόλουθο τύπο

$$S_D^* = 1/D, \quad S_D^* \in [1, n] \quad (3)$$

Παρατηρώντας τον τύπο (3) διαπιστώνουμε ότι όταν τιμές που προκύπτουν από τον δείκτη βρίσκονται κοντά στην τιμή 1, τότε στο δείγμα δεν υπάρχει ποικιλότητα στα είδη. Επιπλέον δεν υπάρχει άνω φράγμα στις τιμές του δείκτη όταν το $n \rightarrow \infty$.

Ο δείκτης Shannon χρησιμοποιείται ως δείκτης μέτρησης ποικιλομορφίας στην περίπτωση που τα δεδομένα είναι ομαδοποιημένα. Ας υποθέσουμε ότι ο πληθυσμός αποτελείται από n στοιχεία τα οποία είναι χωρισμένα σε k ομάδες, τότε ο δείκτης του Shannon υπολογίζεται από τον ακόλουθο τύπο

$$H = -\sum_{i=1}^k p_i \cdot \ln p_i, \quad H \in [0, \ln k] \quad (4)$$

Όταν οι τιμές που έχουν υπολογιστεί για τον δείκτη H βρίσκονται κοντά στο 0, αυτό φανερώνει ότι δεν υπάρχει μεγάλη ποικιλομορφία.

Ο δείκτης του Margalef ή species richness δίνεται από τον τύπο

$$d = (S - 1) / \ln N, \quad d \in [0, (N - 1) / \ln N] \quad (5)$$

Στον παραπάνω τύπο με S θεωρούμε τον αριθμό των ειδών και με N τον αριθμό των ατόμων. Ο δείκτης του Margalef έχει τιμές κοντά στο 0 όταν παρατηρηθεί μόνο ένα είδος επομένως δεν υπάρχει μεγάλη ποικιλότητα.

Ο δείκτης ο οποίος χρησιμοποιείται για την μέτρηση της β -ποικιλότητας είναι ο ακόλουθος

$$\beta = (S_1 - c) + (S_2 - c) \quad (6)$$

Όπου S_1 : ο συνολικός αριθμός των ειδών στην πρώτη κοινότητα, S_2 : ο συνολικός αριθμός των ειδών στην δεύτερη κοινότητα και c ο αριθμός των ειδών που είναι

κοινός και στις δύο κοινότητες. Ο δείκτης ποικιλότητας εκφράζει το συνολικό αριθμό των ειδών τα οποία είναι μοναδικά μεταξύ των κοινοτήτων.

Ο δείκτης του Sorensen, είναι ένας δείκτης β-ποικιλότητας ο οποίος ορίζεται από το πηλίκο

$$\beta_s = 2c / (S_1 + S_2) \quad (7)$$

όπου S_i : ο συνολικός αριθμός των ειδών στην i -κοινότητα και c ο αριθμός των ειδών και στις δύο κοινότητες. Εύκολα μπορεί να διαπιστωθεί ότι όταν ο αριθμός των δεικτών οι οποίοι βρίσκονται και στις 2 κοινότητες είναι 0, τότε και ο δείκτης του Sorensen είναι 0.

Τέλος ο δείκτης του Whittaker, είναι επίσης ένας δείκτης β-ποικιλότητας και ισούται με

$$\beta_w = (S/a) - 1 \quad (8)$$

Όπου S παριστάνει τον συνολικό αριθμό των ειδών και a είναι ο μέσος αριθμός των τιμών οι οποίες βρίσκονται εντός των κοινοτήτων.

4. ΜΕΘΟΔΟΣ JACKKNIFE

Η μέθοδος Jackknife αποτελεί μία μέθοδο επαναδειγματοληψίας. Χρησιμοποιήθηκε σε αυτή την εργασία με σκοπό να συγκρίνουμε τους δείκτες ποικιλότητας. Ο δείκτης ποικιλότητας είναι ένα είδος συνάρτησης η οποία εφαρμόζεται για ένα πλήθος δεδομένων και δίνει ως αποτέλεσμα μία τιμή. Εφαρμόζοντας στα δεδομένα την μέθοδο Jackknife μας δίνεται η δυνατότητα να πραγματοποιήσουμε συγκρίσεις οι οποίες δεν βασίζονται σε μία μόνο αριθμητική τιμή αλλά μπορούν να εκτιμηθούν διαστήματα εμπιστοσύνης και να υπολογιστούν σφάλματα.

Στην ενότητα αυτή θα παρουσιαστεί περιληπτικά η διαδικασία της μεθόδου Jackknife και στην επόμενη ενότητα θα πραγματοποιηθεί μία εφαρμογή της μεθόδου αυτής.

Εστω ότι δίνονται n παρατηρήσεις x_1, \dots, x_n . Αρχικά κατασκευάζονται τα Jackknife δείγματα, μεγέθους $n-1$ το κάθε ένα, τα οποία προκύπτουν από τις αρχικές παρατηρήσεις αφαιρώντας την τιμή x_i .

Υπολογίζεται η εκτιμήτρια συνάρτηση για κάθε ένα από τα Jackknife δείγματα

$$\hat{g}_i(x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n) \quad (9)$$

Και στην συνέχεια υπολογίζεται η ψευδοτιμή για κάθε ένα από τα Jackknife δείγματα η οποία θα ισούται με

$$\hat{g}_i^* = n \cdot \hat{g} - (n-1) \cdot \hat{g}_i \quad (10)$$

όπου \hat{g} συμβολίζει την εκτίμηση με όλες τις n παρατηρήσεις.

Η Jackknife εκτιμήτρια είναι ίση με τον μέσο όρο των ψευδοτιμών των Jackknife δειγμάτων και επομένως θα δίνεται από τον ακόλουθο τύπο

$$J(\hat{g}) = \frac{1}{n} \sum \hat{g}_i^* \quad (11)$$

Το τυπικό σφάλμα της jackknife εκτιμήτριας είναι ίσο με

$$\hat{\sigma}_{\text{jackknife}}(\hat{g}) = \left[\frac{1}{n(n-1)} \cdot \sum_{i=1}^n (\hat{g}_i^* - J(\hat{g}))^2 \right]^{1/2} \quad (12)$$

Τέλος το προσεγγιστικό διάστημα εμπιστοσύνης δίνεται από τις παρακάτω σχέσεις

$$J(\hat{g}) \pm t_{0.975, n-1} \cdot (\hat{\sigma}_{\text{jackknife}}(\hat{g})) \quad (13)$$

5. ΧΡΗΣΗ JACKKNIFE ΓΙΑ ΣΥΓΚΡΙΣΗ ΔΕΙΚΤΩΝ

Στην εφαρμογή που ακολουθεί χρησιμοποιούνται δεδομένα από μία έρευνα, τα οποία έχουν προκύψει ως αποτέλεσμα δειγματοληψίας κατά την διάρκεια δύο ετών (2000-2002). Οι δειγματοληψίες πραγματοποιήθηκαν σε μία λίμνη με σκοπό να εξεταστεί η περιεκτικότητα μικροοργανισμών στην συγκεκριμένη λίμνη σε διαφορετικές χρονικές στιγμές. Συνολικά πραγματοποιήθηκαν 76 μετρήσεις, σε διαφορετικές χρονικές περιόδους. Το πόρισμα της έρευνας για την οποία πραγματοποιήθηκε η δειγματοληψία, δεν αποτελεί αντικείμενο της παρούσας εργασίας. Σκοπός της εφαρμογής είναι να εξεταστεί ποιος από τους δείκτες ποικιλότητας μετρά με μεγαλύτερη ακρίβεια την ποικιλότητα στην περιοχή αυτή.

Στην συγκεκριμένη δειγματοληψία το είδος της ποικιλότητας είναι η α – ποικιλότητα. Επομένως θα συγκριθούν οι δείκτες Simpson, Margalef και Shannon. Οι παραπάνω δείκτες δεν έχουν το ίδιο πεδίο τιμών συνεπώς, δεν μπορούμε να συγκρίνουμε την τιμή ενός δείκτη με την τιμή ενός άλλου δείκτη, για τον λόγο αυτό θα χρησιμοποιήσουμε την Jackknife μέθοδο.

Η μέθοδος Jackknife υπολογίζει το σφάλμα και το διάστημα εμπιστοσύνης για κάθε εκτιμητή χωριστά. Θα θεωρήσουμε ότι καλύτερος εκτιμητής είναι αυτός ο οποίος αποδίδει για το ίδιο δείγμα μικρότερο σφάλμα. Στην ακόλουθη εφαρμογή θα θεωρήσουμε τους δείκτες Simpson, Margalef και Shannon ως τρεις διαφορετικούς εκτιμητές. Ας υποθέσουμε ότι το αποτέλεσμα της πρώτης μέτρησης (πίνακας 1) είναι ένα δείγμα στο οποίο θα εφαρμόσουμε την μέθοδο Jackknife.

Πίνακας 1. Αποτελέσματα 1^{ns} μέτρησης

ΕΙΔΟΣ (Λατινική Ονομασία)	Αποτέλεσμα 1^{ns} μέτρησης
Πλεκόπτερα (<i>Plecoptera</i>)	6,32
Εφημερόπτερα (<i>Ephemeroptera</i>)	42,32
Τριχόπτερα (<i>Trichoptera</i>)	0,95
Οδοντόγναθα (<i>Odonata</i>)	0,06
Δίπτερα (<i>Diptera</i>)	17,85
Κολεόπτερα (<i>Coleoptera</i>)	0,53
Μεγαλόπτερα (<i>Megaloptera</i>)	0,00
Αμφίποδα (<i>Amphipoda</i>)	0,00
Ισόποδα (<i>Isopoda</i>)	0,00
Δίθυρα (<i>Bivalia</i>)	0,00
Ολιγόχαιτοι (<i>oligochaeta</i>)	6,68

Ετερόπτερα (<i>Heteroptera</i>)	0,00
Γαστερόποδα (<i>Gastropoda</i>)	0,00
Πλατυέλμινθες (<i>Platyhelminthes</i>)	0,00
Βδέλλες (<i>Hirudinea</i>)	0,00
Δεκάποδα (<i>Decapoda</i>)	0,00
Αραχνοειδή (<i>Arachnida</i>)	0,00
Νευρόπτερα (<i>Neuroptera</i>)	0,00
Λεπιδόπτερα (<i>Lepidoptera</i>)	0,00
Άγνωστα (<i>Unknown</i>)	25,30

Αρχικά υπολογίζουμε τον δείκτη Simpson για τα είκοσι είδη του πίνακα 1. Στην συνέχεια υπολογίζεται η εκτιμητή συνάρτηση (δείκτης Simpson) για κάθε ένα από τα είκοσι Jackknife δείγματα (σχέση 9), αυτό σημαίνει ότι ο δείκτης Simpson θα υπολογιστεί είκοσι φορές, έτσι ώστε κάθε φορά να αφαιρείται ένα διαφορετικό είδος από το αρχικό δείγμα. Έπειτα θα υπολογιστούν οι ψευδοτιμές, ο Jackknife εκτιμητής το σφάλμα και το αντίστοιχο δ.ε. (σχέση 10,11,12). Η παραπάνω διαδικασία θα πρέπει να επαναληφθεί για τον δείκτη Shannon και για τον δείκτη Margalef. Στο τέλος θα πρέπει να συγκριθεί το σφάλμα κάθε εκτιμητή για τα ίδια δεδομένα. Καλύτερος εκτιμητής θα είναι αυτός που υπολογίζει την ποικιλότητα με λιγότερο σφάλμα. Παρατηρώντας όμως τον δείκτη Shannon διαπιστώνεται από τον ορισμό του ότι δεν ορίζεται όταν κάποιο είδος δεν εμφανίζεται, επομένως το $p_i = 0$. Η σύγκριση των τριών δεικτών θα περιοριστεί μόνο για τα 8 είδη που εμφανίζονται. Τα συμπεράσματα της σύγκρισης περιγράφονται στον πίνακα 2.

Πίνακας 2. Εκτιμητές Jackknife, σφάλμα και δ.ε. για τους δείκτες κατά την 1^η μέτρηση

Δείκτης Simpson	
Εκτιμητής Jackknife	J=0.67773
Σφάλμα	$\sigma=0.00108$
Διάστημα εμπιστοσύνης	(0.67517 , 0.68028)
Δείκτης Shannon	
Εκτιμητής Jackknife	J=1.32323
Σφάλμα	$\sigma=0.00267$
Διάστημα εμπιστοσύνης	(1.31691, 1.32951)
Δείκτης Margalef	
Εκτιμητής Jackknife	J=1.34862
Σφάλμα	$\sigma=0.00163$
Διάστημα εμπιστοσύνης	(1.34476 , 1.35247)

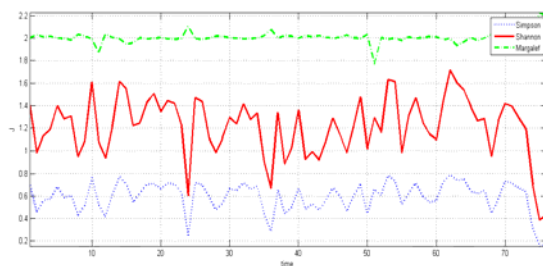
Διαπιστώνεται ότι η τιμή με το μικρότερο σφάλμα έχει υπολογιστεί για τον δείκτη Simpson και ισούται με $\sigma = 0,00108$. Το συμπέρασμα αυτό αφορά μόνο ένα δείγμα το οποίο είναι αποτέλεσμα της πρώτης μέτρησης. Επαναλαμβάνουμε την ίδια διαδικασία για τις υπόλοιπες μετρήσεις, ορισμένα από τα αποτελέσματα περιγράφονται στον πίνακα 3. #

Πίνακας 3. Εκτιμητές Jackknife, σφάλμα για τους δείκτες κατά τις 76 μετρήσεις

	J Simpson	σ Simpson	J Shannon	σ Shannon	J Margalef	σ Margalef
1 ^η μέτρηση	0,677739706	0,001081134	1,323232336	0,002670782	1,348626707	0,001633093
2 ^η μέτρηση	0,442672455	0,002939334	0,915515275	0,005411711	1,376223818	0,004628692
3 ^η μέτρηση	0,553360654	0,001940623	1,079508597	0,003795162	1,578483847	0,003495377
4 ^η μέτρηση	0,573269885	0,002110661	1,188588415	0,00434892	2,015731532	0,004615385
5 ^η μέτρηση	0,687194599	0,001171865	1,398149085	0,002667678	2,000533466	0,002544255
...
75 ^η μέτρηση	0,15456311	0,00501856	0,386812013	0,010577963	2,224772177	0,028334978
76 ^η μέτρηση	0,175192937	0,005694318	0,433582229	0,011428864	2,194652626	0,025214593

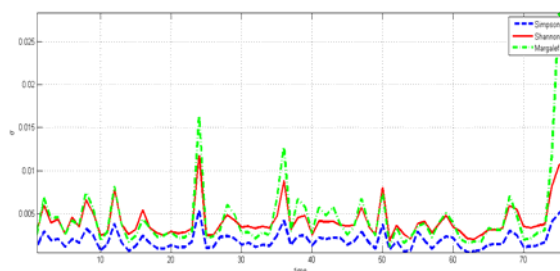
Τα δεδομένα του παραπάνω πίνακα συνοψίζονται στις ακόλουθες γραφικές παραστάσεις.

Εικόνα 1 : Γράφημα εκτιμητών Jackknife ανά δείκτη για τις 76 μετρήσεις.



#

Εικόνα 2 : Γράφημα σφαλμάτων ανά δείκτη για τις 76 μετρήσεις

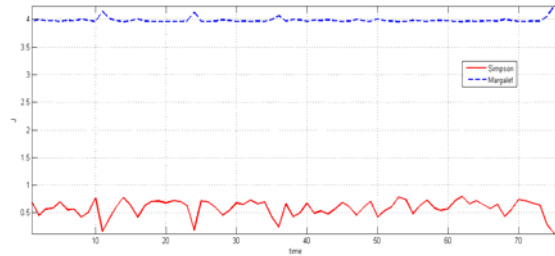


#

Διαπιστώνεται από την εικόνα 2 ότι οι τιμές του σφάλματος για τον δείκτη Simpson είναι μικρότερες για όλες τις μετρήσεις. Συνεπώς ο δείκτης Simpson είναι προτιμότερος για τον υπολογισμό α - ποικιλότητας.

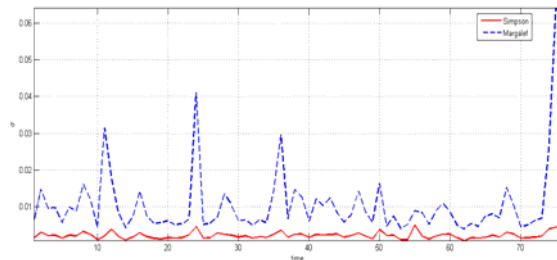
Το παραπάνω συμπέρασμα προέκυψε έχοντας παραλείψει τις μηδενικές τιμές από τις μετρήσεις. Οι δείκτες Simpson και Margalef μπορούν να υπολογιστούν και στην περίπτωση που κατά την μέτρηση προκύψουν μηδενικές τιμές. #

Εικόνα 3: Γράφημα εκτιμητών Jackknife ανά δείκτη συμπεριλαμβανομένου και των μηδενικών τιμών για τις 76 μετρήσεις



Συγκρίθηκαν οι δείκτες Simpson και Margalef για να διαπιστωθεί αν οι μηδενικές τιμές επηρεάζουν την συμπεριφορά τους.

Εικόνα 4: Γράφημα σφαλμάτων ανά δείκτη συμπεριλαμβανομένου και των μηδενικών τιμών για τις 76 μετρήσεις



Από την εικόνα 4 διαπιστώνεται ότι οι τιμές των σφαλμάτων του δείκτη Simpson είναι μικρότερες για όλες τις μετρήσεις, επομένως ο δείκτης Simpson είναι προτιμότερος για την μέτρηση α – ποικιλότητας.

6. ΣΥΜΠΕΡΑΣΜΑΤΑ

Στην εργασία αυτή περιγράφηκαν αρχικά τα είδη ποικιλότητας, οι δείκτες ποικιλότητας και η μέθοδος Jackknife. Στην συνέχεια συγκρίθηκαν οι δείκτες α – ποικιλότητας. Ο δείκτης Simpson συμπεραίνεται ότι είναι προτιμότερος για τον υπολογισμό της α – ποικιλότητας σε μία περιοχή, διότι το σφάλμα που υπολογίστηκε για αυτόν μέσω της μεθόδου Jackknife είναι μικρότερο συγκριτικά με τα σφάλματα που αποδίδουν οι υπόλοιποι δείκτες.

ABSTRACT

Diversity indices are a statistic which intend to measure the diversity of any population in which each member belongs to a unique group. Diversity indices are of great interest because they can be used in many fields of study. For example they are used in biology in order to measure biodiversity, in economics, in demography and in

information science. In this paper our purpose is to use Jackknife method in order to compare diversity indices and choose the most suitable among them.

ΑΝΑΦΟΡΕΣ

- Zahl Samuel. (1977). Jackknifing an index of diversity. *Ecology*, **58**, 907-913
- Peet K. R. (1975). Relative diversity indices. *Ecology*, **56**, 496-498
- Hill M. O. (1973). Diversity and evenness : a unifying notation and its consequences. *Ecology*, **57**, 427-432
- Φωκιανός, Κ. Χαράλάμπους, Χ. (2008). *Εισαγωγή στην R*, Εκδόσεις Πανεπιστημίου Κύπρου, Λευκωσία
- Φαρμάκης, Ν. (2009). *Εισαγωγή στη Δειγματοληψία*, Εκδόσεις Χριστοδουλίδη, Θεσσαλονίκη

ΕΦΑΡΜΟΓΗ ΤΗΣ ΑΝΑΛΥΣΗΣ ΕΠΙΒΙΩΣΗΣ ΓΙΑ ΤΗΝ ΑΞΙΟΛΟΓΗΣΗ ΤΗΣ ΑΥΞΗΣΗΣ ΤΩΝ ΔΕΝΤΡΩΝ ΕΛΑΤΗΣ (*Abies borisii-regis*) ΣΤΟ ΠΑΝΕΠΙΣΤΗΜΙΑΚΟ ΔΑΣΟΣ ΠΕΡΤΟΥΛΙΟΥ

Κιτικίδου Κυριακή¹, Μάτης Κωνσταντίνος²

¹ Δημοκρίτειο Πανεπιστήμιο Θράκης, Τμήμα Δασολογίας και Διαχείρισης Περιβάλλοντος και Φυσικών Πόρων

kkitikid@fmenr.duth.gr

² Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης, Σχολή Δασολογίας και Φυσικού Περιβάλλοντος
matis@for.auth.gr

ΠΕΡΙΛΗΨΗ

Παραδοσιακά, η αύξηση των δέντρων αξιολογείται στις δασικές απογραφές με διάφορες μετρήσεις υψών και διαμέτρων. Σε αυτή την εργασία προτείνεται η εφαρμογή της ανάλυσης επιβίωσης, με σκοπό την ανάλυση της αύξησης των δέντρων. Μετρήσεις ατομικής μείωσης της διαμέτρου των δέντρων καθ' ύψος χρησιμοποιήθηκαν για την εκτίμηση συναρτήσεων επιβίωσης και στιγμιαίου κινδύνου, για την ελάτη (*Abies borisii-regis*) του Πανεπιστημιακού Δάσους Περτουλίου. Αυτές οι εκτιμήσεις παρείχαν ενδείξεις αυξημένου ρυθμού μείωσης της διαμέτρου σε ύψος 9.3 – 13.3m. Η τεχνική της ανάλυσης επιβίωσης, χρησιμοποιώντας τη μείωση της διαμέτρου καθ' ύψος, θα μπορούσε να βοηθήσει τους δασολόγους να αξιολογήσουν τις τοπικές τάσεις αύξησης των δέντρων.

Λέξεις Κλειδιά: Ανάλυση επιβίωσης, αύξηση δέντρου, δασική απογραφή, συνάρτηση στιγμιαίου κινδύνου.

1. ΕΙΣΑΓΩΓΗ

Ιστορικά, η δασική επιστήμη είχε επικεντρωθεί στην κατάρτιση μοντέλων για την ατομική αύξηση του δέντρου, με στόχο την ενσωμάτωσή τους σε μοντέλα αύξησης και παραγωγής ξυλείας (Stage, 1973, Daniels and Burkhardt, 1975, Hamilton and Edwards, 1976, Monserud, 1976 και Buchman et al., 1983). Παρόλο που άλλες επιστήμες, οι οποίες παρακολουθούν πληθυσμούς ζωντανών οργανισμών, όπως η κτηνιατρική και η ιατρική, είχαν αναπτύξει μεθόδους αξιολόγησης της αύξησης πέρα από την ατομική, στη δασική επιστήμη εφαρμόζονται σχετικά λίγες μεθοδολογίες για την αξιολόγηση της αύξησης των δέντρων, αν αυτά μελετηθούν ως πληθυσμός, δηλαδή της αύξησης της συστάδας (Hawkes, 2000). Οι κοινά χρησιμοποιούμενες μέθοδοι αξιολόγησης της αύξησης στερούνται μεθοδολογίας για την ενσωμάτωση της χρονικά εξαρτώμενης φύσης της αύξησης του δέντρου, ελέγχων υποθέσεων,

χειρισμού των αποκομμένων παρατηρήσεων (censored observations) (δεδομένων που αφορούν σε δέντρα που δεν είχαν ολοκληρωμένη παρακολούθηση) και ελέγχων για την επίδραση διάφορων μεταβλητών.

Οι αναλυτικές μέθοδοι που αναπτύχθηκαν στην ιατρική επιστήμη, για τη συλλογική έκφραση της επιβίωσης, θα μπορούσαν να δώσουν τη βάση για την ανάπτυξη μεθόδων εκτίμησης της δασικής αύξησης (Hargcombe, 1987 και Zens and Peart, 2003). Η ανάλυση επιβίωσης ορίζεται πιο συχνά ως μια κατηγορία στατιστικών μεθόδων για τη μελέτη της συχνότητας εμφάνισης και της στιγμής εμφάνισης διάφορων γεγονότων – πιο συχνά του θανάτου (Berkson and Gage, 1950, Cox and Oakes, 1984 και Vahtsevanos et al., 2009). Η μοναδικότητα της ανάλυσης επιβίωσης, οφείλεται στο γεγονός ότι επιτρέπει τη χρήση αποκομμένων παρατηρήσεων (ατόμων για τους οποίους δεν έχουμε τον ακριβή χρόνο θανάτου) και διάφορων μεταβλητών, χωρίς τον περιορισμό της κανονικότητας των κατανομών (Collett, 1994 και Vahtsevanos et al., 2009). Πρώτος ο Waters (1969) πρότεινε τη χρήση της ανάλυσης επιβίωσης σε δασικά θέματα αλλά εφαρμογές τέτοιου είδους περιορίστηκαν κυρίως σε δασικές απογραφές ομήλικων δασών (Morse and Kulman, 1984, Amateis et al., 1997, Volney, 1998 και Wyckoff and Clark, 2000), σχεδιασμούς δασικών πειραμάτων (Reams et al., 1988, Burgman et al., 1994 και Preisler and Slaughter, 1997) και καταρτίσεις μαζοπινάκων συστάδων (Rose, 2002). Παρόλο που έχει γίνει σημαντική δουλειά στην ανάπτυξη μεθόδων εκτίμησης συναρτήσεων επιβίωσης και στιγμιαίου κινδύνου σε σχεδιασμούς δασικών πειραμάτων (Preisler and Slaughter, 1997 και Volney, 1998) και σε σχέση με την αύξηση της διαμέτρου (Bigler and Bugmann, 2004a και 2004b, Kitikidou and Apostolopoulou, 2011), η τεχνική της ανάλυσης επιβίωσης δεν εφαρμόστηκε ευρέως στις δασικές απογραφές, λόγω της εγγενούς έλλειψης λεπτομερών πληροφοριών σχετικά με το χρόνο ανόδου (ο χρόνος που χρειάζεται το δέντρο για να περάσει από μια κλάση διαμέτρου στην επόμενη) και την ηλικία των δέντρων (Flewelling and Monserud, 2002).

Σκοπός της παρούσας μελέτης είναι να εκτιμήσει και να ερμηνεύσει τις βασικές συναρτήσεις στη στατιστική μέθοδο της ανάλυσης επιβίωσης Kaplan–Meier με συναρτήσεις επιβίωσης (survival functions) και συναρτήσεις στιγμιαίου κινδύνου (hazard functions), σε μια χρονική σειρά που ορίζεται από τη μείωση της διαμέτρου καθ' ύψος, για την ελάτη (*Abies borisii-regis*) του Πανεπιστημιακού Δάσους Πετρουλίου.

2. ΥΛΙΚΑ ΚΑΙ ΜΕΘΟΔΟΙ

Τα δεδομένα που χρησιμοποιήθηκαν σε αυτή τη μελέτη προέρχονται από τα αρχεία του Εργαστηρίου Δασικής Βιομετρίας του Αριστοτέλειου Πανεπιστημίου Θεσσαλονίκης. Στα 1061 δέντρα του δείγματος μετρήθηκαν η πρεμνική διάμετρος d0.3, η στηθιαία διάμετρος d1.3 και οι διαμέτροι ανά 4 m πάνω από το στηθιαίο ύψος d5.3, d9.3, d13.3 κτλ ως την τελευταία μετρούμενη διάμετρο πριν το συνολικό ύψος h. Βασικά περιγραφικά στατιστικά των δεδομένων δίνονται στον Πίνακα 1.

Πίνακας 1. Περιγραφικά στατιστικά του δείγματος

Μεταβλητή	min (m)	max (m)	Μέσος όρος (m)	Τυπική απόκλιση (m)
d0.3	0,14	0,93	0,3967	0,1569
d1.3	0,13	0,87	0,3375	0,1363
h	5,40	35,10	17,4985	5,8806

Στη συνέχεια υπολογίστηκε η μείωση της διαμέτρου από το ένα διαδοχικό ύψος στο άλλο Δd_i ($i=0.3, 1.3, 5.3, 9.3, \dots$) και βρέθηκε η μέγιστη μείωση διαμέτρου για κάθε δέντρο $\max \Delta d$. Η μεταβλητή $\max \Delta d_cat$ για το δέντρο j ($j = 1, 2, \dots, 1061$) παίρνει την τιμή 0 αν $d0.3 - d1.3 = \max$, 1 αν $d1.3 - d5.3 = \max$, 2 αν $d5.3 - d9.3 = \max$, 3 αν $d9.3 - d13.3 = \max$, 4 αν $d13.3 - d17.3 = \max$, 5 αν $d17.3 - d21.3 = \max$, 6 αν $d21.3 - d25.3 = \max$, 7 αν $d25.3 - d29.3 = \max$, 8 αν $d29.3 - d33.3 = \max$. Η μεταβλητή $status$ για το δέντρο j ($j = 1, 2, \dots, 1061$) παίρνει την τιμή 0 αν $\max \Delta d > \overline{\max \Delta d}$ και 1 αν $\max \Delta d \leq \overline{\max \Delta d}$.

Οι Kaplan–Meier συναρτήσεις επιβίωσης και οι συναρτήσεις στιγμιαίου κινδύνου, χρησιμοποιούνται για την ποσοτικοποίηση της κατανομής πιθανότητας της θνησιμότητας σε έναν πληθυσμό (Muenchow, 1986). Η συνάρτηση επιβίωσης ορίζεται ως εξής (Berkson and Gage, 1950, Cox and Oakes, 1984 και Collett, 1994):

$$S(t) = P(T > t) \quad (1)$$

όπου $S(t)$ είναι η πιθανότητα να επέλθει ο θάνατος σε χρονική στιγμή μεταγενέστερη της χρονικής στιγμής t . Η τυχαία μεταβλητή T , που μετρά το χρόνο μέχρι το θάνατο, δεν έχει κανένα περιορισμό εκτός από το να έχει τιμή μεγαλύτερη από το μηδέν.

Η συνάρτηση στιγμιαίου κινδύνου εκφράζει τη στιγμιαία θνησιμότητα και επομένως είναι μια δεσμευμένη πιθανότητα (conditional probability) που ορίζεται ως εξής (Cox and Oakes, 1984 και Collett, 1994):

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T \leq t + \Delta t | T \geq t)}{\Delta t} \quad (2)$$

όπου $h(t)$ είναι η πιθανότητα να συμβεί ένας θάνατος ακριβώς τη χρονική στιγμή t , με δεδομένο ότι δε συνέβη πριν από αυτή. Σημειώνουμε πως οι συναρτήσεις (1) και (2) είναι μαθηματικές συναρτήσεις όπου η μια προκύπτει από την άλλη.

Η μηδενική υπόθεση, ότι οι συναρτήσεις επιβίωσης είναι οι ίδιες για δυο ομάδες ατόμων, μπορούν να ελεγχθούν με το μη παραμετρικό logrank τεστ (Landau and Everitt, 2004):

$$U_L = \sum_{j=1}^r (d_j - e_j) \quad (3)$$

όπου U_L είναι το άθροισμα των διαφορών (ή αποστάσεων) μεταξύ του αριθμού των παρατηρηθέντων d και του αριθμού των αναμενόμενων γεγονότων (θανάτων) e στις χρονικές στιγμές (έστω r) που έχει συμβεί ένα (τουλάχιστον) γεγονός (θάνατος). Υψώνοντας στο τετράγωνο και διαιρώντας το U_L με την εκτιμώμενη διακύμανση υπολογίζουμε το στατιστικό χ^2 του logrank τεστ. Με το stratified logrank test, όπου η ελεγχοσυνάρτηση U_L υπολογίζεται κατά παράγοντα, μπορούμε να εξετάσουμε αν διάφορες μεταβλητές σχετίζονται με τους χρόνους επιβίωσης. Στην παρούσα έρευνα, ως παράγοντας χρησιμοποιήθηκε η μεταβλητή maxΔd_cat.

Όπως φαίνεται στους μαθηματικούς τύπους που χρησιμοποιούνται στην ανάλυση επιβίωσης, ο χρόνος που συμβαίνει ένα γεγονός είναι η καθοριστική συνιστώσα. Συνεπώς, το μεγαλύτερο εμπόδιο που αναφέρεται στην εφαρμογή της μεθόδου στις δασικές απογραφές είναι η έλλειψη μετρήσεων ηλικιών των δέντρων και ακριβών χρόνων θνησιμότητας (Flewelling and Monserud, 2002). Ωστόσο, η γνώση της ηλικίας δεν είναι απαραίτητη για την εφαρμογή της ανάλυσης επιβίωσης. Οποιαδήποτε μέτρηση που δείχνει αλλαγές στην κατάσταση των δέντρων σε επόμενο χρόνο, μπορεί να αντικαταστήσει την παραδοσιακή ανάλυση επιβίωσης με μετρήσεις ηλικιών (Harcombe, 1987). Σε αυτή την εργασία, σε όρους διαμέτρου, το χαρακτηριστικό που εξετάζεται στην ανάλυση επιβίωσης είναι η μέγιστη μείωση στη διάμετρο από το ένα διαδοχικό ύψος i προς το άλλο, δηλαδή η maxΔd. Αν για ένα δέντρο η μέγιστη μείωση της διαμέτρου είναι μεγαλύτερη από το μέσο όρο μέγιστης μείωσης διαμέτρου, θεωρήσαμε ότι η μείωση της διαμέτρου είναι ραγδαία, άρα το δέντρο έχει κακή τάση αύξησης και στη μεταβλητή status παίρνει τιμή μηδέν (αντίστοιχα με το γεγονός του θανάτου σε ιατρικά δεδομένα). Αντίθετα, αν για ένα δέντρο η μέγιστη μείωση της διαμέτρου είναι μικρότερη από το μέσο όρο μέγιστης μείωσης διαμέτρου, θεωρήσαμε ότι η μείωση της διαμέτρου είναι λιγότερο ραγδαία, άρα το δέντρο έχει καλύτερη τάση αύξησης και στη μεταβλητή status παίρνει τιμή ένα (αντίστοιχα με την επιβίωση σε ιατρικά δεδομένα). Για όλους, λοιπόν, τους προηγούμενους μαθηματικούς τύπους, αντικαθιστούμε το χρόνο με την maxΔd.

Η ανάλυση επιβίωσης μπορεί να γίνει με χρήση διάφορων στατιστικών πακέτων H/Y. Σε αυτή τη μελέτη χρησιμοποιήθηκε η ανάλυση SURVIVAL του SPSS (Landau and Everitt, 2004).

3. ΑΠΟΤΕΛΕΣΜΑΤΑ

Το πλήθος των δέντρων ανά κατηγορία maxΔd_cat και ο αριθμός των δέντρων με ραγδαία μείωση διαμέτρου (status = 0) δίνονται στον Πίνακα 2.

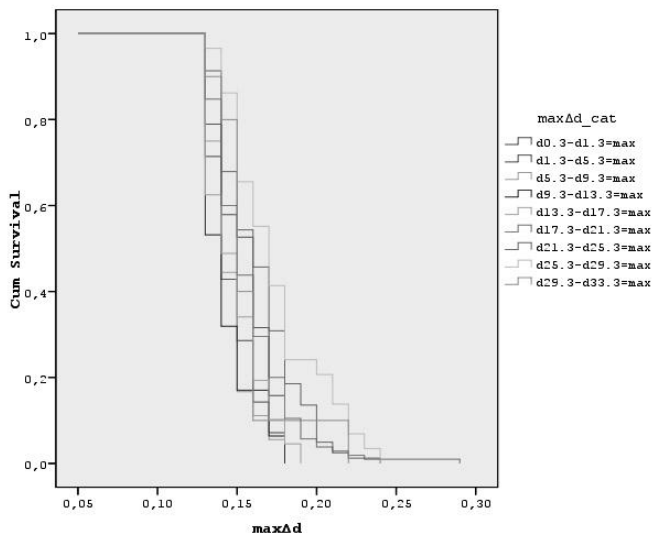
Πίνακας 2. Πλήθος δέντρων ανά κατηγορία μέγιστης μείωσης διαμέτρου

maxΔd_cat	Πλήθος δέντρων	Αποκομμένες παρατηρήσεις (status = 1)	
		Πλήθος	Ποσοστό %
d0.3-d1.3=max	41	22	53,7%
d1.3-d5.3=max	114	100	87,7%
d5.3-d9.3=max	214	178	83,2%
d9.3-d13.3=max	194	147	75,8%

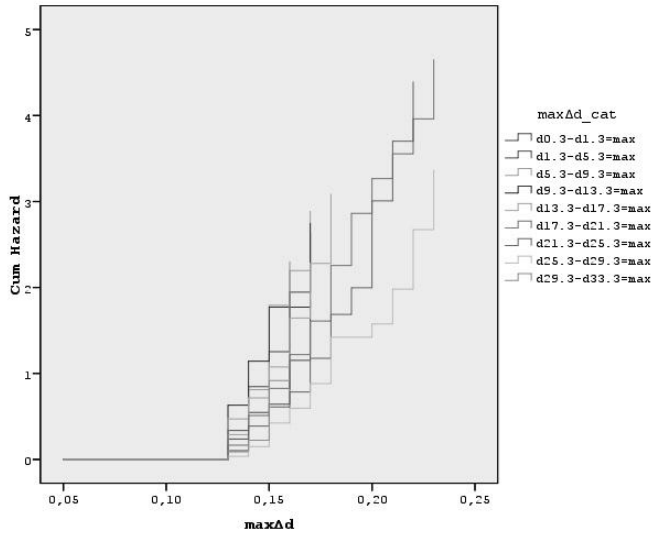
d13.3-d17.3=max	193	105	54,4%
d17.3-d21.3=max	159	54	34,0%
d21.3-d25.3=max	104	23	22,1%
d25.3-d29.3=max	29	0	0,0%
d29.3-d33.3=max	13	3	23,1%
Σύνολο	1061	632	59,6%

Οι συναρτήσεις επιβίωσης και στιγμιαίου κινδύνου εκτιμήθηκαν ξεχωριστά για τις εννέα τιμές της maxΔd_cat (εικόνες 1 και 2). Θα μπορούσαμε να πούμε πως στην κατηγορία d25.3 – d29.3 εμφανίζεται λιγότερο ραγδαία μείωση της διαμέτρου (εικόνα 1), ενώ αυξημένος κίνδυνος ραγδαίας μείωσης διαμέτρου σημειώνεται στην κατηγορία d9.3 – d13.3 (εικόνα 2).

Εικόνα 1. Kaplan–Meier συναρτήσεις επιβίωσης για τις εννέα κατηγορίες μέγιστης μείωσης διαμέτρου



Εικόνα 2. Συναρτήσεις στιγμιαίου κινδύνου για τις εννέα κατηγορίες μέγιστης μείωσης διαμέτρου



Το τεστ logrank που έγινε για τον έλεγχο της επίδρασης του ύψους στη μέγιστη μείωση διαμέτρου (Πίνακας 3), έδειξε στατιστικά σημαντική επίδραση ($p < 0,05$).

Πίνακας 3. Logrank τεστ για την ομοιογένεια των κατανομών μέγιστης μείωσης διαμέτρου ανάμεσα στα επίπεδα των κατηγοριών maxΔd

Παράγοντας	χ^2	Βαθμοί ελευθερίας	Τιμή p
maxΔd_cat	61,447	8	0,000

4. ΣΥΖΗΤΗΣΗ - ΣΥΜΠΕΡΑΣΜΑΤΑ

Ως διαμήκης (longitudinal) μπορεί να οριστεί κάθε μεταβλητή που μετράει τη μεταβολή από μια κατάσταση σε μια άλλη (Collett, 1994). Η μεγαλύτερη δυσκολία στην εφαρμογή της ανάλυσης επιβίωσης στις δασικές απογραφές είναι να βρεθούν κατάλληλες διαμήκεις μεταβλητές για την ποσοτικοποίηση πιθανοτήτων επιβίωσης και διακινδύνευσης θνησιμότητας (Zens and Peart, 2003). Αν ως διαμήκεις μεταβλητές χρησιμοποιηθούν ο χρόνος ανόδου ή η ηλικία, ενδεχόμενα να προκύψουν διάφορα προβλήματα, όπως αναφέρονται στη βιβλιογραφία σχετικά με τη δασική ανάλυση επιβίωσης (Harcombe, 1987, Burgman et al., 1994, Preisler and Slaughter, 1997 και Volney, 1998). Πρώτα απ' όλα, οι περισσότερες παρατηρήσεις είναι αποκομμένες. Ο ακριβής χρόνος θανάτου του δέντρου είναι αβέβαιος, με τις επαναλήψεις των μετρήσεων σε μια επόμενη δασική απογραφή να χρησιμοποιούνται συχνά ως διαμήκη δεδομένα. Κατά δεύτερο, η συνάρτηση επιβίωσης εξαρτάται μερικά από το πότε και πού γίνονται οι μετρήσεις. Για παράδειγμα, αν τα περισσότερα νεκρά δέντρα εντοπίζονται σε μια συγκεκριμένη περιοχή της χώρας,

όπου γίνεται η απογραφή σε συγκεκριμένη χρονική στιγμή, τότε η εκτίμηση της επιβίωσης θα είναι μεροληπτική, αν ο χρόνος χρησιμοποιηθεί ως μεταβλητή. Κατά τρίτο, η ηλικία του δέντρου είναι δύσκολο να εκτιμηθεί σε ευρείας κλίμακας δασικές απογραφές. Σε προηγούμενες μελέτες, σε ομήλικες συστάδες, μικρής σχετικά έκτασης, καταρτίστηκαν ευέλικτα μοντέλα επιβίωσης / στιγμιαίου κινδύνου, με χρήση της ηλικίας του δέντρου και του χρόνου ανόδου (Burgman et al., 1994, Preisler and Slaughter, 1997 και Volney, 1998).

Υπάρχουν διάφορες μέθοδοι εκτίμησης συναρτήσεων επιβίωσης και στιγμιαίου κινδύνου, καθεμιά από τις οποίες έχει πλεονεκτήματα και μειονεκτήματα. Σε αυτή τη μελέτη χρησιμοποιήθηκε η μέθοδος εκτίμησης Kaplan–Meier, μια τεχνική που προτείνουν οι Harcombe (1987) και Woodall et al. (2005), για μια αρχική αξιολόγηση της αύξησης των δέντρων. Θα μπορούσαμε να πούμε πως στην κατηγορία d25.3 – d29.3 εμφανίζεται λιγότερο ραγδαία μείωση της διαμέτρου, ενώ αυξημένος κίνδυνος ραγδαίας μείωσης διαμέτρου σημειώνεται στην κατηγορία d9.3 – d13.3.

Όσον αφορά στο logrank τεστ, αυτό επιτρέπει τον έλεγχο υποθέσεων σχετικά με τη σύγκριση των καμπυλών επιβίωσης και στιγμιαίου κινδύνου. Σε αυτή την εργασία το logrank τεστ έδειξε πως το ύψος του δέντρου επηρεάζει σημαντικά το ρυθμό μείωσης της διαμέτρου. Η δυνατότητα να συσχετιστούν τα μεμονωμένα γνωρίσματα των δέντρων με τον κίνδυνο κακής αύξησης έχει μεγάλη πιθανά οφέλη για τη μελέτη των φυσικών διεργασιών στα δασικά οικοσυστήματα (Zens and Peart, 2003).

ABSTRACT

Tree growth has traditionally been assessed in forest inventories through summaries height and diameter measurements. This study proposes the application of survival analysis for the purpose of analyzing tree growth. Individual tree taper measurements were used to estimate survival and hazard functions for the fir trees (*Abies borisii-regis*) of the University Forest of Pertouli (Northern Greece). These estimates provided indications of high degree of taper above 25 m of total tree height. Survival analysis technique, by using tree taper, could help foresters to assess regional tree growth trends.

ΑΝΑΦΟΡΕΣ

- Amateis R., Burkhart H. and Walsh T. (1997). Modeling survival in juvenile and mature loblolly pine plantations. *Forest Ecology and Management*, **13**, 170–174.
- Berkson J. and Gage R. (1950). Calculation of survival rates of cancer. Proceedings of Staff Meetings of the Mayo Clinic, vol. 25, 270–286.
- Bigler C. and Bugmann H. (2004a). Assessing the performance of theoretical and empirical tree mortality models using tree–ring series of Norway spruce. *Ecological Modelling*, **174**, 225–239.
- Bigler C. and Bugmann H. (2004b). Predicting the time of tree death using dendrochronological data. *Ecological Applications*, **14**, 902–914.

- Buchman R., Pederson S. and Walters N. (1983). A tree survival model with application to species of the Great Lakes region. *Canadian Journal of Forest Research*, **13**, 601–608.
- Burgman M., Incoll W., Ades P., Ferguson I., Fletcher T. and Wohlers A. (1994). Mortality models for mountain and alpine ash. *Forest Ecology and Management*, **67**, 319–327.
- Collett D. (1994). *Modelling Survival Data in Medical Research*, New York: Chapman and Hall/CRC.
- Cox D. and Oakes D. (1984). *Analysis of Survival Data*, London: Chapman and Hall.
- Daniels R. and Burkhart H. (1975). Simulation of Individual Tree Growth and Development In Managed Loblolly Plantations. Virginia Polytechnic Institute and State University, Blacksburg Publication FWS–5–75.
- Flewelling J. and Monserud R. (2002). Comparing methods for modeling tree mortality. In: N.L. Crookston and R.N. Havis, Editors, Second Forest Vegetation Simulator Conference. RMRS–P–25 USDA Forest Service, Rocky Mountain Research Station, Ft. Collins, CO, 169–177.
- Hamilton D. and Edwards B. (1976). Modeling the Probability of Individual Tree Mortality. Research Paper INT–185. USDA Forest Service, Intermountain Forest and Range Experiment Station, Ogden, UT.
- Harcombe P. (1987). Tree life tables. *BioScience* **37**, 557–568.
- Hawkes C. (2000). Woody plant mortality algorithms: description, problems, and progress. *Ecological Modelling*, **126**(2-3), 225–248.
- Kitikidou K. and Apostolopoulou E. (2011). Applying survival analysis for assessment of forests sustainable development. *Renewable & Sustainable Energy Reviews*, **15**, 851-855.
- Landau S. and Everitt B. (2004). *A handbook of statistical analyses using SPSS*, New York: Chapman and Hall/CRC.
- Monserud R. (1976). Simulation of forest tree mortality. *Forest Science*, **22**, 438–444.
- Morse B. and Kulman H. (1984). Plantation white spruce mortality: estimates based on aerial photography and analysis using a life–table format. *Canadian Journal of Forest Research*, **14**, 195–200.
- Muenchow G. (1986). Ecological use of failure time analysis. *Ecology*, **67**, 246–250.
- Preisler H. and Slaughter G. (1997). Stochastic model for tree survival in stands affected by annosum root disease. *Forest Science*, **43**, 78–86.
- Reams G., Brann T. and Halteman W. (1988). A nonparametric survival model for balsam fir during a spruce budworm outbreak. *Canadian Journal of Forest Research*, **18**, 787–793.
- Rose C. (2002). Modeling and allocating forestry survival: a loblolly pine case study. Univ. Georgia. Dissertation.
- Stage A. (1973). Prognosis model for stand development. Research Paper INT–137. USDA Forest Service, Intermountain Forest and Range Experiment Station, Ogden, UT.
- Vahntsevanos K., Kyrgidis A., Verrou E., Katodritou E., Triaridis S., Andreadis C., Boukovinas I., Koloutsos G., Teleioudis Z., Kitikidou K., Paraskevopoulos P.,

- Zervas K. and Antoniadis K. (2009). Longitudinal Cohort Study of Risk Factors in Cancer Patients of Bisphosphonate-Related Osteonecrosis of the Jaw (Treatment-Related Complications). *Journal of Clinical Oncology*, **27**(32), 5356–5362.
- Volney W. (1998). Ten-year tree mortality following a jack pine budworm outbreak in Saskatchewan. *Canadian Journal of Forest Research*, **28**, 1784–1793.
- Waters W. (1969). Life-table approach to analysis of insect impact. *Journal of Forestry*, **67**, 300–304.
- Woodall C., Grambsch P. and Thomas W. (2005). Applying survival analysis to a large-scale forest inventory for assessment of tree mortality in Minnesota. *Ecological Modelling*, **189**, 199–208.
- Wyckoff P. and Clark J. (2000). Predicting tree mortality from diameter growth: a comparison of maximum likelihood and Bayesian approaches. *Canadian Journal of Forest Research*, **30**, 156–167.
- Zens M. and Peart D. (2003). Dealing with death data: individual hazard, mortality, and bias. *Trends in Ecology Evolution*, **18**, 366–373.

Η ΔΕΙΓΜΑΤΙΚΗ ΔΙΑΣΠΟΡΑ ΚΑΙ ΤΟ ΤΙΜΗΜΑ ΤΗΣ ΑΜΕΡΟΛΗΨΙΑΣ ΤΗΣ

Σταύρος Κουρούκλης

Τμήμα Μαθηματικών, Πανεπιστήμιο Πατρών
stavros@math.upatras.gr

ΠΕΡΙΛΗΨΗ

Είναι ευρέως γνωστό ότι η δειγματική διασπορά $S^2 = (n-1)^{-1} \sum_{i=1}^n (X_i - \bar{X})^2$ είναι αμερόληπτος εκτιμητής της διασποράς, σ^2 , μιας κατανομής όταν οι παρατηρήσεις X_1, \dots, X_n , $n \geq 2$, αποτελούν τυχαίο δείγμα από αυτήν την κατανομή. Χρησιμοποιώντας την τεχνική της τεχνητής αύξησης δείγματος, ο Yatracos (2005) κατασκεύασε τον εκτιμητή $S_2 = (n+2)(n(n+1))^{-1} \sum_{i=1}^n (X_i - \bar{X})^2$ και απέδειξε ότι έχει μικρότερο μέσο τετραγωνικό σφάλμα (ΜΤΣ) από τον S^2 για κάθε κατανομή (με πεπερασμένη τέταρτη ροπή) και για κάθε $n \geq 2$. Στην εργασία αυτή, κατασκευάζεται ο εκτιμητής $S_1 = n(n(n-1)+2)^{-1} \sum_{i=1}^n (X_i - \bar{X})^2$ και αποδεικνύεται ότι έχει τις εξής ιδιότητες: (α) ΜΤΣ μικρότερο από το ΜΤΣ του S_2 . (β) Είναι αποδεκτός στην κλάση των εκτιμητών $C = \{c \sum_{i=1}^n (X_i - \bar{X})^2 : c > 0\}$. Η τεχνική κατασκευής του S_1 βασίζεται στην κλασική ιδέα του Stein (1964), είναι πολύ απλή και μπορεί να διδαχθεί ακόμη και σε προπτυχιακό επίπεδο. Τέλος, σε αντίθεση με τη δειγματική διασπορά, αποδεικνύεται ότι ο εκτιμητής μεθόδου ροπών $S_0^2 = n^{-1} \sum_{i=1}^n (X_i - \bar{X})^2$ είναι αποδεκτός στην κλάση C .

1. ΕΙΣΑΓΩΓΗ

Έστω X_1, \dots, X_n , $n \geq 2$, τυχαίο δείγμα από την κατανομή $F \in \mathcal{F} = \{F : \text{η } F \text{ έχει πεπερασμένη τέταρτηροπή και δεν στηρίζεται σε ένα μόνο σημείο}\}$.

Δύο συνήθεις εκτιμητές της διασποράς, σ^2 , της F είναι η δειγματική διασπορά $S^2 = (n-1)^{-1} \sum_{i=1}^n (X_i - \bar{X})^2$, η οποία ως γνωστόν είναι αμερόληπτη, και ο εκτιμητής μεθόδου ροπών $S_0^2 = n^{-1} \sum_{i=1}^n (X_i - \bar{X})^2$. Στην εργασία αυτή θα ασχοληθούμε κυρίως με τον S^2 . Στην περίπτωση που η F είναι η κανονική κατανομή με άγνωστη μέση τιμή, ο S^2 είναι, επί πλέον, ο αμερόληπτος εκτιμητής ελάχιστης διασποράς του σ^2 . Παρά τη βέλτιστη αυτή ιδιότητά του, είναι επίσης γνωστό ότι ο εκτιμητής $(n+1)^{-1} \sum_{i=1}^n (X_i - \bar{X})^2$ έχει μικρότερο ΜΤΣ από τον S^2 , και επομένως ο S^2 είναι

μη αποδεκτός στην περίπτωση της κανονικής κατανομής. Για αυτήν την περίπτωση, υπάρχει στη βιβλιογραφία πληθώρα καλύτερων εκτιμητών από τον S^2 , βλ. Stein (1964), Brown (1968), Strawderman (1974), Brewster and Zidek (1974), Maruyama (1998), Maruyama and Strawderman (2006). Όμως, ο Yatracos (2005), χρησιμοποιώντας την τεχνική της τεχνητής αύξησης δείγματος, κατασκεύασε τον εκτιμητή $S_2 = c_2 S^2$, $c_2 = (n+2)(n-1)(n(n+1))^{-1}$ και απέδειξε ότι έχει μικρότερο ΜΤΣ από τον S^2 , για κάθε $F \in \mathbb{F}$ και κάθε $n \geq 2$. Το αποτέλεσμα αυτό (εκτός από αναπάντεχο) είναι και εξαιρετικά ενδιαφέρον γιατί αποδεικνύει για πρώτη φορά στη βιβλιογραφία την καθολική μη αποδεκτικότητα του S^2 , δηλαδή ανεξάρτητα της υποκείμενης κατανομής.

Στην εργασία αυτή, το πρόβλημα εκτίμησης του σ^2 προσεγγίζεται από την πλευρά της Στατιστικής Θεωρίας Αποφάσεων με συνάρτηση ζημίας το τετραγωνικό σφάλμα. Χρησιμοποιώντας την ιδέα του Stein (1964), κατασκευάζεται ο εκτιμητής $S_1 = c_1 S^2$, $c_1 = n(n-1)(n(n-1)+2)^{-1}$ και αποδεικνύεται ότι έχει μικρότερο ΜΤΣ από κάθε εκτιμητή cS^2 , $c > c_1$, για κάθε $F \in \mathbb{F}$ και κάθε $n \geq 2$. Ειδικά, επειδή $1 > c_1$ και $c_2 > c_1$, ο S_1 είναι καλύτερος από τον S^2 αλλά και από τον S_2 . Επί πλέον, αποδεικνύεται ότι ο S_1 δεν μπορεί να βελτιωθεί από εκτιμητή της μορφής cS^2 , $c > 0$, δηλαδή είναι αποδεκτός στην κλάση $C = \{cS^2 : c > 0\}$. Ακολουθώς, χαρακτηρίζονται οι εκτιμητές της κλάσης C που βελτιώνουν τον S^2 για κάθε $F \in \mathbb{F}$ και δεδομένο (αλλά αυθαίρετο) n . Τέλος, συγκρίνονται οι εκτιμητές S^2 και S_0^2 (με κριτήριο το ΜΤΣ) και αποδεικνύεται η αποδεκτικότητα του S_0^2 στην κλάση C . Υπενθυμίζεται ότι ένας εκτιμητής λέγεται αποδεκτός σε μια κλάση εκτιμητών αν δεν υπάρχει κάποιος εκτιμητής στην κλάση αυτή με μικρότερο ή ίσο ΜΤΣ για κάθε $F \in \mathbb{F}$ και γνησίως μικρότερο ΜΤΣ για τουλάχιστον ένα $F \in \mathbb{F}$.

2. ΜΗ ΑΠΟΔΕΚΤΙΚΟΤΗΤΑ ΤΗΣ ΔΕΙΓΜΑΤΙΚΗΣ ΔΙΑΣΠΟΡΑΣ

Αρχικά, αναζητούμε το στοιχείο της κλάσης C , αν υπάρχει, με το μικρότερο ΜΤΣ. Η συνάρτηση $E(cS^2 - \sigma^2)^2$ ελαχιστοποιείται ως προς c για την τιμή $c_F = ES^2 / ES^4$. Λαμβάνοντας υπ' όψη τις σχέσεις

$$ES^2 = \sigma^2, \quad V = VarS^2 = \frac{\mu_4 - \sigma^4}{n} + \frac{2\sigma^4}{n(n-1)}, \quad (1)$$

όπου $\mu_4 = E(X_1 - \mu)^4$ και $\mu = EX_1$, βλ. Serfling (1980, p.185), προκύπτει ότι

$$c_F = \frac{n(n-1)}{n(n-1) + 2 + (n-1)(a_4 - 1)}, \quad (2)$$

όπου $a_4 = \mu_4 / \sigma^4$ είναι η κύρτωση της F . Επειδή, γενικά, η κύρτωση a_4 εξαρτάται από άγνωστες παραμέτρους της F , συμπεραίνουμε ότι δεν υπάρχει βέλτιστο στοιχείο στην κλάση C . Ακολουθώντας την ιδέα του Stein (1964), θα αναζητήσουμε (άνω) φράγμα του c_F μη εξαρτώμενο από την F . Αυτό είναι πολύ εύκολο να βρεθεί ως εξής. Θέτουμε $Z = (X_1 - \mu)^2$ οπότε $\mu_4 = EZ^2$ και $\sigma^4 = (EZ)^2$. Τώρα, επειδή $EY^2 \geq (EY)^2$ για κάθε τυχαία μεταβλητή Y , έχουμε $\mu_4 \geq \sigma^4$, δηλαδή $a_4 \geq 1$ και επομένως

$$c_F \leq \frac{n(n-1)}{n(n-1)+2} = c_1. \quad (3)$$

Μάλιστα, το άνω φράγμα c_1 είναι ακριβές αφού η ισότητα $c_F = c_1$ επιτυγχάνεται όταν η Z είναι σταθερά (με πιθανότητα 1).

Ορίζουμε τώρα τον εκτιμητή του σ^2 , $S_1 = c_1 S^2$.

Πρόταση 1. Για κάθε $F \in \mathbb{F}$ και κάθε $n \geq 2$ ισχύουν τα εξής.

- (α) Το ΜΤΣ του S_1 είναι μικρότερο από το ΜΤΣ του cS^2 , $\forall c > c_1$.
- (β) Το ΜΤΣ του S_1 είναι μικρότερο από το ΜΤΣ του S^2 .
- (γ) Το ΜΤΣ του S_1 είναι μικρότερο από το ΜΤΣ του S_2 .

Απόδειξη. Η συνάρτηση $E(cS^2 - \sigma^2)^2$ είναι γνησίως αύξουσα για $c > c_F$. Άρα, η πρόταση προκύπτει αμέσως από το γεγονός ότι $c_F \leq c_1 < c_2 < 1$.

Παρατήρηση 1. Έστω $\mathbb{F}_0 = \{F \in \mathbb{F} : a_4 \text{ είναι σταθερά μη εξαρτώμενη από άγνωστες παραμέτρους}\}$. Τότε, προφανώς, από την (2), ο βέλτιστος εκτιμητής στην κλάση C για $F \in \mathbb{F}_0$ είναι ο $c_F S^2$. Το αποτέλεσμα αυτό έχει δειχθεί προηγουμένως από τους Searls and Intarapanich (1990). Στην κλάση \mathbb{F}_0 ανήκουν π.χ. οι κατανομές $N(\mu, \sigma^2)$ ($a_4 = 3$), $G(\alpha, \beta)$ με γνωστό α ($a_4 = 3 + 6/\alpha$), $U(0, \theta)$ και $U(-\theta, \theta)$ ($a_4 = 1.8$).

Το ερώτημα που θέτουμε στη συνέχεια είναι κατά πόσον υπάρχει εκτιμητής στην κλάση C καλύτερος από τον S_1 . Η απάντηση στο ερώτημα είναι αρνητική όπως αποδεικνύεται παρακάτω.

Πρόταση 2. Δεν υπάρχουν $n \geq 2$ και $c > 0$, $c \neq c_1$, τέτοια ώστε ο εκτιμητής cS^2 να έχει μικρότερο ή ίσο ΜΤΣ από το ΜΤΣ του S_1 για κάθε $F \in \mathbb{F}$.

Απόδειξη. Έστω ότι υπάρχουν. Τότε, κατ' ανάγκη, $0 < c < c_1$ (αφού $E(cS^2 - \sigma^2)^2$ είναι γνησίως αύξουσα για $c > c_F$, άρα και για $c > c_1$) και $E(cS^2 - \sigma^2)^2 \leq E(c_1S^2 - \sigma^2)^2$ για κάθε $F \in \mathbb{F}$. Ισοδύναμα, έχουμε

$$c^2 \text{Var}S^2 + (1-c)^2 \sigma^4 \leq c_1^2 \text{Var}S^2 + (1-c_1)^2 \sigma^4$$

ή (επειδή $0 < c < c_1$)

$$\frac{\text{Var}S^2}{\sigma^4} \geq \frac{2-c_1-c}{c_1+c} \quad \text{για κάθε } F \in \mathbb{F}.$$

Επομένως, συμπεραίνουμε ότι

$$\inf_{F \in \mathbb{F}} \frac{\text{Var}S^2}{\sigma^4} \geq \frac{2-c_1-c}{c_1+c}.$$

Από την (1), είναι εύκολο να δειχθεί ότι

$$\inf_{F \in \mathbb{F}} \frac{\text{Var}S^2}{\sigma^4} = \frac{2}{n(n-1)}$$

και συνεπώς $2/(n(n-1)) \geq (2-c_1-c)/(c_1+c)$. Αντικαθιστώντας το c_1 από την (3), η τελευταία σχέση δίνει $c \geq c_1$, το οποίο είναι άτοπο.

Από την απόδειξη της Πρότασης 1 προκύπτει ότι κάθε εκτιμητής cS^2 με $c_1 \leq c < 1$ είναι καλύτερος από τον S^2 με κριτήριο το ΜΤΣ. Υπάρχουν άλλοι; Στο ερώτημα δίνει απάντηση η επόμενη πρόταση που παραθέτει όλους τους εκτιμητές cS^2 που είναι καλύτεροι από τον S^2 .

Πρόταση 3. Ο εκτιμητής cS^2 έχει μικρότερο ΜΤΣ από τον S^2 για κάθε $F \in \mathbb{F}$ αν και μόνον αν $c_3 = (n(n-1)-2)/(n(n-1)+2) \leq c < 1$, εκτός από τον c_3S^2 που έχει μικρότερο ΜΤΣ όταν $a_4 > 1$ και ίσο ΜΤΣ με τον S^2 όταν $a_4 = 1$.

Απόδειξη. Ότι $c < 1$ είναι προφανές. Επί πλέον, ισχύει $E(cS^2 - \sigma^2)^2 < E(S^2 - \sigma^2)^2$ αν και μόνον αν $c^2 \text{Var}S^2 + (1-c)^2 \sigma^4 < \text{Var}S^2$ ή ισοδύναμα $(1-c)/(1+c) < \text{Var}S^2 / \sigma^4$. Το αποτέλεσμα, τώρα, προκύπτει από την τελευταία σχέση επειδή $\inf_{F \in \mathbb{F}} \text{Var}S^2 / \sigma^4 = 2/(n(n-1))$ και η τιμή $2/(n(n-1))$ επιτυγχάνεται μόνον αν $a_4 = 1$.

Παρατήρηση 2. Για $n = 2$, η Πρόταση 3 συνεπάγεται ότι ο cS^2 είναι καλύτερος του S^2 , $\forall 0 \leq c < 1$. Ειδικά, ο εκτιμητής 0 (με ΜΤΣ = σ^4) είναι καλύτερος του S^2 (με

ΜΤΣ = $(\mu_4 - \sigma^4)/2 + \sigma^4$!!! Από την άλλη πλευρά, για $n = 2$, ο S_1 συμπίπτει με τον εκτιμητή μεθόδου ροπών S_0^2 .

Κλείνοντας αυτήν την ενότητα, αναφέρουμε ότι η μη αποδεκτικότητα του αμερόληπτου εκτιμητή του σ^2 μπορεί να αποδειχθεί και στο γενικότερο μοντέλο της Ανάλυσης Διασποράς κατά ένα παράγοντα με ανεξάρτητα και ισόνομα σφάλματα. Επίσης, η Πρόταση 1 μπορεί να επεκταθεί στην πολυμεταβλητή ανάλυση για την εκτίμηση του πίνακα συνδιασποράς Σ ως προς τη συνάρτηση ζημίας $\text{tr}(\delta - \Sigma)^2$.

3. Ο ΕΚΤΙΜΗΤΗΣ ΜΕΘΟΔΟΥ ΡΟΠΩΝ

Η πρόταση που ακολουθεί αποδεικνύει ότι, σε αντίθεση με τον S^2 , ο εκτιμητής μεθόδου ροπών $S_0^2 = n^{-1} \sum_{i=1}^n (X_i - \bar{X})^2$ είναι αποδεκτός στην κλάση C .

Πρόταση 4. Δεν υπάρχουν $n \geq 2$ και $c > 0$, $c \neq (n-1)/n$, τέτοια ώστε ο εκτιμητής cS^2 να έχει μικρότερο ή ίσο ΜΤΣ από εκείνο του S_0^2 για κάθε $F \in \mathcal{F}$.

Απόδειξη. Θέτουμε $S_0^2 = c_0 S^2$, $c_0 = (n-1)/n$ και υποθέτουμε ότι υπάρχουν $n \geq 2$ και $c > 0$, $c \neq c_0$, τέτοια ώστε $E(cS^2 - \sigma^2)^2 \leq E(c_0 S^2 - \sigma^2)^2$, για κάθε $F \in \mathcal{F}$. Τότε έχουμε

$$c^2 \text{Var} S^2 + (1-c)^2 \sigma^4 \leq c_0^2 \text{Var} S^2 + (1-c_0)^2 \sigma^4$$

ή

$$(c^2 - c_0^2) \text{Var} S^2 \leq [(1-c_0)^2 - (1-c)^2] \sigma^4. \quad (4)$$

Υπάρχουν δύο περιπτώσεις:

I. $c < c_0$. Τότε η (4) απλοποιείται και δίνει

$$\frac{2-c_0-c}{c_0+c} \leq \frac{\text{Var} S^2}{\sigma^4} \quad \text{για κάθε } F \in \mathcal{F},$$

οπότε

$$\frac{2-c_0-c}{c_0+c} \leq \inf_{F \in \mathcal{F}} \frac{\text{Var} S^2}{\sigma^4} = \frac{2}{n(n-1)}, \quad (5)$$

όπως αναφέρθηκε στην απόδειξη της Πρότασης 2. Όμως, η συνάρτηση $(2-c_0-c)/(c_0+c)$ είναι (γνησίως) φθίνουσα και επομένως η σχέση $c < c_0$ συνεπάγεται $(2-c_0-c)/(c_0+c) > (2-2c_0)/(2c_0) = 1/(n-1) \geq 2/(n(n-1))$, $\forall n \geq 2$, που αντίκειται στην (5).

II. $c > c_0$. Τότε η (4) απλοποιείται και δίνει

$$\frac{2-c_0-c}{c_0+c} \geq \frac{VarS^2}{\sigma^4} \quad \text{για κάθε } F \in \mathbb{F}. \quad (6)$$

Όμως, από την (1) είναι εύκολο να δειχθεί ότι $\sup_{F \in \mathbb{F}} VarS^2 / \sigma^4 = \infty$, που αντίκειται στην (6).

Ολοκληρώνοντας αυτή την ενότητα, θα συγκρίνουμε τους εκτιμητές S^2 και S_0^2 ως προς το ΜΤΣ.

Πρόταση 5. (α) Ο S_0^2 έχει ΜΤΣ μικρότερο από το ΜΤΣ του S^2 αν και μόνον αν $a_4 > (3n^2 - 8n + 3)/(2n^2 - 3n + 1)$.

(β) Για $n = 2, 3, 4$, ο S_0^2 είναι καλύτερος από τον S^2 για κάθε $F \in \mathbb{F}$.

(γ) Ο S_0^2 είναι καλύτερος από τον S^2 για κάθε $F \in \mathbb{F}$ με $a_4 > 3/2$.

Απόδειξη. (α) Προκύπτει αμέσως από τη σχέση $E(cS^2 - \sigma^2)^2 = c^2 VarS^2 + (1-c)^2 \sigma^4$ και την (1).

(β) Για $n = 2, 3, 4$, έχουμε $(3n^2 - 8n + 3)/(2n^2 - 3n + 1) < 1$ ενώ $a_4 \geq 1$.

(γ) Ισχύει $(3n^2 - 8n + 3)/(2n^2 - 3n + 1) < 3/2$.

Παρατήρηση 3. Η κανονική κατανομή έχει $a_4 = 3$, άρα ο S_0^2 είναι καλύτερος από τον S^2 (το οποίο μπορεί, φυσικά, να δειχθεί και με κατ' ευθείαν υπολογισμούς).

ABSTRACT

Yatracos (2005) constructed an estimator of the variance, σ^2 , which has smaller mean squared error (mse) than the sample variance S^2 (i.e., the unbiased estimator) for all distributions with finite fourth moment and all sample sizes. In this work, another estimator of σ^2 is constructed with the following properties: (a) it has smaller mse than Yatracos' (2005) estimator and (b) it is admissible in the class $C = \{cS^2 : c > 0\}$. The method of construction is based on Stein's (1964) classical idea, is very simple and may be taught even in an undergraduate class. Also, for a given (but arbitrary) sample size, all the estimators in C with smaller mse than S^2 are found. Finally, in contrast to S^2 , it is shown that the method of moments estimator of σ^2 is admissible in C .

ΑΝΑΦΟΡΕΣ

- Brewster J.F, Zidek J.V. (1974). Improving on equivariant estimators. *Ann. Statist.*, **2**, 21-38.
- Brown L. (1968). Inadmissibility of the usual estimators of scale parameters in problems with unknown location and scale parameters. *Ann. Math. Statist.*, **39**, 29-48.
- Maruyama Y. (1998). Minimax estimators of a normal variance. *Metrika*, **48**, 209-214.
- Maruyama Y., Strawderman W. (2006). A new class of minimax generalized Bayes estimators of a normal variance. *J. Statist. Plan. Infer.*, **136**, 3822-3836.
- Searls D., Intarapanich P. (1990). A note on an estimator for the variance that utilizes the kurtosis. *The American Statistician*, 295-296.
- Serfling R.J. (1980). *Approximation theorems of mathematical statistics*. Wiley.
- Stein C. (1964). Inadmissibility of the usual estimator for the variance of a normal distribution with unknown mean. *Ann. Inst. Statist. Math.*, **16**, 155-160.
- Strawderman W. (1974). Minimax estimation of powers of the variance of a normal population under squared error loss. *Ann. Statist.*, vol.2, no. 1, 190-198.
- Yatracos Y. (2005). Artificially augmented samples, shrinkage, and mean squared error reduction. *J. Amer. Statist. Assoc.*, **100**, 1168-1175.



ΜΕΤΡΑ ΣΥΝΑΦΕΙΑΣ ΚΑΙ ΜΕΤΡΑ ΣΥΜΜΕΤΡΙΑΣ – ΑΣΥΜΜΕΤΡΙΑΣ ΓΙΑ ΠΙΝΑΚΕΣ ΣΥΝΑΦΕΙΑΣ

Γ. Κουτσοχέρας, Τ. Παπαιωάννου

Τμήμα Στατιστικής & Ασφαλιστικής Επιστήμης, Πανεπιστήμιο Πειραιώς
georgios.koutsocheras@gmail.com, takrap@unipi.gr

ΠΕΡΙΛΗΨΗ

Η παρούσα εργασία αποτελεί μια συγκριτική μελέτη των μέτρων συνάφειας για ονοματικές και διατακτικές μεταβλητές. Παρουσιάζεται επίσης και μια αριθμητική διερεύνηση των μέτρων συμμετρίας – ασυμμετρίας για τετραγωνικούς πίνακες. Συγκεκριμένα, εξετάζονται 21 μέτρα συνάφειας και 23 μέτρα συμμετρίας – ασυμμετρίας. Τα τελευταία μέτρα περιορίζονται σε 11 από 23, χρησιμοποιώντας γενικευμένα μέτρα πληροφορίας από την οικογένεια των μέτρων *power divergence*. Τα προτεινόμενα μέτρα δεν ακολουθούν μια αξιωματική θεωρία, αλλά βασίζονται στις λειτουργικές τους ιδιότητες. Παρόλαυτα, όλα ακολουθούν κάποιες βασικές παραδοχές. Το μέγεθος του δείγματος και η διάσταση του πίνακα, φαίνεται να παίζουν ρόλο στην επιλογή του μέτρου για τη σύγκριση πινάκων. Ειδικότερα, σκοπός της εργασίας είναι η σύγκριση των διαφόρων μέτρων με βάση τις ιδιότητές τους και η παρουσίαση αποτελεσμάτων αριθμητικών διερευνήσεων. Στο πλαίσιο αυτό, εξετάζονται τα πλεονεκτήματα και τα μειονεκτήματα των μέτρων, φιλοδοξώντας να διατυπωθεί μια ξεκάθαρη ερμηνεία και να διευκολυνθεί η διαδικασία επιλογής του κατάλληλου μέτρου για κάθε περίπτωση.

Λέξεις κλειδιά: Μέτρα συνάφειας, Εντροπία, Power divergence, f – divergence, Συμμετρία, Περιθώρια Ομοιογένεια, Ψευδοσυμμετρία.

1. ΕΙΣΑΓΩΓΗ

Πολλά μέτρα συνάφειας για πίνακες συνάφειας έχουν προταθεί στη βιβλιογραφία, διανύοντας μια διαδρομή τριών αιώνων. Η συνάφεια είναι έννοια συνώνυμη της εξάρτησης. Τα μέτρα συμμετρίας – ασυμμετρίας αναφέρονται σε τετραγωνικούς πίνακες. Τα μέτρα αυτά είναι γνωστά και ως μέτρα απομάκρυνσης από το μοντέλο που θέλει να διερευνήσει ο ερευνητής. Το πρόβλημα που αντιμετωπίζουν συνήθως οι ερευνητές και οι χρήστες είναι η επιλογή του καταλληλότερου μέτρου για κάθε περίπτωση, καθώς τα μέτρα συνήθως παίρνουν διαφορετικές τιμές για το ίδιο σύνολο δεδομένων. Οι δυο κατηγορίες των μέτρων αυτών ταξινομούνται σε μέτρα για ονοματικές και μέτρα για διατακτικές μεταβλητές.

Για τις ονοματικές μεταβλητές υπάρχουν τα εξής μέτρα συνάφειας:

- (i) Μέτρα που βασίζονται στην ελεγχοσυνάρτηση χ^2 - Pearson (*Yule ϕ , Tschuprow T , Cramer V , Pearson C , Sakoda C_{adj}*).
- (ii) Μέτρα αναλογικής μείωσης του σφάλματος πρόβλεψης της εξαρτημένης μεταβλητής από τη γνώση της ανεξάρτητης (*Goodman-Kruskal $\lambda_a, \lambda_b, \lambda, \tau_a, \tau_b$*). Τα μέτρα αυτά, αντανακλούν την ποσοστιαία μείωση του σφάλματος από την πρόβλεψη της εξαρτημένης μεταβλητής, δοθείσης της ανεξάρτητης.
- (iii) Μέτρα που βασίζονται στην εντροπία και την αμοιβαία πληροφορία (*Theil $U_{Y|X}$, $U_{Y|X}, U_{sym}$*).

Για τους ορισμούς των μέτρων βλέπε μεταξύ άλλων *Liebetrau* (1983), *Agresti* (2002) και *Κουτσοχέρας* (2010, σελ. 50-68).

Στην περίπτωση των διατακτικών μεταβλητών, τα μέτρα συνάφειας αξιοποιούν την επιπρόσθετη πληροφορία της διάταξης των μεταβλητών, ταξινομώντας τα ζεύγη των παρατηρήσεων σε «σύμφωνα» (*concordant*) και σε «ασύμφωνα» (*discordant*) ζεύγη καθώς και σε ζεύγη με ισοτιμίες (*ties*). Υπάρχουν τα εξής μέτρα συνάφειας:

- (i) Μέτρα που δεν λαμβάνουν υπόψη τις ισοτιμίες (*Goodman – Kruskal γ , Kendall τ_a, τ_c*).
- (ii) Μέτρα που λαμβάνουν υπόψη τις ισοτιμίες (*Kendall τ_b , Somer $D_{Y|X}, D_{X|Y}$*).
- (iii) Μέτρα που βασίζονται στις τάξεις (*rank.s*) των μεταβλητών (*Spearman r_s*).

Για τους ορισμούς των μέτρων βλέπε μεταξύ άλλων *Goodman-Kruskal* (1954), *Sheskin* (2004) και *Κουτσοχέρας* (2010, σελ. 80-88).

Τα μέτρα συμμετρίας – ασυμμετρίας βασίζονται στα μέτρα απόστασης ή απόκλισης (*divergence*), τα οποία προέρχονται από τον χώρο της θεωρίας της πληροφορίας (*Information Theory*). Στην ουσία αξιολογούν την ομοιότητα – ανομοιότητα μεταξύ δυο (διακριτών) κατανομών πιθανότητας **a** και **b**, όπου **a** συνήθως η πραγματική (θεωρητική) και **b** η παρατηρούμενη κατανομή, με την έννοια ότι όσο μικρότερη η τιμή του μέτρου, τόσο πιο δύσκολος ο διαχωρισμός των κατανομών στις οποίες αναφέρεται. Τα μέτρα συμμετρίας – ασυμμετρίας διακρίνονται ανάλογα με: (i) το είδος του μοντέλου συμμετρίας που εξετάζουν και με (ii) το είδος του μέτρου απόκλισης που χρησιμοποιούν.

Βασιζόμενοι στον πρώτο διαχωρισμό έχουμε τα ακόλουθα μοντέλα: *Symmetry (S)*, *Quasi – Symmetry (QS)*, *Marginal Homogeneity (MH)*, *Global Symmetry (GS)*, *Diagonal Symmetry (DS)*, *Conditional Symmetry (CS)*. Για περισσότερες λεπτομέρειες βλέπε *Κουτσοχέρας* (2010, σελ. 96, 138).

Βασιζόμενοι στον δεύτερο διαχωρισμό, έχουμε μέτρα που χρησιμοποιούν για παράδειγμα, την *Kullback-Leibler (KL) divergence*, *Pearsonian distance*, *Gauss discrepancy*, *Shannon entropy*, *Gini concentration*, *Cressie & Read (CR) power divergence*, *Patil & Taillie (PT) diversity index* [βλέπε *Pardo*, (2006)].

Για δεδομένο μοντέλο συμμετρίας ή ασυμμετρίας, τα μέτρα κατασκευάζονται χρησιμοποιώντας είτε την τυποποιημένη απόκλιση *KL*, είτε την απόσταση *Pearson* αντίστοιχα, μεταξύ της παρατηρηθείσης κατανομής και της κατανομής του μοντέλου (η οποία πρέπει να εκτιμάται από τα δεδομένα). Ένας τρίτος τρόπος κατασκευής μέτρων απομάκρυνσης από το μοντέλο, είναι η χρήση του σταθμισμένου μέσου όρου ή δεσμευμένης μέσης τιμής της εντροπίας του *Shannon* ή της *Ευκλείδειας* απόστασης

(*Gauss discrepancy*) για όλα τα μη-διαγώνια ζεύγη κελιών του πίνακα [βλέπε *Tomizawa* (1994), *Tomizawa et al.* (2005)]. Τα μέτρα θα συμβολίζονται με φ , αν στηρίζονται στην απόσταση KL και με ψ , αν στηρίζονται στην απόσταση χ^2 - *Pearson*, όπου (\cdot) θα υποδεικνύει το εκάστοτε υπό εξέταση μοντέλο. Αν αντί της απόκλισης KL ή χ^2 - *Pearson*, χρησιμοποιηθεί η απόκλιση δύναμης (*power divergence*) των CR ή η ποικιλότητα PT , τότε προκύπτουν γενικευμένα μέτρα τα οποία θα συμβολίζονται με $\Phi^{(\lambda)}$, όπου λ η παράμετρος του μέτρου. Περαιτέρω, για την κατασκευή μέτρων συμμετρίας – ασυμμετρίας οι *Kateri & Papaioannou* (2007), πρότειναν τη χρήση της απόκλισης *Csiszar* γνωστής και ως f - *divergence*. Η απόκλιση αυτή δίνει μια ευρύτερη κλάση μέτρων απόκλισης η οποία εμπεριέχει τα μέτρα του *Tomizawa*. Σε κάθε περίπτωση πρέπει να προσδιορίζεται η κυρτή συνάρτηση f .

Δεν υπάρχει ένα σύνολο ιδιοτήτων ή αξιωμάτων τα οποία πρέπει να τηρεί ένα μέτρο συνάφειας ή συμμετρίας – ασυμμετρίας. Συνήθως ένα μέτρο θα πρέπει να κυμαίνεται στο $[0,1]$ ή $[-1,1]$, όπου η τιμή 0 σημαίνει πλήρη έλλειψη συνάφειας και για τις τιμές -1 & 1 πλήρη συνάφεια, στην περίπτωση των μέτρων συνάφειας, ενώ η τιμή 0 σημαίνει ότι τα δεδομένα έχουν τη δομή του υπό εξέταση μοντέλου και η τιμή 1 σημαίνει πλήρη απομάκρυνση από το μοντέλο, στην περίπτωση των μέτρων συμμετρίας – ασυμμετρίας.

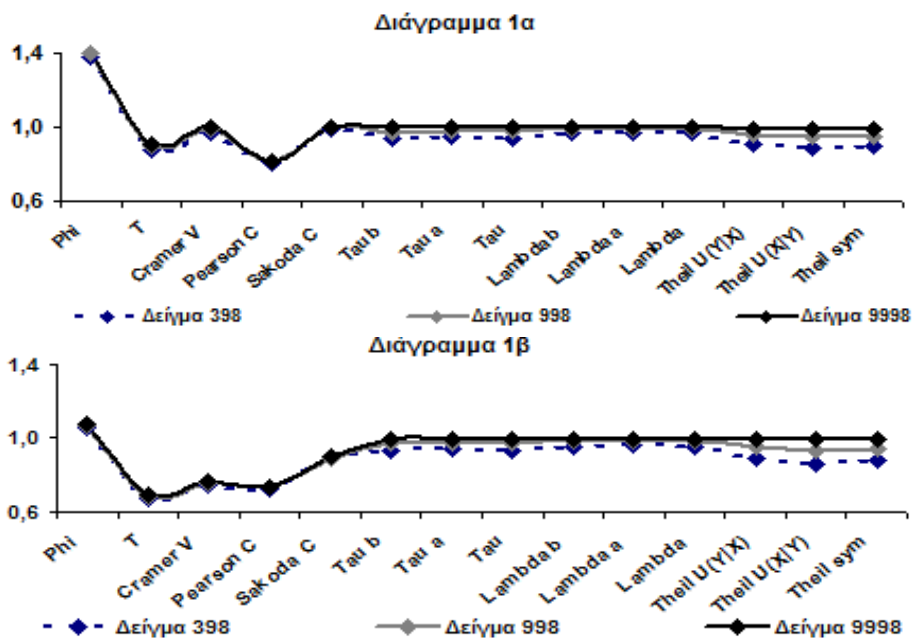
2. ΜΕΤΡΑ ΣΥΝΑΦΕΙΑΣ ΓΙΑ ΟΝΟΜΑΤΙΚΕΣ ΜΕΤΑΒΛΗΤΕΣ

Στην ενότητα αυτή παρουσιάζεται μια αριθμητική διερεύνηση των 14 μέτρων συνάφειας για ονοματικές μεταβλητές που αναφέρονται στην εισαγωγή, μαζί με την ερμηνεία τους και σχετικές παρατηρήσεις.

Υπολογίζονται τα 14 μέτρα της κατηγορίας αυτής, σε πίνακα συνάφειας που αναφέρεται στη μελέτη του χρώματος ματιών και μαλλιών [*Goodman-Kruskal*, (1954)]. Προέκυψαν 14 διαφορετικές τιμές για το βαθμό συνάφειας, κυμαινόμενες από $U_{XY} = 0.075$ μέχρι $C_{adj} = 0.452$, μια αξιόλογη διαφορά. Για λεπτομέρειες βλέπε *Κουτσοχέρας* (2010, σελ. 73).

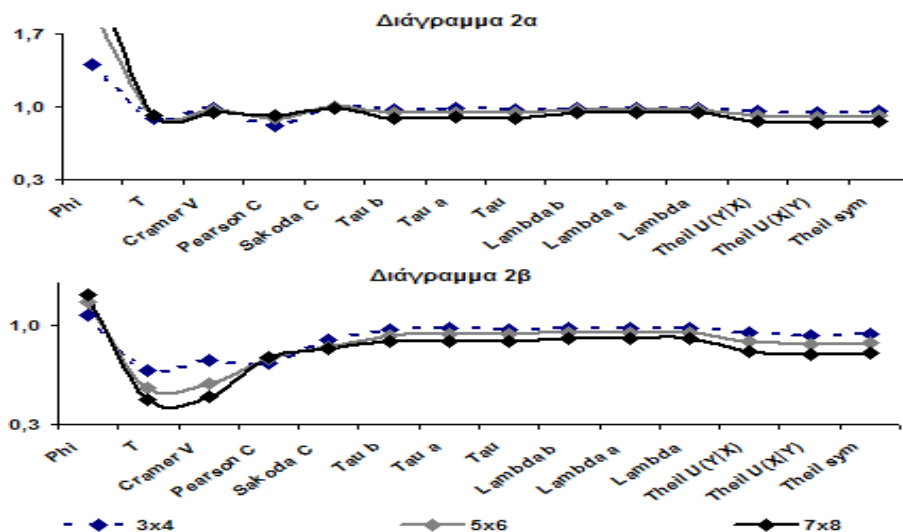
Στη συνέχεια εξετάζεται η συμπεριφορά των μέτρων σε συνθήκες πλήρους συνάφειας 2 μορφών, θεωρώντας είτε ότι όλες οι παρατηρήσεις συγκεντρώνονται (α) στα κελιά της κυρίας διαγωνίου είτε (β) σε 2 οποιαδήποτε κελιά του πίνακα, που ανήκουν σε διαφορετική γραμμή και στήλη. Η ανάλυση εφαρμόζεται σε έναν 3×4 πίνακα για 3 διαφορετικά μεγέθη δείγματος ($n = 398, 998, 9998$). Τα αποτελέσματα δίνονται στα Διαγράμματα 1α και 1β αντίστοιχα. Η τιμή 1 δεν «πιάνεται» από όλα τα μέτρα.

Διάγραμμα 1. Μέτρα συνάφειας σε συνθήκες πλήρους συνάφειας 2 μορφών (α), (β)



Στη συνέχεια και στο ίδιο πλαίσιο συγκρίνονται πίνακες διαφορετικής διάστασης, δηλαδή 3×4 , 5×6 , 7×8 . Τα αποτελέσματα δίνονται στο Διάγραμμα 2α με όλες τις παρατηρήσεις συγκεντρωμένες στα κελιά της κυρίας διαγωνίου και στο Διάγραμμα 2β με όλες τις παρατηρήσεις συγκεντρωμένες σε δυο κελιά του πίνακα. Το μέγεθος του δείγματος είναι $n = 998$.

Διάγραμμα 2. Μέτρα συνάφειας σε πίνακες διαφορετικής διάστασης

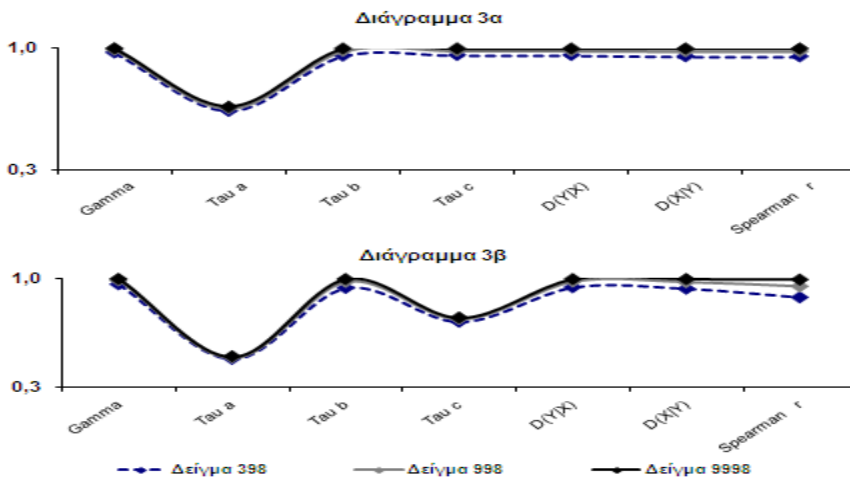


Από την αριθμητική διερεύνηση προκύπτει ότι τα μέτρα συνάφειας για ονοματικές μεταβλητές δεν είναι κατάλληλα για σύγκριση πινάκων με διαφορετική διάσταση. Κάποια μέτρα υποεκτιμούν ή υπερεκτιμούν το βαθμό συνάφειας, ενώ κάποια χρησιμοποιούν λιγότερη πληροφορία. Τέλος, κάποια από τα μέτρα επηρεάζονται από το μέγεθος του δείγματος.

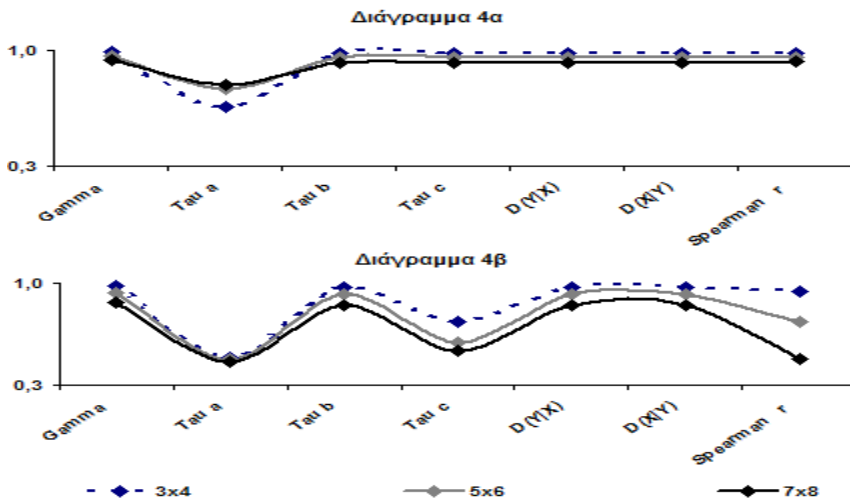
3. ΜΕΤΡΑ ΣΥΝΑΦΕΙΑΣ ΓΙΑ ΔΙΑΤΑΚΤΙΚΕΣ ΜΕΤΑΒΛΗΤΕΣ

Στην ενότητα αυτή παρουσιάζεται μια αριθμητική διερεύνηση των 7 μέτρων συνάφειας για διατακτικές μεταβλητές που αναφέρονται στην εισαγωγή, μαζί με την ερμηνεία τους και σχετικές παρατηρήσεις. Η αριθμητική διερεύνηση των μέτρων συνάφειας για διατακτικές μεταβλητές, ακολουθεί τα βήματα της ενότητας 2 και είναι στο ίδιο πλαίσιο. Θεωρώντας τα δεδομένα του πίνακα συνάφειας που αναφέρεται στη σύγκριση του βάρους νεογέννητων και της σειράς γέννησης [Sheskin, (2004)], οι τιμές των μέτρων κυμαίνονται από $\tau_a = 0.383$ μέχρι $\gamma = 0.699$, πάλι μια αξιόλογη διαφορά. Για λεπτομέρειες βλέπε Κουτσοχέρας (2010, σελ. 91). Τα Διαγράμματα 3 και 4 είναι αυτονόητα και ακολουθούν τα Διαγράμματα 1 και 2.

Διάγραμμα 3. Μέτρα συνάφειας σε συνθήκες πλήρους συνάφειας 2 μορφών (α), (β)



Διάγραμμα 4. Μέτρα συνάφειας σε πίνακες διαφορετικής διάστασης



Από την αριθμητική διερεύνηση προκύπτει ότι τα μέτρα συνάφειας για διατακτικές μεταβλητές δεν είναι κατάλληλα για σύγκριση πινάκων με διαφορετική διάσταση. Κάποια μέτρα ακόμη και σε συνθήκες πλήρους συνάφειας δεν πετυχαίνουν την τιμή 1. Επίσης, κάποια από τα μέτρα επηρεάζονται από το μέγεθος του δείγματος. Γενικά, όταν στα πρωτογενή δεδομένα υπάρχουν πολλές ισοτιμίες (*ties*), είναι απαραίτητο να χρησιμοποιούνται μέτρα που λαμβάνουν υπόψη τις ισοτιμίες.

4. ΜΕΤΡΑ ΣΥΜΜΕΤΡΙΑΣ - ΑΣΥΜΜΕΤΡΙΑΣ

Στην ενότητα αυτή παρουσιάζεται μια αριθμητική διερεύνηση 23 μέτρων συμμετρίας – ασυμμετρίας για ονοματικές και διατακτικές μεταβλητές, μαζί με την ερμηνεία τους και σχετικές παρατηρήσεις. Για τον συμβολισμό των μοντέλων βλέπε Agresti (2002). Συνοπτικά τα μοντέλα συμμετρίας – ασυμμετρίας έχουν ως εξής [βλέπε Κουτσοχέρας (2010, σελ. 96, 138)]:

Μοντέλα για Ονοματικές μεταβλητές:

$$(S): \pi_{ij} = \pi_{ji}, \text{ για } i \neq j \quad (MH): \pi_{i.} = \pi_{.i}, \text{ για } i \neq j$$

$$(QS): \pi_{ij} = \rho_{ij} a_i \beta_j, \text{ όπου } \rho_{ij} = \rho_{ji} \text{ για } i \neq j \quad \text{ή} \quad (QS): \pi_{ij} = 2a_i \pi_{ij}^S / (a_i + a_j)$$

$$\text{ή} \quad (QS): D_{ijk} = D_{kji}, \text{ όπου } D_{ijk} = \pi_{ij} \pi_{jk} \pi_{ki}, D_{kji} = \pi_{kj} \pi_{ji} \pi_{ik}, \text{ για } i < j < k$$

$$\text{ή} \quad (QS): Q_{ijk} = Q_{kji}, \text{ όπου } Q_{ijk} = \pi_{ij}^C \pi_{jk}^C \pi_{ki}^C, Q_{kji} = \pi_{kj}^C \pi_{ji}^C \pi_{ik}^C, \text{ για } i < j < k,$$

$$\text{όπου } \pi_{ij}^C = \pi_{ij} / (\pi_{ij} + \pi_{ji}) \quad \text{ή σε όρους γειτονικών odds ratio} \quad (QS): \theta_{(i < j, j < k)} = \theta_{(j < k, i < j)},$$

$$\text{όπου } \theta_{(i < j, j < k)} = (\pi_{ij} \pi_{jk}) / (\pi_{jj} \pi_{ik}), \text{ για } i < j < k \quad \text{ή αλλιώς} \quad (QS): \theta_{ij} = \theta_{ji},$$

$$\text{όπου } \theta_{ij} = (\pi_{ij} \pi_{i+1, j+1}) / (\pi_{i, j+1} \pi_{i+1, j}), \text{ για } i, j = 1, \dots, R-1.$$

Μοντέλα για Διατακτικές μεταβλητές:

$$(GS): \delta^U = \delta^L \text{ όπου } \delta^U = \sum_{i < j} \pi_{ij}, \delta^L = \sum_{i > j} \pi_{ij}$$

$$(CS): \pi_{ij} = \delta \pi_{ji}, \delta \in R^+, \text{ για } i > j \quad (DS): \pi_{ij} = \delta_{i-j} \pi_{ji}, \delta \in (0, 2), \text{ για } i > j$$

$$(MH): G_{1(i)} = G_{2(i)} \text{ όπου } G_{1(i)} = \sum_{s=1}^i \sum_{t=i+1}^R \pi_{st}, G_{2(i)} = \sum_{s=i+1}^R \sum_{t=1}^i \pi_{st}$$

Υπολογίζονται 16 μέτρα για ονοματικές μεταβλητές σε πίνακα συνάφειας που αναφέρεται στη σύγκριση της Δανέζικης κοινής γνώμης στο δημοψήφισμα για την εισαγωγή της χώρας στην ΕΥ [Tomizawa, (1994)] και 7 μέτρα για διατακτικές μεταβλητές, σε πίνακα συνάφειας που αναφέρεται στη σύγκριση του επαγγέλματος πατέρα – υιού [Tomizawa et al. (2005)]. Για όλα τα μοντέλα, (i) τα μέτρα φ που βασίζονται στην απόκλιση *KL* και στη σταθμισμένη εντροπία *Shannon* ταυτίζονται, (ii) ομοίως και τα μέτρα ψ που βασίζονται στην απόκλιση *Pearson* και στη σταθμισμένη απόσταση *Gauss* ή *Gini*. Τα μέτρα φ και ψ πετυχαίνουν κοντινές τιμές για το ίδιο σύνολο δεδομένων, για παράδειγμα από 0.035 έως 0.048, αλλά δεν ταυτίζονται. Για περισσότερες λεπτομέρειες βλέπε Κουτσοχέρας (2010, σελ. 161-162).

Στη συνέχεια εξετάζονται κάποια χαρακτηριστικά ακραία παραδείγματα τετραγωνικών πινάκων για να ερμηνεύσουμε τα μοντέλα και τη συμπεριφορά των μέτρων. Συγκεκριμένα, στο Διάγραμμα 5 εξετάζονται μοντέλα για ονοματικές μεταβλητές για τις εξής περιπτώσεις:

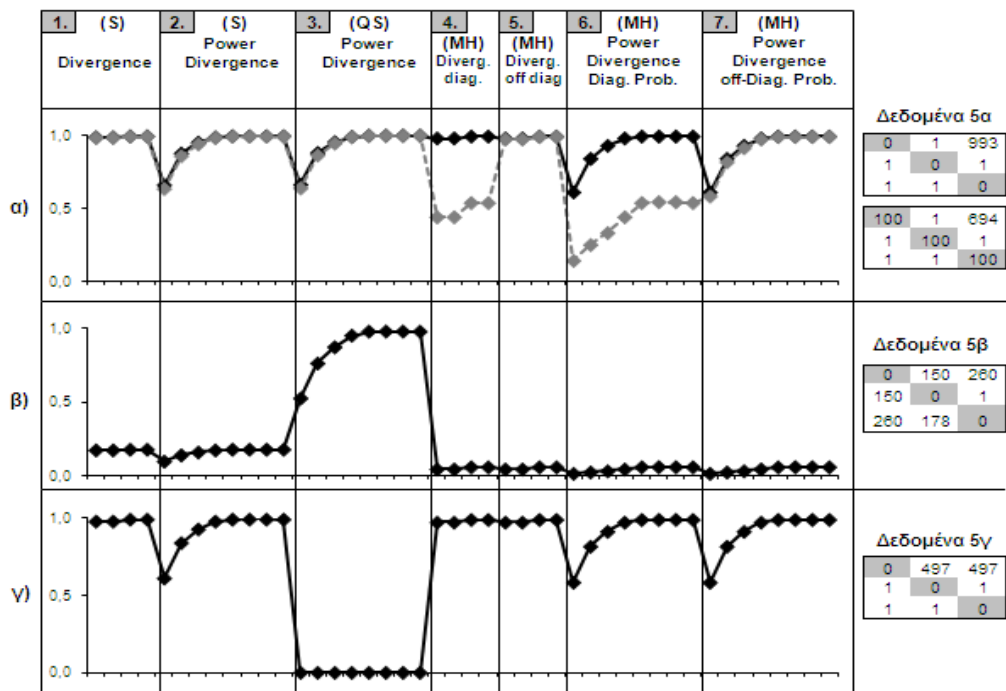
(α) δεδομένα που δεν ακολουθούν (*S*), (*QS*), (*MH*).

(β) δεδομένα που δεν ακολουθούν (*QS*) αλλά σχεδόν ακολουθούν (*S*), (*MH*).

(γ) δεδομένα που δεν ακολουθούν (*S*), (*MH*) και ακολουθούν (*QS*).

Η ανάλυση εφαρμόζεται σε έναν 3×3 πίνακα με μέγεθος δείγματος $n = 998$ και για τις ακόλουθες τιμές της παραμέτρου $\lambda = -0.8, -0.6, -0.4, 0, 1, 1.4, 1.6, 2$. Τα αποτελέσματα της ανάλυσης αυτής παρατίθενται στα διαγράμματα 5α, 5β και 5γ, αντίστοιχα.

Διάγραμμα 5. Μέτρα συμμετρίας-ασυμμετρίας για ονοματικές μεταβλητές



1. δυο μέτρα $\varphi_{(S)}$ και δυο μέτρα $\psi_{(S)}$ 2. ένα μέτρο CR, $\Phi_{(S)}^{(\lambda)}$ 3. ένα μέτρο CR, $\Phi_{(QS)}^{(\lambda)}$ 4. δυο μέτρα $\varphi_{(MH)}$ και δυο μέτρα $\psi_{(MH)}$ που μοντελοποιούν τις διαγώνιες πιθανότητες 5. δυο μέτρα $\varphi_{(MH)}$ και δυο μέτρα $\psi_{(MH)}$ που δεν μοντελοποιούν τις διαγώνιες πιθανότητες 6. ένα μέτρο CR, $\Phi_{(MH)}^{(\lambda)}$ που μοντελοποιεί τις διαγώνιες πιθανότητες 7. ένα μέτρο CR, $\Phi_{(MH)}^{(\lambda)}$ που δεν μοντελοποιεί τις διαγώνιες πιθανότητες. Τα λ παίρνουν 8 τιμές.

Στο Διάγραμμα 6, εξετάζονται μοντέλα για διατακτικές μεταβλητές για τις εξής περιπτώσεις:

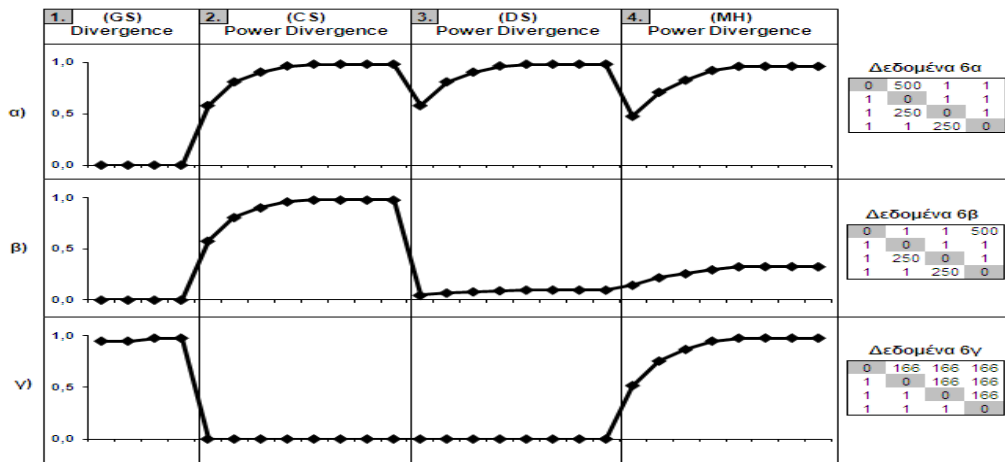
(α) δεδομένα που δεν ακολουθούν (CS), (DS), (MH) και ακολουθούν (GS).

(β) δεδομένα που δεν ακολουθούν (CS) και ακολουθούν (GS), (DS), (MH).

(γ) δεδομένα που δεν ακολουθούν (GS), (MH) και ακολουθούν (CS), (DS).

Η ανάλυση εφαρμόζεται σε έναν 4×4 πίνακα με μέγεθος δείγματος $n = 998$ και για τις ακόλουθες τιμές της παραμέτρου $\lambda = -0.8, -0.6, -0.4, 0, 1, 1.4, 1.6, 2$. Τα αποτελέσματα της ανάλυσης αυτής παρατίθενται στα διαγράμματα 6α, 6β και 6γ, αντίστοιχα.

Διάγραμμα 6. Μέτρα συμμετρίας-ασυμμετρίας για διατακτικές μεταβλητές



1. δυο μέτρα $\phi_{(GS)}$ και δυο μέτρα $\psi_{(GS)}$ 2. ένα μέτρο $CR, \Phi_{(CS)}^{(\lambda)}$ 3. ένα μέτρο $CR, \Phi_{(DS)}^{(\lambda)}$ 4. ένα μέτρο $CR, \Phi_{(MH)}^{(\lambda)}$. Τα λ παίρνουν 8 τιμές.

Τα μέτρα εκφράζουν τον βαθμό απομάκρυνσης της παρατηρηθείσας κατανομής πιθανότητας από τη θεωρητική κατανομή του μοντέλου που εξετάζουν. Όλα τα μέτρα μοντελοποιούν τις πιθανότητες των μη-διαγώνιων κελιών. Όμως για το μοντέλο της (MH), έχουν προταθεί και μέτρα που μοντελοποιούν και τις πιθανότητες της κυρίας διαγώνιου. Από το Διάγραμμα 5α, στήλες 4, 6, διαπιστώνεται ότι τα μέτρα αυτά ίσως να μην είναι τόσο κατάλληλα καθώς υποεκτιμούν τον βαθμό απομάκρυνσης από τη (MH), όταν παρατηρήσεις συγκεντρώνονται στη διαγώνιο. Στην περίπτωση διατακτικών μεταβλητών τα μέτρα μοντελοποιούν τις πιθανότητες του άνω και κάτω τριγώνου συνολικά, έτσι ώστε να μην παραμένουν αμετάβλητα σε αναδιατάξεις γραμμών και στηλών. Παρατηρούμε ότι για $\lambda < 0$ τα μέτρα $\Phi^{(\lambda)}$ υποεκτιμούν τον βαθμό απομάκρυνσης από το υπό εξέταση μοντέλο. Σημειώνεται ότι για $\lambda = 0$, $\Phi^{(0)} = \phi$ και για $\lambda = 1$, $\Phi^{(1)} = \psi$, όπου (\cdot) το εκάστοτε υπό εξέταση μοντέλο. Κανένα μέτρο δεν επηρεάζεται από το μέγεθος του δείγματος για πίνακα της ίδιας μορφής ή δομής (αποτελέσματα δεν δίνονται λόγω χώρου) και επομένως είναι όλα κατάλληλα για τη σύγκριση πινάκων με διαφορετικό μέγεθος δείγματος.

5. ΣΥΜΠΕΡΑΣΜΑΤΑ

Στην παρούσα εργασία εξετάστηκαν 21 μέτρα συνάφειας για ονοματικές και διατακτικές μεταβλητές. Τα μέτρα συνάφειας δεν είναι κατάλληλα για τη σύγκριση πινάκων με διαφορετική διάσταση, ενώ κάποια από αυτά επηρεάζονται από το μέγεθος τους δείγματος. Για τον υπολογισμό της συνάφειας σε ένα δοσμένο πίνακα, προτείνουμε τον υπολογισμό, ως καταλληλότερων και χρησιμότερων, των μέτρων Goodman – Kruskal (τ_a, τ_b, τ), Cramer V, Sakoda C_{adj} , Theil U στην περίπτωση ονοματικών μεταβλητών και των μέτρων Kendall τ_b και Somer D στην περίπτωση

διατακτικών μεταβλητών. Στη συνέχεια εξετάστηκαν 23 μέτρα συμμετρίας – ασυμμετρίας. Η ερμηνεία κάποιων μοντέλων δεν είναι άμεση και απλή. Υπολογίζοντας τα μέτρα σε επιλεγμένους πίνακες, ερμηνεύσαμε καλύτερα τα μοντέλα. Διαπιστώθηκε ότι χρησιμοποιώντας τα μέτρα απόκλισης δύναμης (*power divergence*) CR ή τον δείκτη ποικιλότητας PT , τα 23 προτεινόμενα μέτρα μπορούν να περιοριστούν σε 11. Η επιλογή της καταλληλότερης τιμής για την παράμετρο λ παραμένει ένα ανοικτό ερώτημα. Όλα τα μέτρα είναι κατάλληλα για τη σύγκριση πινάκων με διαφορετικό μέγεθος δείγματος καθώς δεν επηρεάζονται από αυτό.

ABSTRACT

This paper is a comparative study of measures of association for contingency tables with nominal or ordinal classification variables. Also, a numerical investigation of measures of symmetry – asymmetry for square contingency tables is presented. Specifically, 21 measures of association and 23 measures of symmetry – asymmetry are examined. The latter can be reduced to 11 from 23, by applying power divergence type measures. The development of all measures is primarily operational, in the sense that the various measures are established more on account of their properties rather than on account of an axiomatic theory. However, all measures are based on some fundamental assumptions. The sample size and the dimension of the table seem to play a role on measure selection, when comparing several tables. The aim of the paper is to compare the various measures in terms of their properties and to present the results of numerical investigations. In that sense, our ambition is to provide a clear interpretation of each measure, facilitating the process of selecting the most appropriate measure per case.

ΑΝΑΦΟΡΕΣ

- Agresti A. (2002). *Categorical Data Analysis*, New York: John Wiley.
- Goodman L. and Kruskal W. (1954). Measures of Association for Cross Classifications. *Journal of the American Statistical Association*, **49**, 732-764.
- Kateri M. and Papaioannou T. (2007). Measures of Symmetry-Asymmetry for Square Contingency Tables. *Technical Report*, **4**, University of Piraeus.
- Κουτσοχέρας Γ. (2010). Μέτρα συνάφειας και μέτρα συμμετρίας-ασυμμετρίας για πίνακες συνάφειας. *Διπλωματική Εργασία*, Πανεπιστήμιο Πειραιώς.
- Liebetrau A. (1983). *Measures of Association*, SAGE Publications Inc.
- Pardo L. (2006). *Statistical Inference Based on Divergence Measures*, Chapman & Hall.
- Sheskin D. (2004). *Handbook of Parametric and Nonparametric Statistical Procedures*, Chapman & Hall: 3rd ed.
- Tomizawa S. (1994). Two kinds of Measures of Departure from Symmetry in Square Contingency Tables having Nominal Categories. *Statistica Sinica*, **4**, 325-334.
- Tomizawa S., Miyamoto N. and Yamane S. (2005). Power Divergence type Measure of Departure from Diagonals-Parameter Symmetry for Square Contingency Tables with Ordered Categories. *Statistics*, **39**, 107-115.

ΣΥΜΒΟΛΟΣΕΙΡΕΣ ΦΡΑΓΜΕΝΟΥ ΜΗΚΟΥΣ ΣΕ ΔΥΑΔΙΚΕΣ ΑΚΟΛΟΥΘΙΕΣ

Φ. Σ. Μακρή¹, Ζ. Μ. Ψυλλάκης²

¹Τμήμα Μαθηματικών, Πανεπιστήμιο Πατρών
makri@math.upatras.gr

²Τμήμα Φυσικής, Πανεπιστήμιο Πατρών
psillaki@physics.upatras.gr

ΠΕΡΙΛΗΨΗ

Σε μία ακολουθία τυχαίων μεταβλητών με δύο δυνατά αποτελέσματα (Επιτυχία-Αποτυχία) ορίζουμε την τυχαία μεταβλητή που απαριθμεί τις συμβολοσειρές οι οποίες σχηματίζονται από ροές αποτυχιών φραγμένου μήκους μεταξύ δύο διαδοχικών επιτυχιών. Θεωρώντας ακολουθίες ανεξάρτητων και ισόνομων δυαδικών τυχαίων μεταβλητών προσδιορίζουμε τη συνάρτηση πιθανότητας της τόσο μέσω συνδυαστικής ανάλυσης όσο και μέσω αναδρομικών σχέσεων. Μέσω αναδρομικών σχέσεων δίνουμε επίσης την πιθανογεννήτρια και τη ροπογεννήτρια συνάρτηση καθώς και τις παραγοντικές ροπές και τις ροπές περί το μηδέν.

Λέξεις Κλειδιά: Πειράματα Bernoulli, Ροές, Συμβολοσειρές.

1. ΕΙΣΑΓΩΓΗ ΚΑΙ ΠΡΟΚΑΤΑΡΚΤΙΚΕΣ ΕΝΝΟΙΕΣ

Θεωρούμε μια ακολουθία X_1, X_2, \dots, X_n , $n \geq 2$, από δυαδικές [Επιτυχία (S ή 1)-Αποτυχία (F ή 0)] τυχαίες μεταβλητές (τ.μ.) διατεταγμένες σε μια γραμμή. Δοθέντων δύο ακεραίων αριθμών k και ℓ , με $0 \leq k \leq \ell \leq n - 2$, ορίζουμε την τ.μ. $M_{n;k,\ell}$ που απαριθμεί τις συμβολοσειρές (σχηματισμούς) $S \underbrace{FF \dots F}_{\geq k, \leq \ell} S$ που σχηματίζονται

από ροές αποτυχιών, με μήκος τουλάχιστον k και το πολύ ℓ , μεταξύ δύο διαδοχικών επιτυχιών. Στην απαρίθμηση θεωρούμε ότι μια επιτυχία η οποία δεν βρίσκεται σε άκρο της ακολουθίας μπορεί να συνεισφέρει στη καταμέτρηση δύο συμβολοσειρών, δηλαδή εκείνης που τελειώνει και εκείνης που αρχίζει με αυτήν.

Εάν στην απαρίθμηση συμπεριλάβουμε επιπλέον και συμβολοσειρές της μορφής $S \underbrace{FF \dots F}_{\geq k, \leq \ell}$ που (τυχόν) προκύπτουν μετά τη τελευταία επιτυχία, τότε την τ.μ. που

τις απαριθμεί τη συμβολίζουμε με $N_{n;k,\ell}$, $0 \leq k \leq \ell \leq n - 1$ και

$$N_{n;k,\ell} = M_{n;k,\ell} + I_n, \quad I_n = \sum_{j=k}^{\ell} \prod_{i=0}^{j-1} (1 - X_{n-i}) X_{n-j}. \quad (1)$$

Η $N_{n;k,\ell}$ είναι ιδιαίτερα χρήσιμη για την εξαγωγή αναδρομικών σχέσεων. Στην ακολουθία των $n = 25$ πειραμάτων $FSFFFSSFFSSSSFFSSFFSFFSSFF$ έχουμε ότι $M_{25;1,2} = 3$, $N_{25;1,2} = 4$ και $M_{25;1,3} = 4$, $N_{25;1,3} = 5$.

Η μελέτη συμβολοσειρών φραγμένου μήκους μέσω της $M_{n;k,\ell}$ γενικεύει πρόσφατα αποτελέσματα διαφόρων ερευνητών σε παρόμοιες συμβολοσειρές με μήκος ροής αποτυχιών το πολύ ίσο ($M_{n;0,\ell}$) ακριβώς ίσο ($M_{n;k,k}$) και τουλάχιστον ίσο ($M_{n;k,n-2}$), με ένα μη αρνητικό ακέραιο αριθμό (βλέπε, Sarkar et al. (2004), Sen and Goyal (2004), Holst (2007, 2008), Dafnis et al. (2010) και Makri and Psillakis (2012)). Η $M_{n;0,0}$ ταυτίζεται με την τ.μ. $M_n^{(2)}$ του Ling (1988) που απαριθμεί επικαλυπτόμενες ροές επιτυχιών μήκους 2 (με επικαλυπτόμενο μέρος μήκους το πολύ 1) και έχει μελετηθεί από πολλούς ερευνητές, βλ. Balakrishnan and Koutras (2002).

Τα αποτελέσματα της εργασίας εξάγονται για δυαδικές ακολουθίες ανεξάρτητων και ισόνομων τ.μ. και παρουσιάζονται στην ενότητα 2. Συγκεκριμένα, στην ενότητα 2.1 προσδιορίζουμε μέσω αναδρομικών σχέσεων τη συνάρτηση πιθανότητας, την πιθανογεννήτρια και τη ροπογεννήτρια συνάρτηση καθώς και τις παραγοντικές ροπές και τις ροπές περί το μηδέν των τ.μ. $N_{n;k,\ell}$ και $M_{n;k,\ell}$. Στην ενότητα 2.2 χρησιμοποιώντας συνδυαστική ανάλυση εκφράζουμε τη συνάρτηση πιθανότητας της $M_{n;k,\ell}$ μέσω αθροισμάτων διωνυμικών συντελεστών. Η πλειονότητα των αποτελεσμάτων, της ενότητας 2, δίνεται σε μορφές κατάλληλες για επέκτασή τους σε δυαδικές ακολουθίες γενικότερης δομής όπως σε ακολουθίες ανταλλάξιμων τ.μ. και ανεξάρτητων, όχι κατ' ανάγκη ισόνομων τ.μ. Τέλος, στην ενότητα 3 παρουσιάζουμε μια ενδεικτική αριθμητική εφαρμογή. Σκοπός αυτής είναι να δώσει βοηθητικά αριθμητικά στοιχεία για περαιτέρω κατανόηση των θεωρητικών αποτελεσμάτων.

2. ΚΥΡΙΑ ΑΠΟΤΕΛΕΣΜΑΤΑ

Έστω X_1, X_2, \dots, X_n , $n \geq 2$, μια ακολουθία από ανεξάρτητες και ισόνομες δυαδικές τ.μ. με σταθερή πιθανότητα επιτυχίας $P(X_i = 1) = p = 1 - P(X_i = 0) = 1 - q$, $i = 1, 2, \dots, n$.

2.1 Αναδρομικές σχέσεις για τις τ.μ. $N_{n;k,\ell}$ και $M_{n;k,\ell}$

Αρχικά προσδιορίζουμε αναδρομικές σχέσεις για την συνάρτηση πιθανότητας (Θεώρημα 1), την πιθανογεννήτρια συνάρτηση (Πρόταση 1) και τη ροπογεννήτρια συνάρτηση (Πρόταση 2) της τ.μ. $N_{n;k,\ell}$. Στη συνέχεια χρησιμοποιώντας τα αποτελέσματα αυτά δίνουμε αναδρομικές σχέσεις (Θεώρημα 2 και Προτάσεις 3, 4) για τις αντίστοιχες συναρτήσεις της τ.μ. $M_{n;k,\ell}$.

Θεώρημα 1. Η συνάρτηση πιθανότητας $g_n(x) = P(N_{n;k,\ell} = x)$, $x = 0, 1, \dots, \lfloor \frac{n}{k+1} \rfloor$, της τ.μ. $N_{n;k,\ell}$, $0 \leq k \leq \ell \leq n - 1$, ικανοποιεί για $n \geq k + 1$ το αναδρομικό σχήμα

$$g_n(x) = g_{n-1}(x) - pq^k [g_{n-k-1}(x) - g_{n-k-1}(x-1)] - pq^{\ell+1} [g_{n-\ell-2}(x-1) - g_{n-\ell-2}(x)] \quad (2)$$

με αρχικές συνθήκες $g_n(x) = 0$ αν $x < 0$ ή $x > \lfloor \frac{n}{k+1} \rfloor$, $g_n(x) = 0$ αν $n < 0$ και $g_n(x) = \delta_{x,0}$ για $0 \leq n < k + 1$.

Απόδειξη Για $x < 0$, $x > \lfloor \frac{n}{k+1} \rfloor$ και $n < 0$ το αποτέλεσμα είναι άμεσο. Για $n \geq k+1$ το αποτέλεσμα προκύπτει αν παρατηρήσουμε ότι

$$P(N_{n;k,\ell} = r+1, N_{n-1;k,\ell} = r) = pq^k P(N_{n-1-k;k,\ell} = r), \quad r = 0, 1, \dots$$

$$P(N_{n;k,\ell} = r-1, N_{n-1;k,\ell} = r) = pq^{\ell+1} P(N_{n-\ell-2;k,\ell} = r-1), \quad r = 1, 2, \dots$$

Από το αναδρομικό σχήμα για την συνάρτηση πιθανότητας $g_n(x)$ προκύπτουν αναδρομικές σχέσεις για την πιθανογενήτρια συνάρτηση $\psi_n(t)$, τη ροπογεννήτρια συνάρτηση $\theta_n(t)$ καθώς και τις παραγοντικές ροπές $\rho_{n,r}$ και τις ροπές περί το μηδέν $\nu_{n,r}$ της $N_{n;k,\ell}$.

Πρόταση 1. Η πιθανογενήτρια συνάρτηση $\psi_n(t) = E(t^{N_{n;k,\ell}})$, $t \in R$ της τ.μ. $N_{n;k,\ell}$ ικανοποιεί το αναδρομικό σχήμα

$$\psi_n(t) = \psi_{n-1}(t) - (1-t)pq^k [\psi_{n-1-k}(t) - q^{\ell-k+1}\psi_{n-\ell-2}(t)], \quad n \geq k+1 \quad (3)$$

με την αρχική συνθήκη $\psi_n(t) = 1$ αν $0 \leq n < k+1$ και τη σύμβαση $\psi_n(t) = 0$ αν $n < 0$.

Πόρισμα 1. Οι παραγοντικές ροπές $\rho_{n,r} = E(N_{n;k,\ell}(N_{n;k,\ell}-1) \cdots (N_{n;k,\ell}-r+1))$, $r \geq 1$, της τ.μ. $N_{n;k,\ell}$ δίνονται από το αναδρομικό σχήμα

$$\rho_{n,r} = \rho_{n-1,r} + rpq^k (\rho_{n-k-1,r-1} - q^{\ell-k+1}\rho_{n-\ell-2,r-1}), \quad n \geq k+1 \quad (4)$$

με $\rho_{n,0} = 1$ για $n \geq 0$, $\rho_{n,r} = 0$ για $n < 0$, $r \geq 0$ και $\rho_{n,r} = 0$ για $0 \leq n < k+1$, $r \geq 1$.

Πρόταση 2. Η ροπογεννήτρια συνάρτηση $\theta_n(t) = E(e^{tN_{n;k,\ell}}) = \psi_n(e^t)$, $t \in R$ της τ.μ. $N_{n;k,\ell}$ ικανοποιεί το αναδρομικό σχήμα

$$\theta_n(t) = \theta_{n-1}(t) - (1-e^t)pq^k [\theta_{n-k-1}(t) - q^{\ell-k+1}\theta_{n-\ell-2}(t)], \quad n \geq k+1 \quad (5)$$

με την αρχική συνθήκη $\theta_n(t) = 1$ αν $0 \leq n < k+1$ και τη σύμβαση $\theta_n(t) = 0$ αν $n < 0$.

Πόρισμα 2. Οι ροπές $\nu_{n,r} = E(N_{n;k,\ell}^r)$, $r \geq 1$, της τ.μ. $N_{n;k,\ell}$ δίνονται από το αναδρομικό σχήμα

$$\nu_{n,r} = \nu_{n-1,r} + pq^k \sum_{i=0}^{r-1} \binom{r}{i} (\nu_{n-k-1,i} - q^{\ell-k+1}\nu_{n-\ell-2,i}), \quad n \geq k+1 \quad (6)$$

με $\nu_{n,0} = 1$ για $n \geq 0$, $\nu_{n,r} = 0$ για $n < 0$, $r \geq 0$ και $\nu_{n,r} = 0$ για $0 \leq n < k+1$, $r \geq 1$.

Θεώρημα 2. Η συνάρτηση πιθανότητας $f_n(x) = P(M_{n;k,\ell} = x)$, $x = 0, 1, \dots, \lfloor \frac{n-1}{k+1} \rfloor$, της τ.μ. $M_{n;k,\ell}$, $0 \leq k \leq \ell \leq n-2$, ικανοποιεί για $n \geq k+2$ το αναδρομικό σχήμα

$$f_n(x) = pg_{n-1}(x) + qf_{n-1}(x) \quad (7)$$

με αρχικές συνθήκες $f_n(x) = 0$ αν $x < 0$ ή $x > \lfloor \frac{n-1}{k+1} \rfloor$, και $f_n(x) = \delta_{x,0}$ για $0 \leq n < k+2$.

Απόδειξη Από το Θεώρημα Ολικής Πιθανότητας έχουμε ότι $P(M_{n;k,\ell} = x) = P(M_{n;k,\ell} = x | X_n = 1)P(X_n = 1) + P(M_{n;k,\ell} = x | X_n = 0)P(X_n = 0) = P(N_{n-1;k,\ell} = x)p + P(M_{n-1;k,\ell} = x)q$.

Πρόταση 3. Η πιθανογεννήτρια συνάρτηση $\phi_n(t) = E(t^{M_{n;k,\ell}})$, $t \in R$ της τ.μ. $M_{n;k,\ell}$ ικανοποιεί το αναδρομικό σχήμα

$$\phi_n(t) = p\psi_{n-1}(t) + q\phi_{n-1}(t), \quad n \geq k+2 \quad (8)$$

με την αρχική συνθήκη $\phi_n(t) = 1$ αν $0 \leq n < k+2$.

Πόρισμα 3. Οι παραγοντικές ροπές $\pi_{n,r} = E(M_{n;k,\ell}(M_{n;k,\ell}-1) \cdots (M_{n;k,\ell}-r+1))$, $r \geq 1$, της τ.μ. $M_{n;k,\ell}$ δίνονται από το αναδρομικό σχήμα

$$\pi_{n,r} = p\rho_{n-1,r} + q\pi_{n-1,r}, \quad n \geq k+2 \quad (9)$$

με $\pi_{n,0} = 1$, για $n \geq 0$, $\pi_{n,r} = 0$ για $0 \leq n < k+2$ και $r \geq 1$.

Χρησιμοποιώντας το Πόρισμα 3 μια εναλλακτική έκφραση για τη συνάρτηση πιθανότητας $f_n(x)$ μπορεί να δοθεί από τη σχέση

$$f_n(x) = \frac{1}{x!} \sum_{r \geq x} (-1)^{r-x} \frac{\pi_{n,r}}{(r-x)!} = \sum_{r \geq x} (-1)^{r-x} \binom{r}{x} \pi'_{n,r} \quad (10)$$

όπου $\pi'_{n,r} = E\left(\binom{M_{n;k,\ell}}{r}\right) = \frac{\pi_{n,r}}{r!}$ είναι η r -τάξης διωνυμική ροπή της $M_{n;k,\ell}$.

Πρόταση 4. Η ροπογεννήτρια συνάρτηση $\eta_n(t) = E(e^{tM_{n;k,\ell}}) = \phi_n(e^t)$, $t \in R$ της τ.μ. $M_{n;k,\ell}$ ικανοποιεί το αναδρομικό σχήμα

$$\eta_n(t) = p\theta_{n-1}(t) + q\eta_{n-1}(t), \quad n \geq k+2 \quad (11)$$

με την αρχική συνθήκη $\eta_n(t) = 1$ αν $0 \leq n < k+2$.

Πόρισμα 4. Οι ροπές περί το μηδέν $\mu_{n,r} = E(M_{n;k,\ell}^r)$, $r \geq 1$, της τ.μ. $M_{n;k,\ell}$ δίνονται από το αναδρομικό σχήμα

$$\mu_{n,r} = p\nu_{n-1,r} + q\mu_{n-1,r}, \quad n \geq k+2 \quad (12)$$

με $\mu_{n,0} = 1$ για $n \geq 0$ και $\mu_{n,r} = 0$ για $0 \leq n < k+2$, $r \geq 1$.

Εναλλακτικές αναδρομικές εκφράσεις για την $\phi_n(t)$, $f_n(x)$ και $\mu_{n,r}$ των τ.μ. $M_{n;0,\ell}$, με $0 \leq \ell \leq n-2$, $M_{n;k,k}$, με $0 \leq k \leq n-2$ και $M_{n;k,n-2}$, με $0 \leq k \leq n-2$ έχουν δοθεί από τους Dafnis et al. (2010) χρησιμοποιώντας τη μέθοδο εμβάπτισης των τ.μ. αυτών σε Μαρκοβιανή αλυσίδα.

2.2 Συνάρτηση πιθανότητας της $M_{n;k,\ell}$ μέσω συνδυαστικής ανάλυσης

Στην ενότητα αυτή προσδιορίζουμε με συνδυαστική ανάλυση τη συνάρτηση πιθανότητας της $M_{n;k,\ell}$ (Θεώρημα 3) μέσω συντελεστών απαρίθμησης διανομών όμοιων

σφαιρών σε διακεκριμένες κάλπες περιορισμένης χωρητικότητας (Λήμμα 1, 2).

Λήμμα 1. Έστω $C_{i,r-i}(\alpha, m, k_1 - 1, k_2 - 1)$ ο αριθμός των διαφορετικών διανομών α όμοιων σφαιρών σε m διακεκριμένες κάλπες, i συγκεκριμένες από τις οποίες έχουν χωρητικότητα $k_1 - 1$ και $r - i$ επίσης συγκεκριμένες κάλπες έχουν χωρητικότητα $k_2 - 1$, $0 \leq r \leq m$, $0 \leq i \leq r$, $k_1 \geq 1$, $k_2 \geq 1$. Τότε,

$$C_{i,r-i}(\alpha, m, k_1 - 1, k_2 - 1) = \sum_{j_1=0}^{\lfloor \frac{\alpha}{k_1} \rfloor} \sum_{j_2=0}^{\lfloor \frac{\alpha - k_1 j_1}{k_2} \rfloor} (-1)^{j_1+j_2} \binom{i}{j_1} \binom{r-i}{j_2} \binom{\alpha + m - k_1 j_1 - k_2 j_2 - 1}{m-1}. \quad (13)$$

Απόδειξη Το αποτέλεσμα προκύπτει αν αναπτύξουμε τη γεννήτρια συνάρτηση $g(t) = (1 - t^{k_1})^i (1 - t^{k_2})^{r-i} (1 - t)^{-m}$ των $C_{i,r-i}(\alpha, m, k_1 - 1, k_2 - 1)$.

Για $k_1 = k_2 = 1$ η σχέση (13) γίνεται

$$C_{i,r-i}(\alpha, m, 0, 0) = \binom{\alpha + m - r - 1}{\alpha}. \quad (14)$$

Λήμμα 2. (Makri, Philippou and Psillakis, 2007) Ο αριθμός των διανομών α όμοιων σφαιρών σε m κάλπες, εκ των οποίων r συγκεκριμένες έχουν χωρητικότητα $k - 1$, δίνεται από την

$$H_r(\alpha, m, k - 1) = \sum_{j=0}^{\lfloor \alpha/k \rfloor} (-1)^j \binom{r}{j} \binom{\alpha - kj + m - 1}{\alpha - kj}. \quad (15)$$

Θεώρημα 3. Η συνάρτηση πιθανότητας $P(M_{n;k,\ell} = x)$, $x = 0, 1, \dots, \lfloor \frac{n-1}{k+1} \rfloor$, $n \geq k + 2$, δίνεται από τη σχέση

(α) Για $k \geq 1$

$$P(M_{n;k,\ell} = x) = \sum_{y=x+1}^{n-kx} p^y q^{n-y} \sum_{z=0}^{y-1-x} \binom{y-1}{x} \binom{y-1-x}{z} C_{x,z}(\alpha, y+1, \ell - k, k - 1) + q^n \delta_{x,0},$$

$$\alpha = n - y - kx - (\ell + 1)(y - 1 - x - z). \quad (16)$$

(β) Για $k = 0$

$$P(M_{n;0,\ell} = x) = \sum_{y=x+1}^n p^y q^{n-y} \binom{y-1}{x} H_x(\alpha, y+1, \ell) + q^n \delta_{x,0},$$

$$\alpha = n - y - (\ell + 1)(y - 1 - x). \quad (17)$$

Απόδειξη Έστω Y_n η τ.μ. που παριστάνει τον αριθμό των επιτυχιών στην ακολουθία των n πειραμάτων Bernoulli. Για $k \geq 1$ παρατηρούμε ότι μια συμβολοσειρά $S \underbrace{F \dots F}_k S$ εμφανίζεται στην ακολουθία σε μια από τις μορφές $S \underbrace{F \dots F}_k S$, $S \underbrace{F \dots F}_{\geq k-1} S$ ή $S \underbrace{F \dots F}_{\geq \ell+1} S$ τις οποίες ονομάζουμε τύπου A , B ή Γ , αντίστοιχα. Οι y επιτυχίες δημιουργούν $y + 1$ κάλπες. Στη συνέχεια έχοντας υπ' όψιν ότι η διανομή των $n - y$ αποτυχιών πρέπει να γίνει στις $y + 1$ κάλπες έτσι ώστε να δημιουργηθούν x συμβολοσειρές τύπου A και ότι οι ακολουθίες με τον ίδιο αριθμό επιτυχιών (συνεπώς και αποτυχιών) έχουν την ίδια πιθανότητα, βρίσκουμε ότι για $x + 1 \leq y \leq n - kx$, $P(M_{n;k,\ell} = x, Y_n = y) = \sum_{z=0}^{y-1-x} \binom{y-1}{x} \binom{y-1-x}{z} C_{x,z} (n - y - kx - (y - 1 - x - z)(\ell + 1), y + 1, \ell - k, k - 1) p^y q^{n-y}$. Λαμβάνοντας υπ' όψιν ότι $P(M_{n;k,\ell} = x, Y_n = 0) = q^n \delta_{x,0}$ και αθροίζοντας ως προς y , προκύπτει το (α) μέρος του Θεωρήματος. Η απόδειξη για $k = 0$ γίνεται θεωρώντας ότι οι δημιουργούμενες συμβολοσειρές είναι δύο τύπων, δηλ. $S \underbrace{F \dots F}_{\leq \ell} S$ και $S \underbrace{F \dots F}_{\geq \ell+1} S$.

Για τις ειδικές περιπτώσεις $M_{n;k,k}$ και $M_{n;k,n-2}$, $0 \leq k \leq n - 2$ της τ.μ. $M_{n;k,\ell}$, $0 \leq k \leq \ell \leq n - 2$, οι Sen and Goyal (2004) προσδιόρισαν τη συνάρτηση πιθανότητας τους μέσω αθροισμάτων διωνυμικών συντελεστών και σχολίασαν ότι και η κατανομή της τ.μ. $M_{n;0,\ell}$, $0 \leq \ell \leq n - 2$, μπορεί να βρεθεί με ανάλογο τρόπο. Ο Holst (2008) προσδιόρισε τις διωνυμικές ροπές $E[(M_{n;0,0}^r)]$, μέσω αθροίσματος διωνυμικών συντελεστών, η χρήση των οποίων στο δεύτερο μέρος της (10) δίνει τη συνάρτηση πιθανότητας της $M_{n;0,0}$ συναρτήσει ενός διπλού αθροίσματος διωνυμικών συντελεστών, ενώ η περίπτωση (β) του Θεωρήματος 3 την εκφράζει συναρτήσει ενός απλού αθροίσματος. Αυτό συμβαίνει διότι $H_r(\alpha, m, 0) = \binom{\alpha+m-r-1}{\alpha}$.

3. ΑΡΙΘΜΗΤΙΚΗ ΕΦΑΡΜΟΓΗ

Στη συνέχεια δίνουμε ένα (αριθμητικό) παράδειγμα στο οποίο έχουν χρησιμοποιηθεί όλοι σχεδόν οι τύποι της ενότητας 2. Σκοπός αυτού είναι να δώσει βοηθητικά αριθμητικά στοιχεία για περαιτέρω κατανόηση των θεωρητικών αποτελεσμάτων.

Ο Πίνακας 1 δίνει τις τιμές των $f_n(r)$, $\pi_{n,r}$, $\mu_{n,r}$, για $r = 0, 1, \dots, \lfloor (n - 1)/(k + 1) \rfloor$ καθώς και της $v^2 = V(M_{n;k,\ell}) = \mu_{n,2} - m^2 = \pi_{n,2} + m(1 - m)$, με $m = E(M_{n;k,\ell}) = \pi_{n,1} = \mu_{n,1}$, για τις ενδεικτικές τιμές $n = 11$, $k = 1$, $\ell = 2, 3$ και $p = 1/4, 1/2$. Τα στοιχεία του πίνακα παρουσιάζουν μια εικόνα για τις αριθμητικές τιμές των εμπλεκόμενων συναρτήσεων. Μεταξύ άλλων υποδεικνύουν ότι η μορφή της $f_{11}(r)$ είναι διαφορετική για $p = 1/4$ και $p = 1/2$ και επίσης ότι εμφανίζεται μια μετατόπιση των βαρών της κατανομής προς μεγαλύτερες τιμές (τιμή r του πίνακα) για $p = 1/2$.

Από αριθμητική διερεύνηση της μορφής της $f_{n;k,\ell}(r, p) = P(M_{n;k,\ell} = r)$ για διάφορες τιμές των n , k , ℓ και p προέκυψε ότι για κάποιες τιμές των n , k και ℓ ισχύει ότι $\delta_{n;k,\ell}(p) = \max\{f_{n;k,\ell}(r, p) : 0 \leq r \leq \lfloor \frac{n-1}{k+1} \rfloor\} - f_{n;k,\ell}(0, p) = 0$ για τιμές του p που ανήκουν σε κάποιο ή κάποια υποδιαστήματα του $(0, 1)$. Για τα

υπόλοιπα υποδιαστήματα, για τα οποία η τιμή $r = 0$ δεν είναι κορυφή, παρατηρείται μια μετατόπιση των βαρών της κατανομής προς μεγαλύτερες τιμές του r .

Συγκεκριμένα, για μια διαμέριση του διαστήματος $[0.001, 0.999]$ σε 2^{13} υποδιαστήματα ίσου μήκους, προέκυψε ότι $\delta_{11;1,\ell}(p) = 0$, $\ell = 2, 3$ όταν $p \in \Delta_\ell = [0.001, p_1^{(\ell)}] \cup [p_2^{(\ell)}, 0.999]$ με $p_1^{(2)} = 0.332124$, $p_1^{(3)} = 0.296307$, $p_2^{(2)} = 0.905194$ και $p_2^{(3)} = 0.906047$. Επομένως οι παρατηρήσεις της προηγούμενης παραγράφου ισχύουν για τις $f_{11;1,2}(r, 0.25)$ και $f_{11;1,2}(r, 0.50)$ διότι $0.25 \in [0.001, 0.332124]$ και $0.50 \notin \Delta_2$. Ανάλογα συμπεράσματα ισχύουν και για τις $f_{11;1,3}(r, 0.25)$ και $f_{11;1,3}(r, 0.50)$.

Πίνακας 1: Αριθμητικά χαρακτηριστικά της $M_{11;1,\ell}$, $\ell = 2, 3$ και $p = 1/4, 1/2$.

p	r	$\ell = 2$			$\ell = 3$		
		$f_{11}(r)$	$\pi_{11,r}$	$\mu_{11,r}$	$f_{11}(r)$	$\pi_{11,r}$	$\mu_{11,r}$
1/4	0	0.516557	1.000000	1.000000	0.431809	1.000000	1.000000
	1	0.312209	0.703125	0.703125	0.321987	0.887695	0.887695
	2	0.127886	0.546570	1.249695	0.178695	0.797316	1.685011
	3	0.038308	0.352910	2.675740	0.061772	0.510375	3.790017
	4	0.004982	0.126532	6.773080	0.005678	0.143217	9.674372
	5	0.000058	0.006952	18.996584	0.000058	0.006952	27.045928
v^2		0.755310			0.897008		
1/2	0	0.148926	1.000000	1.000000	0.094727	1.000000	1.000000
	1	0.318359	1.625000	1.625000	0.261230	1.843750	1.843750
	2	0.322754	2.093750	3.718750	0.382813	2.533203	4.376953
	3	0.179199	1.831055	9.737305	0.228516	2.173828	11.617188
	4	0.030273	0.785156	28.052734	0.032227	0.832031	33.451172
	5	0.000488	0.058594	86.717773	0.000488	0.058594	102.566406
v^2		1.078125			0.977539		

ABSTRACT

Consider a sequence of binary (Success-Failure) random variables ordered on a line. Numbers of failure strings of bounded (constrained) length between two subsequent successes are studied under an overlapping enumeration scheme. Exact probability mass functions are obtained by means of combinatorial analysis and via recursive schemes in the case of a sequence of independent and identically distributed random variables. Recursive schemes of the probability and the moment generating functions are also provided.

Ευχαριστίες: Ευχαριστούμε τους ανώνυμους κριτές για τις χρήσιμες και εύστοχες παρατηρήσεις τους στην αρχική έκδοση της εργασίας.

ΑΝΑΦΟΡΕΣ

- Balakrishnan, N. and Koutras, M. V. (2002). *Runs and Scans with Applications*, New York: John Wiley.
- Dafnis, S. D., Philippou, A. N. and Antzoulakos, D. L. (2010). Distributions of patterns of two successes separated by a string of $k - 2$ failures. *Stat. Papers*, DOI 10.1007/s00362-010-0340-7, 1-22.
- Holst, L. (2007). Counts of failure strings in certain Bernoulli sequences. *J. Appl. Prob.* **44**, 824-830.
- Holst, L. (2008). The number of two consecutive successes in a Hoppe-Polya urn. *J. Appl. Prob.* **45**, 901-906.
- Ling, K. D. (1988). On binomial distribution of order k . *Statist. Probab. Lett.* **6**, 247-250.
- Makri, F. S., Philippou, A. N. and Psillakis, Z. M. (2007). Success run statistics defined on an urn model. *Adv. Appl. Prob.* **39**, 991-1019.
- Makri, F. S. and Psillakis, Z. M. (2012). Counting certain binary strings. *J. Statist. Plann. Inference* **142**, 908-924.
- Sarkar, A., Sen, K. and Anuradha (2004). Waiting time distributions of runs in higher order Markov chains. *Ann. Inst. Statist. Math.* **56**, 317-349.
- Sen, K. and Goyal, B. (2004). Distributions of patterns of two failures separated by success runs of length k . *J. Korean Stat. Soc.* **33**, 35-58.



ΚΑΤΑΣΚΕΥΗ ΜΟΝΤΕΛΟΥ ΓΙΑ ΤΗΝ ΕΚΤΙΜΗΣΗ ΤΗΣ ΕΠΙΛΟΓΗΣ ΚΑΤΕΥΘΥΝΣΗΣ ΣΠΟΥΔΩΝ ΤΩΝ ΜΑΘΗΤΩΝ ΤΗΣ Β ΛΥΚΕΙΟΥ.

Μακρής Γεώργιος, Δαλαμήτρα Ευανθία, Αντωνίου Ιωάννης

Π.Μ.Σ. "Επιστήμη του Διαδικτύου", Σ.Θ.Ε., Τμήμα Μαθηματικών, Α.Π.Θ.

geormak@hotmail.com, dalamitra.eva@gmail.com, iantonio@math.auth.gr

ΠΕΡΙΛΗΨΗ

Κατασκευάζεται ένα μη-παραμετρικό στατιστικό μοντέλο πρόβλεψης της επιλογής κατεύθυνσης σπουδών από τους μαθητές της Β Λυκείου. Το μοντέλο υλοποιείται με νευρωνικό δίκτυο το οποίο μαθαίνει με επίβλεψη, με κατάλληλη επιλογή των εισόδων. Ο πρωταρχικός σκοπός είναι να αναγνωρίζει όλα τα πρότυπα – παραδείγματα εκπαίδευσης σε ποσοστό 100%. Κατόπιν ελέγχεται εάν το μοντέλο μπορεί να χρησιμοποιηθεί σε νέα δεδομένα και σε ποιο βαθμό μπορεί να ανταποκριθεί (τι ποσοστό επιτυχίας έχει σε νέα δεδομένα). Η αξιολόγηση γίνεται μετά την εκπαίδευση του νευρωνικού δικτύου με βάση τη σχέση συστήματος εισόδου – εξόδου (black box). Δηλαδή δίδονται οι εισόδοι και συγκρίνονται οι έξοδοι με το πραγματικό αποτέλεσμα, για τα διαθέσιμα δεδομένα. Με βάση αυτά το μοντέλο έδωσε προβλέψεις με επιτυχία 95%.

Λέξεις Κλειδιά: Μοντέλο , Νευρωνικό Δίκτυο , Matlab.

1. ΕΙΣΑΓΩΓΗ

Στόχος της παρούσας έρευνας είναι να κατασκευαστεί ένα μη-παραμετρικό στατιστικό μοντέλο που μπορεί να προβλέψει σε ικανοποιητικό ποσοστό την κατεύθυνση σπουδών που πρόκειται να ακολουθήσουν οι μαθητές που τελειώνουν την Α' Λυκείου.

Το μοντέλο που παρουσιάζεται προέκυψε μετά από πολλές προσπάθειες και δοκιμές σε διάφορα μοντέλα πρόβλεψης όχι μόνο Νευρωνικών Δικτύων. Έγινε προσπάθεια κατασκευής γραμμικών μοντέλων παλινδρόμησης αλλά και εκθετικών μοντέλων παλινδρόμησης όμως τα μοντέλα δεν ήταν δυνατόν να έχουν πολύ καλή προσαρμογή στα δεδομένα. Τα καλύτερα αποτελέσματα πρόβλεψης όμως είχε το μοντέλο που παρουσιάζεται.

Στην αρχή γίνεται μία μικρή παρουσίαση των νευρωνικών δικτύων. Κατόπιν θα παρουσιαστεί το μοντέλο που κατασκευάστηκε και προτείνεται. Ακολουθούν η αξιολόγηση του μοντέλου και τα συμπεράσματα.

2. ΝΕΥΡΩΝΙΚΑ ΔΙΚΤΥΑ

Το Τεχνητό Νευρωνικό Δίκτυο (ΤΝΔ) είναι μία αρχιτεκτονική δομή -δίκτυο- αποτελούμενη από ένα πλήθος διασυνδεδεμένων μονάδων - τεχνητοί νευρώνες. Κάθε μονάδα χαρακτηρίζεται από εισόδους και εξόδους και υλοποιεί τοπικά έναν απλό υπολογισμό. Κάθε σύνδεση μεταξύ δύο μονάδων χαρακτηρίζεται από μια τιμή βάρους. Οι τιμές των βαρών των συνδέσεων αποτελούν τη γνώση που είναι αποθηκευμένη στο δίκτυο και καθορίζουν τη λειτουργικότητα του [Ματσατσίνης(2010)]. Η έξοδος κάθε μονάδας καθορίζεται από τον τύπο της μονάδας, τη διασύνδεση με τις υπόλοιπες μονάδες και πιθανώς κάποιες εξωτερικές εισόδους. Πέρα από μία πιθανή δεδομένη εκ κατασκευής λειτουργική ικανότητα ενός δικτύου, συνήθως ένα δίκτυο αναπτύσσει μία συνολική λειτουργικότητα μέσω μίας μορφής εκπαίδευσης [Haykin(1999)].

2.1 Δομή του νευρωνικού δικτύου

Οι νευρώνες είναι τα δομικά στοιχεία του δικτύου. Κάθε τέτοιος κόμβος δέχεται ένα σύνολο αριθμητικών εισόδων από διαφορετικές πηγές (είτε από άλλους νευρώνες, είτε από το περιβάλλον), επιτελεί έναν υπολογισμό με βάση αυτές τις εισόδους και παράγει μία έξοδο. Η εν λόγω έξοδος είτε κατευθύνεται στο περιβάλλον, είτε τροφοδοτείται ως είσοδος σε άλλους νευρώνες του δικτύου. Υπάρχουν τρεις τύποι νευρώνων: οι νευρώνες εισόδου, οι νευρώνες εξόδου και οι υπολογιστικοί νευρώνες ή κρυμμένοι νευρώνες. Οι νευρώνες εισόδου δεν επιτελούν κανέναν υπολογισμό, μεσολαβούν απλώς ανάμεσα στις περιβαλλοντικές εισόδους του δικτύου και στους υπολογιστικούς νευρώνες. Οι νευρώνες εξόδου διοχετεύουν στο περιβάλλον τις τελικές αριθμητικές εξόδους του δικτύου. Οι υπολογιστικοί νευρώνες πολλαπλασιάζουν κάθε είσοδό τους με το αντίστοιχο συναπτικό βάρος και υπολογίζουν το ολικό άθροισμα των γινομένων. Το άθροισμα αυτό τροφοδοτείται ως όρισμα στη συνάρτηση ενεργοποίησης, την οποία υλοποιεί εσωτερικά κάθε κόμβος. Η τιμή που λαμβάνει η συνάρτηση για το εν λόγω όρισμα είναι και η έξοδος του νευρώνα για τις τρέχουσες εισόδους και βάρη [Rumelhart(1996),Αντωνίου(2010)].

Στόχος ενός νευρωνικού δικτύου είναι να προβλέψει ένα γεγονός για το οποίο διατίθενται διαχρονικά στοιχεία ή να ομαδοποιήσει ένα συγκεκριμένο όγκο δεδομένων. Η διαδικασία κατά την οποία ένα δίκτυο συντίθεται και παράγει αποτέλεσμα μπορεί να χωριστεί σε ένα αριθμό βημάτων, τα οποία περιγράφονται στη συνέχεια και αποτέλεσαν και το μεθοδολογικό πλαίσιο της συγκεκριμένης εργασίας [Αντωνίου(2010)].

Βήμα 1^ο: Συγκέντρωση των δεδομένων

Αρχικά συγκεντρώνονται όσα στοιχεία κρίνονται ότι είναι απαραίτητα για το πρόβλημα που μελετάται. Ανάλογα με τη μορφή του δικτύου καθορίζεται και η ποσότητα των δεδομένων. Γενικά για να λειτουργήσει ένα δίκτυο

ικανοποιητικά χρειάζεται αρκετό όγκο δεδομένων. Αυτό δε σημαίνει όμως ότι όσο περισσότερα δεδομένα διατίθενται στο δίκτυο τόσο καλύτερα γιατί μπορεί να αποβεί χρονοβόρο και δαπανηρό. Τα δεδομένα θα πρέπει να χωρισθούν σε δύο αρχεία. Το ένα θα αποτελεί το αρχείο εκπαίδευσης (training file) και το άλλο το αρχείο με το οποίο θα ελεγχθεί το δίκτυο (test file).

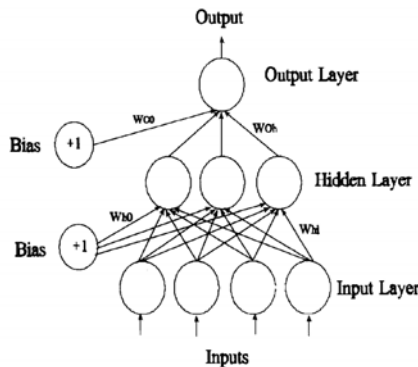
Βήμα 2^ο: Μετασχηματισμός των δεδομένων.

Μετασχηματίζονται όλα τα δεδομένα ώστε να έχουν μέση τιμή 0 και τυπική απόκλιση 1. [Rumelhart(1996)]

2.2 Δίκτυα MLP

Ένα MLP (Multi Layer Perceptron - πολύ επίπεδο Perceptron) αποτελείται γενικά από διάφορα στρώματα ή επίπεδα κόμβων. Το πρώτο ή το χαμηλότερο στρώμα είναι ένα στρώμα εισόδου όπου παραλαμβάνει τις εξωτερικές πληροφορίες. Το τελευταίο ή το υψηλότερο στρώμα είναι ένα στρώμα εξόδου όπου λαμβάνεται η λύση του προβλήματος. Το στρώμα εισόδου και το στρώμα εξόδου χωρίζονται από ένα ή περισσότερα ενδιάμεσα στρώματα αποκαλούμενα κρυμμένα στρώματα. Οι κόμβοι στα παρακείμενα στρώματα συνήθως συνδέονται πλήρως με ακυκλικά τόξα από ένα χαμηλότερο στρώμα σε ένα υψηλότερο στρώμα. Στο σχήμα που ακολουθεί παρουσιάζεται ένα πλήρως συνδεδεμένο προδρομικής τροφοδότησης δίκτυο με δύο κρυμμένα επίπεδα [Rumelhart (1996)]:

Εικόνα 1. Δίκτυο MLP-B



3. Κατασκευή του Μοντέλου

Δεδομένα

1.214 μαθητές της Α' τάξης Λυκείου (1.214 διανύσματα των 9+1 τιμών). Για κάθε μαθητή καταγράφηκαν:

- Οι γραπτές βαθμολογίες στα επιλεγμένα 8 μαθήματα.
- Ο Μέσος Όρος (ΜΟ) βαθμολογίας της Α' Λυκείου.
- Η κατεύθυνση που επέλεξαν το επόμενο έτος.

Το δείγμα των μαθητών προέρχεται από Λύκεια όλης της Ελλάδας. Συγκεκριμένα από ένα Λύκειο κάθε Νομού. Το Λύκειο του κάθε Νομού προέκυψε με κλήρωση [Φαρμάκης(1992)].

Παραδοχές:

Η επιλογή κατεύθυνσης στη Β' Λυκείου από τους μαθητές εξαρτάται από την επίδοσή τους (γραπτά) στις εξετάσεις της Α' Λυκείου. Επιλέγονται τα μαθήματα : αρχαία ελληνικά , νεοελληνική λογοτεχνία , έκθεση , ιστορία , άλγεβρα , γεωμετρία , φυσική , χημεία και τον τελικό Μέσο Όρο (ΜΟ) προαγωγής του μαθητή από την Α' Τάξη. Η επιλογή των παραπάνω μαθημάτων έγινε με γνώμονα την άμεση σχέση που έχουν με τα αντίστοιχα μαθήματα κατεύθυνσης της Γ' τάξης Λυκείου που εξετάζονται πανελλαδικά.

Δημιουργία Μοντέλου

Επιλέγεται για το δίκτυο μας ένα νευρωνικό δίκτυο MLP-B με δύο επίπεδα νευρώνων. Το κρυφό επίπεδο (hidden layer) και το επίπεδο εξόδου (out layer). Το επίπεδο εξόδου έχει μόνο ένα νευρώνα εξόδου. Οι εισοδοί στο νευρωνικό δίκτυο είναι οι βαθμοί των γραπτών σε 8 μαθήματα και ο Μέσος Όρος (ΜΟ) βαθμολογίας. (9 εισοδοί + 1 σταθερής τάσης (=1) = 10 Σύνολο εισόδων στο δίκτυο). Γίνεται μετασχηματισμός των δεδομένων: η κάθε τιμή μετατρέπεται σε νέα με τον τύπο

$$x'_{i,k} = \frac{x_{i,k} - \bar{x}_k}{s_k}, \quad i=1...1214, k=1..9, \text{ όπου : } x'_{i,k} \text{ ο νέος βαθμός του } i \text{ μαθητή στο}$$

k μάθημα, $x_{i,k}$ ο κανονικός βαθμός του i μαθητή στο k μάθημα, \bar{x}_k ο μέσος όρος των βαθμών του k μαθήματος όλων των μαθητών και s_k η τυπική απόκλιση των βαθμών του k μαθήματος όλων των μαθητών.

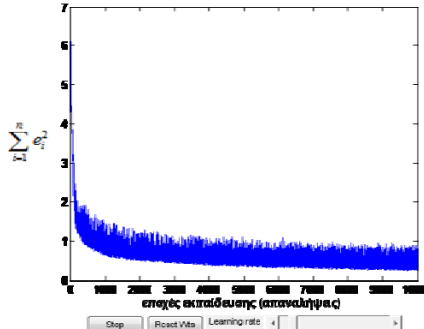
Το MLP-B και με μικρό αριθμό κρυφών νευρώνων είναι δυνατόν να προσεγγίσει τα δεδομένα σε πολύ μεγάλο ποσοστό , όμως χρειάζεται μεγάλο αριθμό εποχών εκπαίδευσης. Ο αριθμός των εποχών εκπαίδευσης είναι ο αριθμός των επαναλήψεων που γίνονται για να επαναπροσδιοριστούν οι πίνακες των βαρών εισόδου W_i (είναι ο τελικός πίνακας των βαρών εισόδου, έχει διάσταση: [πλήθος εισόδων] x [πλήθος κρυφών νευρώνων]) και εξόδου W_o (είναι ο τελικός πίνακας των βαρών εξόδου, έχει διάσταση: [πλήθος κρυφών νευρώνων] x [πλήθος εξόδων]) του δικτύου.

Εφαρμόζεται επαναληπτικά ο αλγόριθμος του υπολογισμού των βαρών δοκιμάζοντας κάθε φορά διαφορετικό αριθμό κρυφών νευρώνων. Ξεκινώντας από έναν κρυφό νευρώνα υπολογίζεται ο πίνακας των βαρών εισόδου και εξόδου του δικτύου και, εάν δεν υπάρξει 100% προσαρμογή του μοντέλου στα δεδομένα εκπαίδευσης, τότε αυξάνεται ο αριθμός των κρυφών νευρώνων κατά έναν. Οι επαναλήψεις – εποχές εκπαίδευσης έχουν οριστεί σε 10.000.

Για να δοκιμαστεί και να προσομοιωθεί το νευρωνικό δίκτυο αναπτύχθηκε λογισμικό σε MatLab στο οποίο είναι δυνατή η εκπαίδευση των νευρώνων

παραμετρικά με βάση τα στοιχεία που δίνονται. Το λογισμικό είναι γενικού σκοπού, που σημαίνει ότι μπορεί να εκπαιδευτεί για οποιοδήποτε παράδειγμα (ανεξαρτήτως του τρέχοντος προβλήματος). Το σχήμα 1 είναι στιγμιότυπο της εφαρμογής που αναπτύχθηκε στο MatLab και παριστάνει το άθροισμα των τετραγώνων των σφαλμάτων του δικτύου εκπαίδευσης σε σχέση με τις εποχές εκπαίδευσης (επαναλήψεις) του νευρωνικού δικτύου.

Σχήμα 1: Εκπαίδευση Νευρωνικού δικτύου με 4 κρυφούς νευρώνες, 10.000 εποχές εκπαίδευσης και 0.1 ρυθμός μάθησης

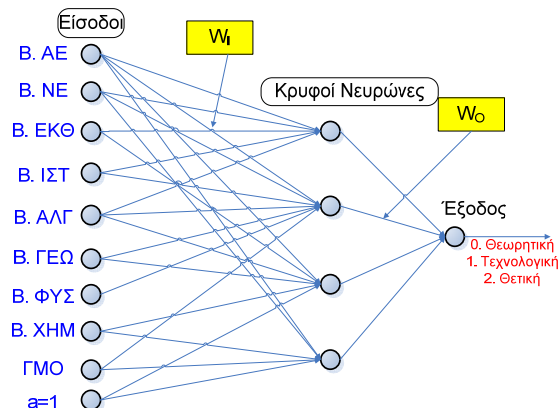


Σωστές Έξοδοι :	1.214
Πλήθος Δειγμάτων :	1.214
Ποσοστό Επιτυχίας =	100.000000 %

Η σύγκλιση του νευρωνικού δικτύου γίνεται σε περίπου 1.000 εποχές εκπαίδευσης. Οι 10.000 εποχές επιλέχθηκαν σε περίπτωση που δεν είχαμε σύγκλιση γρήγορα.

Προφανώς με 4 κρυφούς νευρώνες έχουμε 100% επιτυχία στα δεδομένα εκπαίδευσης. Το νευρωνικό δίκτυο με 4 κρυφούς νευρώνες έχει την παρακάτω μορφή:

Εικόνα 2. Αναπαράσταση του Νευρωνικού δικτύου που επιλέχθηκε.



Η έξοδος του νευρώνα θα δίνεται από τη σχέση :

$$NetWork\ Output = \left[\tanh(Inputs \cdot W_i) \right] \cdot W_o \quad (1)$$

Το διάνυσμα *Inputs* περιέχει τις βαθμολογίες των μαθημάτων με την σειρά που περιγράφονται παραπάνω:

$$Inputs = (\beta_1 \quad \beta_2 \quad \beta_3 \quad \beta_4 \quad \beta_5 \quad \beta_6 \quad \beta_7 \quad \beta_8 \quad MO \quad 1)$$

Έχουμε $NetWork\ Output \in \{0,1,2\}$. Η έξοδος 0 αντιστοιχεί στη Θεωρητική κατεύθυνση, η 1 στην Τεχνολογική και η 2 στη Θετική (οι τιμές των βαρών υπολογίστηκαν με επιθυμητή έξοδο του δικτύου: $\{0,1,2\}$).

Επιλέχθηκε (μετά από δοκιμές σε διάφορες συναρτήσεις μεταφοράς) η μη γραμμική συνάρτηση μεταφοράς $\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$ για το κρυφό επίπεδο (γιατί έδινε καλύτερα αποτελέσματα και είχε γρηγορότερη σύγκλιση) και η γραμμική του «αθροίσματος» για το επίπεδο εξόδου (Διαμαντάρας, 2007). Η συνάρτηση μεταφοράς \tanh εφαρμόζεται σε καθένα από τα στοιχεία του πίνακα: $(Inputs \cdot W_i)$

Ο πίνακας των βαρών της εισόδου του δικτύου (διαστάσεων : 10x4) υπολογίστηκε και είναι :

$$W_i = \begin{pmatrix} 0,796113 & -3,684245 & 0,119323 & 2,595828 \\ -4,332786 & -0,881493 & -2,319520 & 3,483173 \\ -1,475764 & -0,803497 & -0,732227 & 4,102054 \\ -4,545607 & -0,454925 & 5,304011 & 0,477580 \\ -5,542296 & -1,885284 & 3,139690 & 3,931324 \\ -1,021515 & 1,056403 & -8,108079 & -0,948392 \\ 7,916003 & -0,079651 & -1,829020 & -4,393408 \\ 3,825567 & 0,965811 & -4,622280 & -9,629415 \\ 5,841470 & 0,730410 & -0,960588 & 2,410040 \\ -0,905999 & -3,542296 & 6,180753 & -1,374190 \end{pmatrix}$$

Με αντίστοιχο πίνακα βαρών (διαστάσεων : 4x1) για την έξοδο του δικτύου :

$$W_o = \begin{pmatrix} -1,113206 \\ -0,084276 \\ -1,050391 \\ -1,198570 \end{pmatrix}$$

4. ΑΞΙΟΛΟΓΗΣΗ

Η αξιολόγηση γίνεται με τη συνταγή των μηχανικών, με το μαύρο κουτί (black box testing), με δεδομένες εισόδους συγκρίνονται οι έξοδοι του μοντέλου με τα επιθυμητά αποτελέσματα. Δηλαδή συγκρίνεται η έξοδος (η εκτιμώμενη

κατεύθυνση σπουδών) που δίνει το μοντέλο σε νέα δεδομένα με την πραγματική κατεύθυνση σπουδών των δεδομένων.

Αξιολόγηση του μοντέλου

Το μοντέλο που κατασκευάστηκε αναγνωρίζει τα πρότυπα σε ποσοστό 100%.

Θέλουμε να δοκιμάσουμε το μοντέλο εάν μπορεί να αναγνωρίσει νέα δεδομένα και το ποσοστό επιτυχίας που θα έχει πάνω σε αυτά.

Το νευρωνικό δικτύου για τα νέα δεδομένα των 465 μαθητών δίνει τα παρακάτω αποτελέσματα:

Σωστές Έξοδοι : 439
Πλήθος Δειγμάτων : 465
Ποσοστό Επιτυχίας = **94.41 %**

Το μοντέλο έχει **94.41 %** ποσοστό επιτυχίας στα νέα στοιχεία.

Το ποσοστό επιτυχίας θα μπορούσε να είναι και μικρότερο από αυτό που βγήκε με τα τρέχοντα δεδομένα.

Οι λόγοι που πιθανόν να υπάρξουν και χαμηλότερα ποσοστά επιτυχίας είναι :

- Τα δεδομένα ελέγχου – test του νευρωνικού δικτύου προέρχονται από επιλογές των μαθητών που δε γνώριζαν την επίδοσή τους στις εξετάσεις (η πρόθεση επιλογής κατεύθυνσης έγινε πριν από αυτές).
- Η επίδοση σε μία εξέταση μπορεί να μην ανταποκρίνεται στις πραγματικές δυνάμεις του μαθητή.

Δεν λήφθηκαν υπόψη στις εισόδους :

- Εάν η αρχική επιλογή τους ήταν σωστή η όχι.
- Εάν τελικά θα πετύχουν οι μαθητές από την επιλογή που έχουν κάνει.
- Η επιλογή κατεύθυνσης από τους μαθητές γίνεται και με άλλους παράγοντες (π.χ. επιλογή επαγγέλματος – αγορά εργασίας , κοινωνικό και οικογενειακό περιβάλλον κλπ).
- Οι μαθητές έχουν την τάση στη Γ' Λυκείου να αλλάζουν κατεύθυνση (πολλοί μαθητές της Θετικής Κατεύθυνσης της Β' Λυκείου επιλέγουν Τεχνολογική στη Γ' Λυκείου).

ΣΥΜΠΕΡΑΣΜΑΤΑ

Παρόλο που έγιναν προσπάθειες να κατασκευαστούν γραμμικά και εκθετικά μοντέλα παλινδρόμησης ώστε να μπορούν να προβλέπουν την κατεύθυνση σπουδών που θα ακολουθήσουν οι μαθητές στη Β' Λυκείου, λαμβάνοντας υπόψη τις παραδοχές που χρησιμοποιήθηκαν για το μοντέλο που κατασκευάστηκε με χρήση Νευρωνικών Δικτύων, τα μοντέλα αυτά δεν μπόρεσαν να έχουν 100% προσαρμογή στα δεδομένα. Τα μοντέλα επίσης είχαν 60-70 % επιτυχία σε νέα δεδομένα.

Χρησιμοποιώντας ένα νευρωνικό δίκτυο τύπου MLP-B με 1 κρυφό επίπεδο 4 νευρώνων δημιουργήθηκε το μη-παραμετρικό στατιστικό μοντέλο :

$$NetWork\ Output = \left[\tanh(Inputs \cdot W_i) \right] \cdot W_o$$

Οι πίνακες των βαρών εισόδου και εξόδου που υπολογίστηκαν δίνουν την δυνατότητα να προβληθεί η κατεύθυνση σπουδών που πρόκειται να ακολουθήσει ένας μαθητής στη Β' Λυκείου σε ποσοστό 94,41%. Τα μόνα δεδομένα που χρειάζεται κανείς είναι οι βαθμοί που έχει γράψει ο μαθητής στα επιλεγόμενα μαθήματα καθώς και ο γενικός μέσος όρος προαγωγής του.

Με βάση το μοντέλο που κατασκευάστηκε είναι δυνατόν να εκτιμηθούν με ακρίβεια κοντά στο 95% οι επιλογές των μαθητών για τη Β' Λυκείου. Έτσι το κάθε Λύκειο μπορεί να εκτιμήσει τις ώρες των καθηγητών που απαιτούνται για να λειτουργήσει η Β' Λυκείου του σχολείου χωρίς να χρειάζεται να περιμένει την ολοκλήρωση των αιτήσεων.

Οι καινοτομίες της εργασίας είναι:

- α) δεν έχει κατασκευαστεί κάποιο μοντέλο για την εκτίμηση της κατεύθυνσης σπουδών που θα ακολουθήσουν οι μαθητές της Β' Λυκείου με βάση τις επιδόσεις τους στα σχετικά μαθήματα
- β) το μη-παραμετρικό στατιστικό μοντέλο αποτελεί νέα εφαρμογή των νευρωνικών δικτύων.

ABSTRACT

We construct a non-parametric Neural Network Model predicting the selection of the direction of specialized studies by students one year before their graduation. The neural network is based on supervised learning with suitable selection of inputs. The objective of the training is to recognize all the test examples of training with 100% success. Afterwards we test the model performance with new data. The assessment is made after the training of the neural network on the basis of the relation between input-output (black box). That is, the inputs are given and the outputs are compared with the available data. The model performance was found to reach almost 95% success.

ΑΝΑΦΟΡΕΣ

- Haykin, S., (1999), *Neural Networks: A Comprehensive Foundation*, Prentice Hall, New York.
- Rumelhart D.E., Hinton G.E. and Williams R.J., (1996), "*Learning internal representation by error propagation*" In *Parallel Distributed Processing: Exploration in the Microstructure of Cognition* (D.E Rumelhart and J.L. McClelland, eds.), vol. 1, chapter 8, Cambridge, MA, MIT Press.
- Αντωνίου Ι, (2010), *Διαλέξεις στο μεταπτυχιακό πρόγραμμα σπουδών "Επιστήμη του διαδικτύου"*, για το μάθημα WS07. Information Processing and Networks
- Διαμαντάρας, Κ. (2007) *Τεχνητά Νευρωνικά Δίκτυα, Κλειδάριθμος*, Αθήνα
- Ματσατσίνης Ν., (2010), *Συστήματα Υποστήριξης Αποφάσεων*, Εκδόσεις Νέων Τεχνολογιών, Αθήνα.
- Φαρμάκης Ν., (1992), *Εισαγωγή στην δειγματοληψία*, Χριστοδουλίδη, Θεσσαλονίκη.



ΣΤΑΤΙΣΤΙΚΗ ΑΝΑΛΥΣΗ ΠΑΡΑΓΟΝΤΩΝ ΠΡΟΚΛΗΣΗΣ ΘΑΝΑΤΗΦΟΡΩΝ ΕΡΓΑΤΙΚΩΝ ΑΤΥΧΗΜΑΤΩΝ

Μ. Μαντάκης, Π. Κατσακιώρη

Τμήμα Μηχανολόγων και Αεροναυπηγών Μηχανικών, Πανεπιστήμιο Πατρών
manata@mech.upatras.gr, pkatsak@upatras.gr

ΠΕΡΙΛΗΨΗ

Στην παρούσα εργασία επιχειρείται η αναγνώριση παραγόντων που προκαλούν τα εργατικά ατυχήματα και αναδεικνύεται η παρουσία εξωτερικών (περιβαλλοντικοί, οργανωτικοί και εργασιακοί) και εσωτερικών αιτιωδών παραγόντων (ανθρώπινα και μεμονωμένα χαρακτηριστικά). Η στατιστική ανάλυση έγινε με τη χρήση του προγράμματος SPSS και εφαρμόστηκε η ανάλυση κύριων συνιστωσών όσον αφορά στην παραγοντική ανάλυση. Από τα αποτελέσματα της ανάλυσης, προέκυψε ότι κάθε εργατικό ατύχημα ήταν το αποτέλεσμα ενός συνόλου αιτιωδών παραγόντων.

Λέξεις Κλειδιά: εργατικά ατυχήματα, παραγοντική ανάλυση, ανάλυση κύριων συνιστωσών.

1. ΕΙΣΑΓΩΓΗ

Τα στατιστικά στοιχεία για τη συχνότητα των εργατικών ατυχημάτων (Ε.Α.) και τις επιπτώσεις τους στην κοινωνία, καταδεικνύουν την ανάγκη επιστημονικής διερεύνησης του συγκεκριμένου θέματος.

Στο σχεδιασμό ολοκληρωμένης επιστημονικής μεθόδου διερεύνησης-πρόληψης, προϋπόθεση είναι η έρευνα για τον προσδιορισμό των αιτιών πρόκλησης των Ε.Α., δηλαδή η αναγνώριση των παραγόντων που επιδρούν καθώς και του τρόπου με τον οποίο αυτοί αλληλεπιδρούν στην πρόκληση ενός Ε.Α..

Η διερεύνηση συνήθως στηρίζεται στη διαδικασία απόδοσης των αιτιών σε παράγοντες όπως το ανθρώπινο λάθος χωρίς να αποκαλύπτει τους πραγματικούς αιτιώδεις παράγοντες πρόκλησης όπως τις υπάρχουσες για μεγάλο χρονικό διάστημα οργανωτικές ανεπάρκειες, τα προβλήματα στο σχεδιασμό, τις ελλείψεις σε εξοπλισμό, ανθρώπινο δυναμικό, εκπαίδευση, διαδικασίες (Κατσακιώρη, 2010).

Σκοπός της παρούσας εργασίας είναι η αναγνώριση παραγόντων που προκαλούν τα εργατικά ατυχήματα αναδεικνύοντας την παρουσία εξωτερικών (περιβαλλοντικοί, οργανωτικοί και εργασιακοί) και εσωτερικών αιτιωδών παραγόντων (ανθρώπινα και μεμονωμένα χαρακτηριστικά).

2. ΟΙ ΜΕΤΑΒΛΗΤΕΣ ΚΑΙ ΤΑ ΔΕΔΟΜΕΝΑ

Μελετήθηκαν 63 θανατηφόρα Ε.Α. (50 στην κατασκευή και 13 στη βιομηχανία), στην Ανατολική Αττική το διάστημα 1999-2003 με τη χρήση ερωτηματολογίου που περιείχε τους αιτιώδεις παράγοντες (Υπουργείο Εργασίας και Κοινωνικής Ασφάλισης, 2004). Επιλέχθηκε το συγκεκριμένο χρονικό διάστημα λόγω της αυξημένης δραστηριότητας που παρατηρήθηκε στην Ανατολική Αττική με την εγκατάσταση και λειτουργία του Διεθνούς Αερολιμένα Αθηνών καθώς και την προετοιμασία για τη διεξαγωγή των Ολυμπιακών Αγώνων. Διερευνήθηκαν εκθέσεις έρευνας που συντάχθηκαν από τους επιθεωρητές εργασίας με βάση ερωτηματολόγιο το οποίο προσαρμόστηκε και τροποποιήθηκε για τους σκοπούς της παρούσας μελέτης σχετικά με ορισμένες ερωτήσεις (Gyekye, 2003).

Για τις ανάγκες της έρευνας, έγινε διάκριση των παραγόντων σε εξωτερικές και εσωτερικές αιτιώδεις μεταβλητές. Οι εξωτερικές αιτιώδεις μεταβλητές αντιστοιχούν στους περιβαλλοντικούς, οργανωτικούς και εργασιακούς παράγοντες. Οι εσωτερικές αιτιώδεις μεταβλητές αντιστοιχούν στα ανθρώπινα και μεμονωμένα χαρακτηριστικά των θυμάτων των ατυχημάτων (Πίνακας 1).

Πίνακας 1. Ερωτηματολόγιο προς συμπλήρωση από τους αρμόδιους επιθεωρητές.

Εξωτερικοί Αιτιώδεις Παράγοντες (περιβαλλοντικοί, οργανωτικοί και εργασιακοί παράγοντες)						
1	excauf 1: Χαμηλά ημερομίσθια	1	2	3	4	5
2	excauf 2: Εξοικονόμηση χρόνου και προβλημάτων	1	2	3	4	5
3	excauf 3: Φόρτος εργασίας	1	2	3	4	5
4	excauf 4: Ανασφαλής ή ελαττωματικός εξοπλισμός και εγκαταστάσεις	1	2	3	4	5
5	excauf 5: Ανεπαρκής κατάρτιση και επίβλεψη	1	2	3	4	5
6	excauf 6: Πίεση από τη διοίκηση/τον εργοδηγό	1	2	3	4	5
7	excauf 7: Έλλειψη συντονισμού	1	2	3	4	5
8	excauf 8: Προγραμματισμός εργασιών και παραγωγής	1	2	3	4	5
9	excauf 9: Το θύμα ήταν νέο στη θέση εργασίας	1	2	3	4	5
10	excauf 10: Έλλειψη νοικοκυρεμένων συνθηκών εργασίας	1	2	3	4	5
11	excauf 11: Έλλειψη σωστού προστατευτικού εξοπλισμού	1	2	3	4	5
12	excauf 12: Ασαφής και δύσκολη εργασία	1	2	3	4	5
Εσωτερικοί Αιτιώδεις Παράγοντες (ανθρώπινα και μεμονωμένα χαρακτηριστικά)						
1	intcauf 1: Έλλειψη δεξιοτήτων και γνώσης	1	2	3	4	5
2	intcauf 2: Υποεκτίμηση της εργασίας	1	2	3	4	5

3	intcauf 3: Έλλειψη προσοχής: επιλογή λάθους δράσης	1	2	3	4	5
4	intcauf 4: Εσφαλμένη εκτίμηση της κατάστασης	1	2	3	4	5
5	intcauf 5: Μη χρήση μέσων ατομικής προστασίας	1	2	3	4	5
6	intcauf 6: Έλλειψη κατανόησης και ικανότητας να ακολουθηθούν οι εργασιακές πρακτικές που έχουν καθοριστεί	1	2	3	4	5
7	intcauf 7: Ροπή σε ατυχήματα	1	2	3	4	5
8	intcauf 8: Απειρία και λάθος εργασιακές προτεραιότητες	1	2	3	4	5
9	intcauf 9: Απροσεξία	1	2	3	4	5
10	intcauf 10: Προσπάθεια εντυπωσιασμού των συναδέρφων	1	2	3	4	5
11	intcauf 11: Άγνοια: έλλειψη συνειδητοποίησης των σχετικών κινδύνων	1	2	3	4	5
12	intcauf 12: Κούραση και μειωμένη επαγρύπνηση	1	2	3	4	5
13	intcauf 13: Σκόπιμες Παραβιάσεις: προφανής αμέλεια τήρησης των εργασιακών διαδικασιών	1	2	3	4	5

Οι συμμετέχοντες επιθεωρητές απάντησαν ένα ερωτηματολόγιο 25 ερωτήσεων χρησιμοποιώντας μία κλίμακα πέντε βαθμών (1=ελάχιστη, 2=πολύ μικρή, 3=ούτε μεγάλη ούτε μικρή, 4=πολύ μεγάλη, 5=μέγιστη) για να βαθμολογήσουν την επίδραση των παραγόντων, ταξινομημένοι σε εξωτερικές και εσωτερικές αιτιώδεις μεταβλητές.

Ο αριθμός των υπό μελέτη θανατηφόρων Ε.Α. φαίνεται στον Πίνακα 2. Τα περισσότερα θανατηφόρα ατυχήματα συνέβησαν το 2002 ενώ τα λιγότερα το 2001.

Οι κυριότερες κατηγορίες ατυχημάτων ήταν η πτώση από ύψος (58,73%), η έκθεση σε ή επαφή με το ηλεκτρικό ρεύμα (12,69%) και η συμπίεση μέσα ή ανάμεσα σε αντικείμενα (11,11%) (Πίνακας 3).

Η στατιστική ανάλυση έγινε με τη χρήση του προγράμματος SPSS έκδοση 11.0 (SPSS Inc Σικάγο, IL, ΗΠΑ) και εφαρμόστηκε η ανάλυση κύριων συνιστωσών όσον αφορά στην παραγοντική ανάλυση (Moser and Kalton, 1972; Jolliffe, 1986; Tabachnick and Fidell, 1996).

Αποκλείστηκαν από την ανάλυση πέντε ερωτήσεις που αναφέρονται στον εξωτερικό αιτιώδη παράγοντα: «χαμηλά ημερομίσθια» και στους εσωτερικούς αιτιώδεις παράγοντες: «ροπή σε ατυχήματα», «προσπάθεια εντυπωσιασμού των συναδέρφων», «κούραση και μειωμένη επαγρύπνηση» και «σκόπιμες παραβιάσεις: προφανής αμέλεια τήρησης των εργασιακών διαδικασιών» επειδή οι επιθεωρητές εργασίας δεν τους σημείωσαν για την πλειοψηφία των ατυχημάτων κατά τη διάρκεια της συμπλήρωσης του ερωτηματολογίου. Η ανάλυση κύριων συνιστωσών με την περιστροφή μέγιστης διασποράς (Varimax rotation) εφαρμόστηκε για τους υπόλοιπους 20 αιτιώδεις παράγοντες (11 εξωτερικοί και 9 εσωτερικοί).

Πίνακας 2. Θανατηφόρα Ε.Α. ανά έτος: 1999-2003 στην περιοχή της Ανατολικής Αττικής.

Έτος	Αριθμός θανατηφόρων Ε.Α.
1999	12
2000	10
2001	9
2002	19
2003	13
Σύνολο	63

Πίνακας 3. Θανατηφόρα Ε.Α. ανά κατηγορία: 1999-2003 στην περιοχή της Ανατολικής Αττικής.

Τύπος ατυχήματος	Αριθμός θανατηφόρων Ε.Α.
Εκθεση σε ή επαφή με ηλεκτρικό ρεύμα	8
Πτώση από ύψος	37
Κτυπήματα από πύπτοντα αντικείμενα (γλιστρήματα, καταρρεύσεις-υποχωρήσεις κλπ.)	6
Συμπίεση μέσα ή ανάμεσα σε αντικείμενα	7
Εκθεση σε ή επαφή με επιβλαβείς ουσίες (εισπνοή, κατάποση ή απορρόφηση επιβλαβών ουσιών)	2
Άλλες κατηγορίες ατυχημάτων	3
Σύνολο	63

3. ΠΕΡΙΓΡΑΦΗ ΤΗΣ ΣΤΑΤΙΣΤΙΚΗΣ ΑΝΑΛΥΣΗΣ

Η διαδικασία της παραγοντικής ανάλυσης ακολουθεί τέσσερα στάδια:

α) τον υπολογισμό και τη διερεύνηση των συντελεστών συσχέτισης: Εάν οι συσχετίσεις είναι χαμηλές (με βάση τα επίπεδα στατιστικής σημαντικότητας), είναι σχεδόν αδύνατο οι μεταβλητές να μοιράζονται κοινούς παράγοντες. Μέτρο της καταλληλότητας της δειγματοληψίας είναι ο δείκτης Kaiser Meyer-Olkin (KMO). Για να έχει νόημα η παραγοντική ανάλυση, θα πρέπει ο δείκτης KMO > 0,40 (Jolliffe, 1986).

β) την εξαγωγή των παραγόντων: Για τον καθορισμό του αριθμού των εξαγόμενων παραγόντων χρησιμοποιούνται διάφορες μέθοδοι για την απόκτηση των εκτιμητών των κύριων παραγόντων (Tabachnick and Fidell, 1996). Στην παρούσα έρευνα, χρησιμοποιείται η ανάλυση σε κύριες συνιστώσες. Σκοπός της ανάλυσης σε κύριες συνιστώσες είναι να δούμε κατά πόσο τα δεδομένα μπορούν να ερμηνευτούν με τη χρήση του ορθογώνιου παραγοντικού μοντέλου και κατά πόσο μπορούν να ερμηνευτούν οι παράγοντες που προκύπτουν.

γ) την περιστροφή των παραγόντων: Συχνά, οι μεταβλητές και οι παράγοντες δε φαίνονται να συσχετίζονται κατά τρόπο ερμηνεύσιμο, ιδιαίτερα όταν οι περισσότεροι παράγοντες συσχετίζονται με πολλές μεταβλητές και γι' αυτό επιχειρείται μέσω της περιστροφής των παραγόντων – δηλαδή των ορθογώνιων αξόνων- ο μετασχηματισμός του αρχικού πίνακα F των παραγόντων (συχνά σύνθετου).

Αρχικά, οι p μεταβλητές γράφονται ως γραμμικός συνδυασμός των k παραγόντων, δηλαδή

$$X - \mu = LF + \varepsilon$$

όπου

X είναι το διάνυσμα των αρχικών μεταβλητών μεγέθους $p \times 1$ (υποθέτω ότι έχω p μεταβλητές)

μ είναι το διάνυσμα των μέσων μεγέθους $p \times 1$

L είναι ένας πίνακας $p \times k$ όπου το L_{ij} είναι η επιβάρυνση του παράγοντα F_j στην μεταβλητή X_i

F είναι ένας $k \times 1$ πίνακας με τους παράγοντες

ε είναι το σφάλμα ή μοναδικός παράγοντας. Το σφάλμα ε_i είναι ο μοναδικός παράγοντας της i μεταβλητής και είναι το μέρος της μεταβλητής το οποίο δεν μπορεί να εξηγηθεί από τους παράγοντες.

Επίσης, κάθε μεταβλητή γράφεται στη μορφή:

$$X_1 = L_{11}F_1 + L_{12}F_2 + \dots + L_{1k}F_k + \varepsilon_1$$

$$X_2 = L_{21}F_1 + L_{22}F_2 + \dots + L_{2k}F_k + \varepsilon_2$$

...

$$X_p = L_{p1}F_1 + L_{p2}F_2 + \dots + L_{pk}F_k + \varepsilon_p$$

Η πιο γνωστή μέθοδος της ορθογωνικής περιστροφής είναι η μέθοδος της περιστροφής μέγιστης διασποράς (Varimax Rotation) με την οποία επιχειρείται να μεγιστοποιηθεί η διασπορά των τετραγώνων των παραγοντικών φορτίων και να ελαχιστοποιηθεί ο αριθμός των μεταβλητών με υψηλά παραγοντικά φορτία σε κάθε παράγοντα, που θα βοηθήσει έτσι στην ερμηνεία των παραγόντων.

δ) τον υπολογισμό παραγοντικών βαθμών: Ο υπολογισμός των παραγοντικών βαθμών (factor scores) για κάθε παρατήρηση βοηθά, μέσω του διαγράμματος διασποράς (score plot) των παρατηρήσεων ως προς δύο παράγοντες (συνήθως τους πρώτους), στον εντοπισμό σχέσεων των παρατηρήσεων.

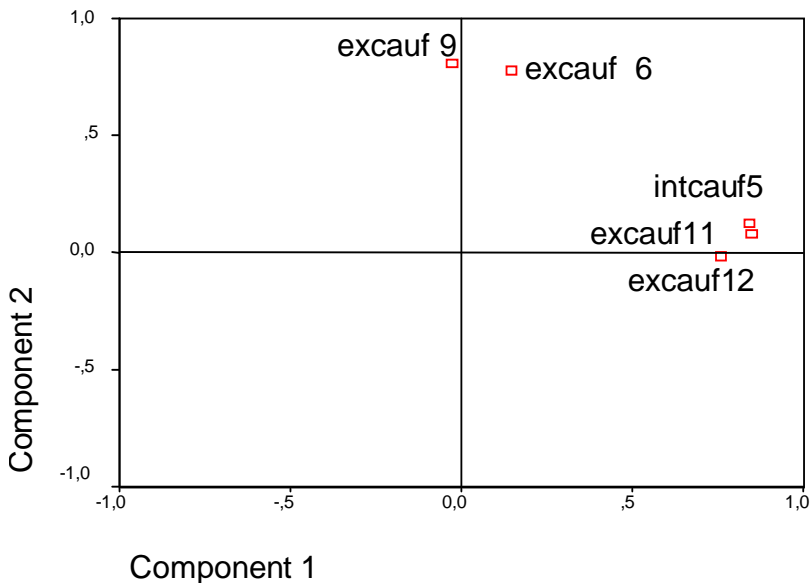
4. ΑΠΟΤΕΛΕΣΜΑΤΑ

Από την εφαρμογή της παραγοντικής ανάλυσης, προκύπτει ο Πίνακας των περιστρεφόμενων παραγόντων (Πίνακας 4) που περιέχει τις επιβαρύνσεις των παραγόντων μετά την περιστροφή. Η γραφική παράσταση των συνιστωσών των περιστρεφόμενων παραγόντων φαίνεται στο Σχήμα 1, στο οποίο οι μεταβλητές έχουν πλησιάσει τους δύο κάθετους άξονες.

Πίνακας 4. Πίνακας Περιστρεφόμενων Παραγόντων.

	Συνιστώσες	
	1	2
excauf 6: Πίεση από τη διοίκηση/τον εργοδηγό	0,147	0,778
excauf 9: Το θύμα ήταν νέο στη θέση εργασίας	-2,766E-02	0,810
excauf 11: Έλλειψη σωστού προστατευτικού εξοπλισμού	0,849	8,096E-02
excauf 12: Ασαφής και δύσκολη εργασία	0,758	-1,632E-02
intcauf 5: Μη χρήση μέσων ατομικής προστασίας	0,843	0,125

Σχήμα 1. Συνιστώσες Περιστρεφόμενων Παραγόντων.



Από το Σχήμα 1, φαίνεται ξεκάθαρα ότι ο Παράγοντας 1 ερμηνεύει τις μεταβλητές excauf11 (0,849), excauf12 (0,758) και intcauf5 (0,843), όπου είναι η «έλλειψη σωστού προστατευτικού εξοπλισμού», «η ασαφής και δύσκολη εργασία» και η «μη χρήση μέσων ατομικής προστασίας» αντίστοιχα. Ο Παράγοντας 2 ερμηνεύει τις μεταβλητές excauf6 (0,778), excauf9 (0,810), όπου είναι η «πίεση από τη διοίκηση/τον εργοδηγό» και «το θύμα ήταν νέο στη θέση εργασίας» αντίστοιχα.

Οι κύριοι παράγοντες που βρέθηκαν σε αυτή την έρευνα να επιδρούν στα ατυχήματα ήταν: «η έλλειψη επαγγελματικής κατάρτισης των εργαζομένων», «η έλλειψη σωστού προστατευτικού εξοπλισμού και επίβλεψης [των εργαζομένων] όσον αφορά στη χρήση του», «η πίεση στην απόδοση των εργαζομένων» καθώς και «η έλλειψη εμπειρίας του εργαζομένου» (Katsakiori et al., 2008).

5. ΣΥΜΠΕΡΑΣΜΑΤΑ

Από τα αποτελέσματα της ανάλυσης προέκυψε ότι κάθε Ε.Α. ήταν το αποτέλεσμα ενός συνόλου αιτιωδών παραγόντων, ιδιαίτερα εξωτερικών παραγόντων (excauf) (δηλ. περιβαλλοντικοί, οργανωτικοί και εργασιακοί παράγοντες).

Ο Παράγοντας 1 συνδέεται με τον προσανατολισμό (orientation) και την επαγγελματική κατάρτιση (job training) των εργαζομένων από τη Διοίκηση. Τα θέματα ασφάλειας είναι ένα από τα θέματα που πρέπει να περιλαμβάνονται στο πρόγραμμα εκπαίδευσης ενός Οργανισμού (Byars and Rue, 2000). Οι φτωχές πρακτικές εργασίας, που προκύπτουν από την έλλειψη προσανατολισμού και επαγγελματικής κατάρτισης από τον Οργανισμό, ήταν οι πιο κοινοί πρόδρομοι του ανθρώπινου λάθους που οδηγεί σε θανατηφόρα ατυχήματα, όπως έδειξε και η έρευνα από τους Feyer et al. (1991; 1997).

Η έλλειψη σωστού προστατευτικού εξοπλισμού (0,849) προκύπτει από τη σκόπιμη πολλές φορές απόφαση της Διοίκησης να παρέχει ένα ανασφαλές περιβάλλον για την εκτέλεση των εργασιών που δεν συμμορφώνεται με τις νομοθετικές απαιτήσεις. Η μη χρήση προστατευτικού εξοπλισμού (0,843) είναι μια μεμονωμένη πρακτική των εργαζομένων, η οποία είναι αναμενόμενη δεδομένου ότι λείπει ο σωστός προστατευτικός εξοπλισμός. Η διοικητική πρακτική και η μεμονωμένη πρακτική των εργαζομένων αλληλεπιδρούν με τη φύση της εργασίας, η οποία περιλαμβάνει ασαφείς και δύσκολους στόχους (0,758).

Ο Παράγοντας 2 εκφράζει τις διοικητικές πρακτικές που σχετίζονται με τις ιδιομορφίες του «προϊόντος». Στον κλάδο των οικοδομικών και τεχνικών έργων, το «προϊόν» είναι το συγκεκριμένο κατασκευαστικό έργο. Η ανησυχία για την εκτέλεση του έργου το συντομότερο δυνατό έγινε πολύ εμφανής τα χρόνια πριν την έναρξη των Ολυμπιακών Αγώνων λόγω της χρονικής πίεσης. Ο ψυχολογικός παράγοντας (η πίεση που ασκείται στον εργαζόμενο για να εκτελέσει την εργασία του όσο το δυνατόν γρηγορότερα) βρέθηκε να είναι αρκετά υψηλός (0,778), ο οποίος σε συνδυασμό με την απειρία του εργαζομένου στη συγκεκριμένη εργασία που βρέθηκε να είναι 0,810 συνέβαλαν αποφασιστικά στην εμφάνιση των ατυχημάτων. Αυτό το συμπέρασμα είναι σε συμφωνία με τις παρατηρήσεις των Wright (1986) και Wagenaar & Groeneweg (1987). Η πίεση στην απόδοση των εργαζομένων είναι μια διοικητική πρακτική. Όταν ένας εργαζόμενος είναι νέος στη θέση εργασίας, είναι περισσότερο εκτεθειμένος στα βασισμένα στη γνώση λάθη, τα οποία εμφανίζονται επειδή υπάρχει έλλειψη πείρας, κάτι που διαπίστωσαν και οι Feyer et al. (1997), δηλ. ότι οι διοικητικές πρακτικές συνδέονται με τα βασισμένα στη γνώση λάθη.

Από αυτή την έρευνα, σχετικά με τους παράγοντες πρόκλησης των Ε.Α., προέκυψαν τα παρακάτω:

➤ *επιβεβαιώθηκε η σημασία του προσανατολισμού και της επαγγελματικής κατάρτισης των εργαζομένων*

➤ *η παροχή σωστού προστατευτικού εξοπλισμού και επίβλεψης των εργαζομένων όσον αφορά στη χρήση του είναι ένα σημαντικό μέτρο που παράγει ευδιάκριτη βελτίωση στις συνθήκες εργασίας*

- η μείωση της πίεσης στην απόδοση των εργαζομένων είναι σημαντική
- η απειρία του εργαζομένου είναι επίσης σημαντική σε κάθε εργασιακό χώρο.

Τα ευρήματα της παρούσας έρευνας μπορούν να χρησιμοποιηθούν για τη βελτίωση των υπηρεσιών που παρέχονται από το Υπουργείο Εργασίας και Κοινωνικής Ασφάλισης όσον αφορά στο σχεδιασμό στρατηγικών πρόληψης. Για τους εργοδότες, η αποκάλυψη πολλαπλών παραγόντων πρόκλησης θα οδηγήσει στην πληρέστερη ενημέρωσή τους και στην υιοθέτηση συμπεριφορών πρόληψης.

ABSTRACT

The present study aims to identify occupational accident causation factors and reveals the presence of external causal variables (environmental, organizational and job factors) and internal causal variables (human and individual characteristics). The statistical factor analysis was performed with the use of SPSS and the method of principal component analysis (PCA) was applied. The results of the study concluded that each occupational accident was the result of a complex network of factors related to each other and not equivalent in causal importance, but these factors in combination contributed to accidents.

ΑΝΑΦΟΡΕΣ

- Κατσακιώρη, Π. (2010). Σχεδιασμός και ανάλυση μεθόδου διερεύνησης παραγόντων και της αλληλεπίδρασής τους στην πρόκληση των εργατικών ατυχημάτων. *Διδακτορική διατριβή, Τμήμα Μηχανολόγων & Αεροναυπηγών Μηχανικών, Πολυτεχνική Σχολή Πανεπιστημίου Πατρών.*
- Υπουργείο Εργασίας και Κοινωνικής Ασφάλισης (2004). Αναγγελθέντα εργατικά ατυχήματα 1999-2003. *Βιβλίο εργατικών ατυχημάτων Κέντρου Πρόληψης Επαγγελματικού Κινδύνου Ανατολικής Αττικής και Βορείου Αιγαίου.*
- Byars, L.L. and Rue, L.W. (2000). *Human Resource Management*, 6th ed., New York: McGraw-Hill.
- Feyer, A.M. and Williamson, A.M. (1991). A classification system for causes of occupational accidents for use in preventive strategies, *Scandinavian Journal of Work, Environment and Health*, **17**, 302-311.
- Feyer, A.M., Williamson, A.M. and Cairns, D.R. (1997). The involvement of human behavior in occupational accidents; errors in context, *Safety Science*, **25**, 55-65.
- Gyekye, S.A. (2003). Causal attributions of Ghanaian industrial workers for accident occurrence: Miners and non-miners perspective. *Journal of Safety Research*, **34**, 533-538.
- Jolliffe, I.T. (1986). *Principal Component Analysis*, Springer-Verlag, New York.
- Katsakiori, P., Manatakis, Em., Goutsos, S. and Athanassiou, G. (2008). Factors attributed to fatal occupational accidents in a period of five years preceding the Athens 2004 Olympic Games. *International Journal of Occupational Safety and Ergonomics*, **14**, 285-292.

- Moser, C.A. and Kalton, G. (1972). *Survey Methods in Social Investigation*, USA: Ashgate Publishing.
- Tabachnick, B.G. and Fidell, L.S. (1996). *Using Multivariate Statistics*, New York: Harper Collins College Publishers.
- Wagenaar, W.A. and Groeneweg, J. (1987). Accidents at sea: multiple causes and impossible consequences. *International Journal of Man-Machine Studies*, **27**, 587-598.
- Wright, C. (1986). Fatal accidents in the oil industry. *Sociological Review*, **4**, 265-289.

ΑΣΥΜΠΤΩΤΙΚΑ ΑΠΟΤΕΛΕΣΜΑΤΑ ΓΙΑ ΧΡΟΝΟΥΣ ΑΝΑΜΟΝΗΣ ΣΧΗΜΑΤΙΣΜΩΝ

Φ. Σ. Μηλιένος¹, Μ. Β. Κούτρας²

¹Τμήμα Στατιστικής και Αναλογιστικών-Χρηματοοικονομικών Μαθηματικών,
Πανεπιστήμιο Αιγαίου
milienos@aegean.gr

²Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης, Πανεπιστήμιο Πειραιά
mkoutras@unipi.gr

ΠΕΡΙΛΗΨΗ

Η μέθοδος της εμφύτευσης τυχαίων μεταβλητών σε Μαρκοβιανή αλυσίδα, έχει αποδειχθεί ένα αποτελεσματικό εργαλείο στην εύρεση ακριβών αποτελεσμάτων, για χρόνους αναμονής σχηματισμών. Με βάση τις ιδιότητες των πινάκων μετάβασης που προκύπτουν, για την υλοποίηση της παραπάνω τεχνικής, έχουν παρουσιαστεί στη βιβλιογραφία διάφορα ασυμπτωτικά αποτελέσματα, κάνοντας ως επί το πλείστον χρήση της ιδιοτιμής Perron-Frobenius. Στην εργασία αυτή εισάγονται κάποια νέα ασυμπτωτικά αποτελέσματα, τα οποία δε βασίζονται στην ιδιοτιμή Perron-Frobenius, αλλά στο ίχνος του πίνακα μετάβασης.

Λέξεις Κλειδιά: Εμφύτευση τυχαίων μεταβλητών σε Μαρκοβιανή αλυσίδα; χρόνοι αναμονής σχηματισμών; ροές επιτυχιών.

1. ΕΙΣΑΓΩΓΗ

Έστω Z_1, Z_2, \dots μια ακολουθία από διακριτές τυχαίες μεταβλητές (τ.μ.), ενώ με \mathcal{E} θα συμβολίζουμε ένα συγκεκριμένο σχηματισμό. Με τον όρο σχηματισμό εννοούμε μία ακολουθία από γράμματα προερχόμενα από το σύνολο τιμών των παραπάνω τ.μ. Για παράδειγμα, εάν οι $Z_i \in \{0, 1, 2\}, \forall i = 1, 2, \dots$ τότε μερικοί απλοί σχηματισμοί (ή αλλιώς, λέξεις) είναι οι $\mathcal{E} = 001, \mathcal{E} = 221$ κ.α. Σύνθετοι σχηματισμοί δημιουργούνται από την ένωση απλών σχηματισμών, π.χ. $\mathcal{E} = \{001, 111\}, \mathcal{E} = \{212, 221\}$ κ.α.

Ας συμβολίζουμε τώρα με X_n το πλήθος των εμφανίσεων του \mathcal{E} στην ακολουθία Z_1, Z_2, \dots, Z_n , για συγκεκριμένο $n \geq 1$, ενώ με T το χρόνο αναμονής μέχρι την πρώτη εμφάνιση του σχηματισμού \mathcal{E} , στην παραπάνω ακολουθία. Τότε εύκολα μπορούμε να διαπιστώσουμε ότι (σύμβαση: $\bar{F}(0) = 1$),

$$\bar{F}(n) = P(T > n) = P(X_n = 0), \quad n = 1, 2, \dots$$

Παράλληλα, είναι γνωστό ότι η απαριθμητρία τ.μ. X_n είναι συνήθως μία εμφυτεύσιμη τ.μ. σε Μαρκοβιανή αλυσίδα (βλ. π.χ. Fu and Lou (2003)). Τότε εύκολα μπορεί να

διαπιστωθεί ότι (βλ. και Fu and Lou (2003) ή Fu and Koutras (1994))

$$\bar{F}(n) = P(T > n) = P(X_n = 0) = \boldsymbol{\pi}_0 \boldsymbol{\Lambda}_0^n \mathbf{1}', \quad n = 0, 1, \dots \quad (1)$$

όπου $\boldsymbol{\Lambda}_0$ είναι ο άνω-αριστερός $s \times s$ υποπίνακας του πίνακα μετάβασης $\boldsymbol{\Lambda}$, που προκύπτει από την Μαρκοβιανή αλυσίδα που χρησιμοποιήθηκε για την εμφύτευση της X_n (υποθέτοντας ότι έχει $s + 1$ καταστάσεις), δηλαδή

$$\boldsymbol{\Lambda} = \begin{pmatrix} \boldsymbol{\Lambda}_0 & \mathbf{h}' \\ \mathbf{0} & 1 \end{pmatrix}$$

ενώ το διάνυσμα $\boldsymbol{\pi}_0$ περιέχει τις πρώτες s πιθανότητες του αντίστοιχου αρχικού διανύσματος $\boldsymbol{\pi}$ και \mathbf{h} είναι το μη μηδενικό διάνυσμα που προκύπτει από τη σχέση $\mathbf{h} = \mathbf{1}' - \boldsymbol{\Lambda}_0 \mathbf{1}'$ (με $\mathbf{1} = (1, 1, \dots, 1) \in \mathbb{R}^s$). Επομένως η (1) μας δίνει την πιθανότητα η Μαρκοβιανή αλυσίδα (με τις $s + 1$ καταστάσεις), να μην περάσει στην κατάσταση απορρόφησης, μετά από n βήματα. Έτσι ενώ ο ακριβής υπολογισμός των πιθανοτήτων \bar{F}_n , μπορεί να πραγματοποιηθεί μέσω της (1), πρόβλημα δημιουργείται όταν το n πάρει πολύ μεγάλες τιμές, καθώς ταυτόχρονα αυξάνονται και οι υπολογιστικές απαιτήσεις της συγκεκριμένης μεθόδου.

Στην περίπτωση αυτή, το ενδιαφέρον εστιάζεται στη μελέτη της ασυμπτωτικής συμπεριφοράς της \bar{F}_n , το οποίο αποτελεί και το αντικείμενο της παρούσας εργασίας. Συγκεκριμένα, η ασυμπτωτική συμπεριφορά της \bar{F}_n , μελετάται συναρτήσει του ίχνους του πίνακα $\boldsymbol{\Lambda}_0^n$, αντί της Perron-Frobenius ιδιοτιμής, που συνήθως συναντάμε στη βιβλιογραφία (βλ. π.χ. Fu *et al.* (2003) και Fu and Johnson (2009)).

2. ΚΥΡΙΑ ΑΠΟΤΕΛΕΣΜΑΤΑ

Αρχικώς, συμβολίζοντας με $\lambda_1, \lambda_2, \dots, \lambda_s$ τις s (όχι κατ' ανάγκην) διαφορετικές ιδιοτιμές του πίνακα $\boldsymbol{\Lambda}_0$, τότε το ίχνος του πίνακα $\boldsymbol{\Lambda}_0^n$, έστω τ_n , δίδεται από τη σχέση

$$\tau_n = \sum_{i=1}^s \lambda_i^n = \text{tr}(\boldsymbol{\Lambda}_0^n), \quad n = 1, 2, \dots$$

Επίσης, με ρ θα συμβολίζουμε την ακτίνα του φασματικού κύκλου του πίνακα $\boldsymbol{\Lambda}_0$ (δηλαδή, $\rho = \rho(\boldsymbol{\Lambda}_0) = \max_{1 \leq i \leq s} |\lambda_i|$). Χρησιμοποιώντας τη σχέση

$$\sum_{n=0}^{\infty} (w \boldsymbol{\Lambda}_0)^n = (\mathbf{I} - w \boldsymbol{\Lambda}_0)^{-1}, \quad |w| < 1/\rho,$$

μπορούμε να δούμε ότι η γεννήτρια συνάρτηση των $\bar{F}_n, n \geq 0$ γίνεται (βλ. (1))

$$G(w) = \sum_{n=0}^{\infty} \bar{F}_n w^n = \boldsymbol{\pi}_0 (\mathbf{I} - w \boldsymbol{\Lambda}_0)^{-1} \mathbf{1}', \quad |w| < 1/\rho. \quad (2)$$

Ας υποθέσουμε τώρα ότι ο πίνακας $\mathbf{\Lambda}_0$ έχει m διαφορετικές ιδιοτιμές, δηλαδή το φάσμα του $\mathbf{\Lambda}_0$ είναι το

$$\sigma(\mathbf{\Lambda}_0) = \{\lambda_1, \lambda_2, \dots, \lambda_m\}$$

όπου $m \leq s$ και πλέον, $|\lambda_1| \geq |\lambda_2| \geq \dots \geq |\lambda_m|$. Δηλαδή, έχουμε αναδιατάξει τον συμβολισμό των ιδιοτιμών $\lambda_1, \lambda_2, \dots, \lambda_s$, έτσι ώστε από εδώ και στο εξής οι m διαφορετικές ιδιοτιμές να παίρνουν τους δείκτες $1, 2, \dots, m$. Επιπλέον οι m διαφορετικές ιδιοτιμές είναι διατεταγμένες σε φθίνουσα σειρά με βάση το μέτρο τους, με την έννοια ότι: $|\lambda_i| \geq |\lambda_j|$ αν και μόνο αν $i \leq j$. Με $u_i, i = 1, 2, \dots, m$ θα συμβολίζουμε τις πολλαπλότητες των ιδιοτιμών $\lambda_i, i = 1, 2, \dots, m$ και θα θεωρούμε ότι ανάμεσα στις ιδιοτιμές με ίδιο μέτρο, υιοθετούμε μια φθίνουσα διάταξη με βάση την πολλαπλότητα τους (δηλαδή, εάν $|\lambda_i| = |\lambda_j|$ και $i < j$, τότε $u_i \geq u_j$).

Το επόμενο λήμμα θα χρησιμοποιηθεί στη συνέχεια ώστε να μελετηθεί η ασυμπτωτική συμπεριφορά των πιθανοτήτων $\bar{F}_n, n \geq 1$.

Λήμμα 2..1. *Εάν η ιδιοτιμή λ_1 του υποστοχαστικού πίνακα $\mathbf{\Lambda}_0$ έχει πολλαπλότητα ίση με ένα ($u_1 = 1$), τότε για ικανοποιητικά μεγάλο n , υπάρχει σταθερά θ τέτοια ώστε*

$$\bar{F}_n - \theta \tau_n = \sum_{i=2}^m p_i(n) \lambda_i^n \quad (3)$$

όπου τα $p_i(n)$ είναι πολυώνυμα του n , βαθμού το πολύ $u_i - 1, i = 2, 3, \dots, m$.

Απόδειξη. Έστω ότι $c_{ij}(w) = (-1)^{i+j} \det(M_{ij}(w))$, με $i, j \in \{1, 2, \dots, s\}$, είναι οι συμπαραγοντες (cofactors) του πίνακα $\mathbf{I} - w\mathbf{\Lambda}_0$ (άρα, οι $c_{ij}(w)$ θα είναι πολυώνυμα του w βαθμού το πολύ $s - 1$). Τότε

$$(\mathbf{I} - w\mathbf{\Lambda}_0)^{-1} = \frac{1}{\det(\mathbf{I} - w\mathbf{\Lambda}_0)} \text{adj}(\mathbf{I} - w\mathbf{\Lambda}_0)$$

όπου $\text{adj}(\mathbf{I} - w\mathbf{\Lambda}_0) = (c_{ij}(w))' = (c_{ji}(w))_{s \times s}$ είναι ο προσαρτημένος (adjoint) πίνακας του $\mathbf{I} - w\mathbf{\Lambda}_0$. Η (2) θα πάρει τότε τη μορφή

$$G(w) = \frac{\boldsymbol{\pi}_0 \text{adj}(\mathbf{I} - w\mathbf{\Lambda}_0) \mathbf{1}'}{\det(\mathbf{I} - w\mathbf{\Lambda}_0)} = \frac{\sum_{i=1}^s \sum_{j=1}^s \pi_{0j} c_{ij}(w)}{\det(\mathbf{I} - w\mathbf{\Lambda}_0)},$$

με $\pi_{0j}, j = 1, 2, \dots, s$ να είναι οι s πρώτες συντεταγμένες του διανύσματος $\boldsymbol{\pi}_0$. Παράλληλα, αφού $\det(\mathbf{I} - w\mathbf{\Lambda}_0) = \prod_{i=1}^m (1 - w\lambda_i)^{u_i}$, συμπεραίνουμε ότι η $G(w)$ (ως λόγος πολυωνύμων του w) ικανοποιεί την παρακάτω ανάλυση σε μερικά κλάσματα (partial fraction decomposition)

$$G(w) = \frac{\sum_{i=1}^s \sum_{j=1}^s \pi_{0j} c_{ij}(w)}{\det(\mathbf{I} - w\mathbf{\Lambda}_0)} = \sum_{i=1}^m \sum_{v=1}^{u_i} \frac{d_{iv}}{(1 - w\lambda_i)^v} + R(w) \quad (4)$$

με τα d_{iv} να είναι σταθερές ανεξάρτητες του n και $R(w)$ ένα πολυώνυμο του w . Να σημειώσουμε ότι οι σταθερές d_{iv} που αντιστοιχούν σε πραγματικές ιδιοτιμές είναι πραγματικές, ενώ αυτές που αντιστοιχούν σε μιγαδικές ιδιοτιμές, είναι εν γένει μιγαδικές. Αναπτύσσοντας τους όρους $(1 - w\lambda_i)^v$ καταλήγουμε στην εξής μορφή

$$G(w) = \sum_{n=0}^{\infty} \bar{F}_n w^n = \sum_{n=0}^{\infty} \left(\sum_{i=1}^m \sum_{v=1}^{u_i} \binom{n+v-1}{n} d_{iv} \lambda_i^n \right) w^n + R(w).$$

Για n μεγαλύτερο από το βαθμό του $R(w)$, μπορούμε εύκολα να πάρουμε ότι (καθώς $u_1 = 1$)

$$\bar{F}_n = \sum_{i=1}^m \sum_{v=1}^{u_i} \binom{n+v-1}{n} d_{iv} \lambda_i^n = d_{11} \lambda_1^n + \sum_{i=2}^m \sum_{v=1}^{u_i} \binom{n+v-1}{n} d_{iv} \lambda_i^n. \quad (5)$$

Η απόδειξη ολοκληρώνεται θέτοντας $\theta = d_{11}$ και βλέποντας ότι

$$\bar{F}_n - \theta \tau_n = \sum_{i=2}^m \left(\sum_{v=1}^{u_i} \binom{n+v-1}{n} d_{iv} - u_i d_{11} \right) \lambda_i^n = \sum_{i=2}^m p_i(n) \lambda_i^n \quad (6)$$

όπου

$$p_i(n) = \sum_{v=1}^{u_i} \binom{n+v-1}{n} d_{iv} - u_i d_{11} = \sum_{v=1}^{u_i} \left(\prod_{j=1}^{v-1} (n+j) \right) \frac{d_{iv}}{(v-1)!} - u_i d_{11}$$

είναι πολυώνυμο του n , βαθμού το πολύ $u_i - 1$, για $i = 2, 3, \dots, m$. □

Στο σημείο αυτό πρέπει να σημειώσουμε ότι η (3) δε θα έχει πρακτική σημασία εάν το θ είναι ένας (καθαρός) μιγαδικός αριθμός. Επομένως, θα εξετάσουμε στη συνέχεια τις περιπτώσεις που το θ γίνεται ένας πραγματικός αριθμός και από μια επιπλέον κανονικοποίηση της ακολουθίας $\bar{F}_n - \theta \tau_n$, θα οδηγούμαστε στην ασυμπτωτική μορφή των πιθανοτήτων \bar{F}_n . Το παρακάτω αποτέλεσμα είναι της παραπάνω φύσεως.

Πρόταση 2..1. Έστω ότι ο πίνακας $\mathbf{\Lambda}_0$ έχει μόνο μια ιδιοτιμή λ_1 στον φασματικό του κύκλο, πολλαπλότητας ίσης με ένα. Τότε, υπάρχει μια πραγματική μη-αρνητική σταθερά θ τέτοια ώστε

$$\bar{F}_n = \theta \tau_n + O(|\lambda_2|^n n^{u_2-1}).$$

Απόδειξη. Από τη θεωρία γνωρίζουμε ότι η ακτίνα ενός μη-αρνητικού πίνακα, ανήκει στο φάσμα του πίνακα (βλ. π.χ. Meyer (2000)). Έτσι, $\rho \in \sigma(\mathbf{\Lambda}_0) = \{\lambda_1, \lambda_2, \dots, \lambda_m\}$ και επειδή έχουμε επιπλέον υποθέσει ότι ο $\mathbf{\Lambda}_0$ έχει μόνο μια ιδιοτιμή στον φασματικό του κύκλο, καταλήγουμε ότι $\rho = \lambda_1 > 0$ και $\rho = \lambda_1 = |\lambda_1| > |\lambda_2| \geq \dots \geq |\lambda_m|$ (η περίπτωση $\rho = \lambda_1 = 0$, αποκλείεται από τις αποθέσεις της προτάσεως). Εάν $\theta = d_{11}$

είναι η σταθερά του Λήμματος 2..1, τότε επειδή $\lambda_1 \in \mathfrak{R}$ έχουμε $\theta \in \mathfrak{R}$. Επίσης, καθώς $u_1 = 1$, η (5) μας οδηγεί στη σχέση

$$\theta + \sum_{i=2}^m \sum_{v=1}^{u_i} \binom{n+v-1}{n} d_{iv} (\lambda_i/\lambda_1)^n = \bar{F}_n/\lambda_1^n \geq 0$$

(για n μεγαλύτερο από το βαθμό του $R(w)$). Για $n \rightarrow \infty$ η τελευταία ανισότητα μας δίνει ότι το θ είναι μια μη-αρνητική σταθερά. Από την (6) και για $|\lambda_2| \neq 0$, παίρνουμε

$$\left| \frac{\bar{F}_n - \theta \tau_n}{\lambda_2^n n^{u_2-1}} \right| \leq \sum_{i=2}^m \left| \frac{p_i(n)}{n^{u_2-1}} \left(\frac{\lambda_i}{\lambda_2} \right)^n \right| \quad (7)$$

και δεν είναι δύσκολο να εξακριβώσουμε ότι το δεξιο μέλος της προηγούμενης ανισότητας, συγκλίνει σε ένα μη-αρνητικό πραγματικό αριθμό, όταν $n \rightarrow \infty$. Συγκεκριμένα, εάν $|\lambda_2| > |\lambda_3| \geq \dots \geq |\lambda_m|$ τότε

$$\lim_{n \rightarrow \infty} \left| p_i(n) \left(\frac{\lambda_i}{\lambda_2} \right)^n \right| = 0 \Rightarrow \lim_{n \rightarrow \infty} \left| \frac{p_i(n)}{n^{u_2-1}} \left(\frac{\lambda_i}{\lambda_2} \right)^n \right| = 0$$

για κάθε $i = 3, 4, \dots, m$, ενώ εάν $|\lambda_2| = |\lambda_3| = \dots = |\lambda_l| > |\lambda_{l+1}| \geq \dots \geq |\lambda_m|$ ($2 \leq l < m$) ισχύει

$$\sum_{i=2}^m \left| \frac{p_i(n)}{n^{u_2-1}} \left(\frac{\lambda_i}{\lambda_2} \right)^n \right| = \sum_{i=2}^l \left| \frac{p_i(n)}{n^{u_2-1}} \right| + \sum_{i=l+1}^m \left| \frac{p_i(n)}{n^{u_2-1}} \left(\frac{\lambda_i}{\lambda_2} \right)^n \right|$$

με το δεύτερο άθροισμα (στο δεξιό μέλος της προηγούμενης ανισότητας) να συγκλίνει στο μηδέν, ενώ το πρώτο σε μια σταθερά. Αφού η ακολουθία $\left| \frac{\bar{F}_n - \theta \tau_n}{\lambda_2^n n^{u_2-1}} \right|$ είναι φραγμένη όταν το $n \rightarrow \infty$ και οι $\bar{F}_n - \theta \tau_n$, $|\lambda_2|^n n^{u_2-1}$ είναι πραγματικοί αριθμοί, τότε

$$\bar{F}_n - \theta \tau_n = O(|\lambda_2|^n n^{u_2-1})$$

και η απόδειξη ολοκληρώνεται, αφού παρατηρήσουμε ότι οδηγούμαστε πάλι στο παραπάνω αποτέλεσμα, και στην περίπτωση $|\lambda_2| = 0$ (βλ. (3)). \square

Αξίζει να σημειώσουμε ότι επειδή ο πίνακας $\mathbf{\Lambda}_0$ είναι υποστοχαστικός, έχουμε $0 < \rho = \rho(\mathbf{\Lambda}_0) \leq 1$, ενώ αυτό που συνήθως συναντάμε είναι ότι $\rho < 1$. Παράλληλα, εάν $|\lambda_1| > |\lambda_2|$, η Πρόταση 2..1 μας δίνει ότι

$$\lim_{n \rightarrow \infty} \frac{\bar{F}_n}{\theta \tau_n} = 1 \quad (8)$$

(για $\theta \tau_n \neq 0$) και επομένως, οι πιθανότητες \bar{F}_n μπορούν ικανοποιητικά να προσεγγιστούν από τη σχέση

$$\bar{F}_n \approx \theta \tau_n. \quad (9)$$

Μια κλάση υποστοχαστικών πινάκων που ικανοποιεί τις προϋποθέσεις της Πρότασης 2..1 είναι οι αρχικοί (primitive) πίνακες. Ένας μη-αρνητικός τετραγωνικός πίνακας \mathbf{A} καλείται αρχικός εάν υπάρχει θετικός ακέραιος n τέτοιος ώστε $\mathbf{A}^n > \mathbf{0}$ (δηλαδή, όλα τα στοιχεία του πίνακα \mathbf{A}^n είναι θετικοί αριθμοί). Από τη θεωρία των Perron-Frobenius (βλ. π.χ. Seneta (1981)) εάν ο \mathbf{A} είναι αρχικός πίνακας, τότε υπάρχει ιδιοτιμή ρ (η ιδιοτιμή Perron-Frobenius) του \mathbf{A} τέτοια ώστε: α) η ρ είναι ένας πραγματικός θετικός αριθμός, β) $\rho > |\lambda|$ για κάθε ιδιοτιμή $\lambda \neq \rho$ και γ) η πολλαπλότητα της ρ είναι ίση με ένα. Το επόμενο αποτέλεσμα αναφέρεται στην κλάση των αρχικών πινάκων.

Πόρισμα 2..1. *Εάν ο πίνακας $\mathbf{\Lambda}_0$ είναι αρχικός, τότε υπάρχει θετική πραγματική σταθερά θ τέτοια ώστε*

$$\bar{F}_n = \theta \tau_n + O(|\lambda_2|^n n^{u_2-1}).$$

Απόδειξη. Αφού ο υποστοχαστικός πίνακας $\mathbf{\Lambda}_0$ είναι αρχικός, έχουμε: $|\lambda_1| = \lambda_1 > |\lambda_2| \geq \dots \geq |\lambda_m| \geq 0$, $u_1 = 1$. Επομένως, η Πρόταση 2..1 μας οδηγεί στο επιθυμητό αποτέλεσμα, όμως για $\theta = d_{11} \geq 0$. Αυτό που πρέπει επιπλέον να αποδείξουμε είναι ότι η σταθερά $\theta = d_{11}$ είναι θετική, και για το λόγο αυτό θα χρησιμοποιήσουμε την παρακάτω σχέση (βλ. π.χ. Seneta (1981))

$$\frac{\text{adj}(\mathbf{I} - (1/\rho)\mathbf{\Lambda}_0)}{\prod_{i=2}^m (1 - (1/\rho)\lambda_i)^{u_i}} = \mathbf{v}'_2 \mathbf{v}_1 \quad (10)$$

όπου $\mathbf{v}_1, \mathbf{v}_2$ είναι το αριστερό και δεξιό θετικό ιδιοδιάνυσμα της $\rho = \lambda_1$, αντιστοίχως (τα οποία πάντα υπάρχουν). Λαμβάνοντας υπόψη ότι ισχύει $u_1 = 1$ και η ανάλυση σε μερικά κλάσματα (4), παίρνουμε

$$d_{11} = \lim_{w \rightarrow \rho^{-1}} \left[\frac{\pi_0 \text{adj}(\mathbf{I} - w\mathbf{\Lambda}_0) \mathbf{1}'}{\prod_{i=1}^m (1 - w\lambda_i)^{u_i}} (1 - w\rho) \right] = \pi_0 \mathbf{v}'_2 \mathbf{v}_1 \mathbf{1}' > 0.$$

και η απόδειξη έχει ολοκληρωθεί. □

Το επόμενο αποτέλεσμα, καλύπτει μια ευρύτερη κλάση πινάκων από αυτή των αρχικών, τους αδιαχώριστους (irreducible) πίνακες (για τον ορισμό και τις ιδιότητες των αδιαχώριστων πινάκων, βλ. π.χ. Meyer (2000)).

Πρόταση 2..2. *Εάν ο πίνακας $\mathbf{\Lambda}_0$ είναι αδιαχώριστος, τότε υπάρχει θετική πραγματική σταθερά θ τέτοια ώστε*

$$\bar{F}_n = \theta \tau_n + O(\rho^n).$$

Απόδειξη. Επειδή ο πίνακας $\mathbf{\Lambda}_0$ είναι αδιαχώριστος, όλες οι ιδιοτιμές που βρίσκονται πάνω στον φασματικό του κύκλο έχουν πολλαπλότητα ίση με 1 (βλ. π.χ. Meyer (2000)). Αν η λ_1 είναι η μοναδική ιδιοτιμή του $\mathbf{\Lambda}_0$ στον φασματικό του κύκλο, τότε

από την Πρόταση 2..1 ή το Πόρισμα 2..1 (ένας μη-αρνητικός αδιαχώριστος πίνακας με μόνο μια ιδιοτιμή στον φασματικό του κύκλο, είναι αρχικός) καταλήγουμε στο επιθυμητό αποτέλεσμα. Ας εξετάσουμε τώρα την περίπτωση

$$\rho = |\lambda_1| = |\lambda_2| = \dots = |\lambda_l| > |\lambda_{l+1}| \geq \dots \geq |\lambda_m|$$

με $u_1 = u_2 = \dots = u_l = 1$ ($2 \leq l \leq m$). Δεδομένου ότι $u_1 = 1$, από το Λήμμα 2..1 με $\theta = d_{11}$ (χ.π.τ.γ. $\rho = \lambda_1$) παίρνουμε την (6), με τα $p_i(n), i = 1, 2, \dots, l$ να είναι πολυώνυμα βαθμού $u_1 - 1 = 0$ (δηλαδή, $p_i(n) = \alpha_i$). Η ανισότητα (7) γίνεται

$$|(\bar{F}_n - \theta\tau_n)/\rho^n| \leq \sum_{i=1}^l |\alpha_i| + \sum_{i=l+1}^m |p_i(n) (\lambda_i/\rho)^n|.$$

Το δεξιό μέλος συγκλίνει στο $\sum_{i=1}^l |\alpha_i|$, ενώ η $\theta = d_{11}$ αποδεικνύεται ότι είναι μια θετική σταθερά, με επιχειρήματα παρόμοια με αυτά της Προτάσεως 2..1 (η (10) ισχύει και για μη-αρνητικούς αδιαχώριστους πίνακες, βλ. π.χ. Seneta (1981)). \square

Αξίζει επίσης να αναφέρουμε ότι αν $\lambda_1 > |\lambda_i|$, για $i = 2, 3, \dots, m$ τότε $\tau_n/\lambda_1^n \rightarrow 1$, καθώς $n \rightarrow \infty$ και τα ασυμπτωτικά αποτελέσματα (8) και (9) μπορεί να διατυπωθούν και ως

$$\lim_{n \rightarrow \infty} \bar{F}_n/\theta\lambda_1^n = 1 \text{ και } \bar{F}_n \approx \theta\lambda_1^n.$$

Τα αποτελέσματα αυτά τα συναντάμε ήδη στη βιβλιογραφία π.χ. στη θεωρία αξιολοπιστίας (βλ. π.χ. Boutsikas and Koutras (2001)), ή σε προβλήματα ροών επιτυχιών (βλ. π.χ. Fu *et al.* (2003)). Επομένως, η Πρόταση 2..1 και το Πόρισμα 2..1 θα μπορούσαν επίσης να διατυπωθούν με χρήση της Perron-Frobenius ιδιοτιμής.

Πίνακας 1: Χρόνος αναμονής μέχρι την εμφάνιση ροής μήκους k , σε μια Bernoulli ακολουθία (με πιθανότητα επιτυχίας p).

$k = 25$, με $p = 0.995, \theta = 8.1538$				
n	τ_n	\bar{F}_n	$\theta\tau_n$	$\theta\lambda_1^n$
30	0.007370	0.095724	0.060096	0.168060
40	0.005534	0.051613	0.045129	0.044068
45	0.003835	0.029558	0.031274	0.022566
65	0.000218	0.001752	0.001774	0.001552

Η αριθμητική μελέτη των προσεγγίσεων μας φανέρωσε (βλ. ενδεικτικά και Πίνακα 1) ότι οι προσεγγίσεις μέσω του ίχνους είναι, πολλές φορές, ακριβέστερες σε σχέση με αυτές που προκύπτουν από την Perron-Frobenius ιδιοτιμή. Είναι γεγονός ότι με το ίχνος του πίνακα $\mathbf{\Lambda}^n$ χρησιμοποιείται περισσότερη πληροφορία (και ίσως εκεί να οφείλονται οι καλύτερες προσεγγίσεις), ενώ θεωρητικώς, ο υπολογισμός του τ_n έχει τις ίδιες υπολογιστικές απαιτήσεις με τον υπολογισμό της Perron-Frobenius ιδιοτιμής. Συγκεκριμένα, γνωρίζοντας το χαρακτηριστικό πολυώνυμο του $\mathbf{\Lambda}_0$ μπορούμε να

χρησιμοποιήσουμε τις Newton ταυτότητες (βλ. π.χ. Meyer (2000)) και μέσα από μια αναδρομική σχέση, να προκύψουν οι τιμές τ_n , για κάθε n .

Τέλος, ανατρέχοντας στη βιβλιογραφία (βλ. π.χ. Fu and Lou (2003)) διαπιστώνουμε ότι τα υπάρχοντα και παρόμοια μορφής προσεγγιστικά αποτελέσματα, αναφέρονται κυρίως σε αρχικούς πίνακες. Έτσι η χρησιμότητα των παραπάνω νέων αποτελεσμάτων αναδεικνύεται και από το γεγονός ότι σε πολλές εφαρμογές, ο πίνακας \mathbf{A}_0 δεν είναι αρχικός. Για παράδειγμα, ο πίνακας που αντιστοιχεί στον χρόνο αναμονής μέχρι την τοποθέτηση σφαιριδίων, σε όλα τα κελιά, σε ένα κλασικό μοντέλο καταλήψεων (βλ. Fu and Lou (2003), σελ. 7), έχει μια ιδιοτιμή στον φασματικό του κύκλο χωρίς να είναι αρχικός (ούτε καν αδιαχώριστος). Επιπλέον, σε πίνακα αδιαχώριστο και όχι αρχικό μπορούμε να καταλήξουμε π.χ. και μέσα από τη θεωρία των τεστ εκκίνησης (βλ. π.χ. Antzoulakos et al (2009)).

ABSTRACT

It is known that the finite Markov chain imbedding technique is a useful tool, for the evaluation of the exact distribution of a pattern waiting time. A number of asymptotic results, based on the properties of the deduced transition probability matrix and the existence of the Perron-Frobenius eigenvalue, have also appeared in the literature. In this article we present asymptotic results that do not require the existence of the Perron-Frobenius eigenvalue, but they are based on the trace of the transition probability matrix.

ΑΝΑΦΟΡΕΣ

- Antzoulakos, D.L., Koutras, M.V. and Rakitzis, A.C. (2009). Start-up demonstration tests based on run and scan statistics. *Journal of Quality Technology*, **41**, 48-59.
- Boutsikas, M.V. and Koutras, M.V. (2001). Reliability approximation for Markov chain imbeddable systems. *Methodology and Computing in Applied Probability*, **2**, 393-411.
- Fu, J.C. and Johnson, B.C. (2009). Approximate probabilities for runs and patterns in i.i.d. and Markov-dependent multistate trials. *Advances in Applied Probability*, **41**, 292-308.
- Fu, J.C. and Koutras M.V. (1994). Distribution theory of runs: a Markov chain approach. *Journal of the American Statistical Association*, **89**, 1050-1058.
- Fu, J.C. and Lou, W.Y.W. (2003). *Distribution Theory of Runs and Patterns and its Applications: A Finite Markov Chain Imbedding Approach*. World Scientific, Singapore.

- Fu, J.C., Wang, L. and Lou, W.Y.W. (2003). On exact and large deviation approximation for the distribution of the longest run in a sequence of two-state Markov dependent trials. *Journal of Applied Probability*, **40**, 346-360.
- Meyer, C. (2000). *Matrix Analysis and Applied Linear Algebra*. Siam, Philadelphia.
- Seneta, E. (1981). *Nonnegative Matrices and Markov chains*. 2nd Edn, Springer, New York.



Εκτίμηση της Συνάρτησης Επιβίωσης με χρήση Μονότονων Κυβικών Spline

Λεωνίδα Μπαντής, Γιάννης Τσιμής, Στέλιος Γεωργίου
Πανεπιστήμιο Αιγαίου, Τμ. Στατιστικής & Αναλογιστικών - Χρημ/κών Μαθηματικών
{lbantis, tsimikas, stgeorgiou}@aegean.gr

ΠΕΡΙΛΗΨΗ

Σε αυτήν την εργασία ασχολούμαστε με την εκτίμηση της συνάρτησης επιβίωσης παρουσία λογοκρίσιμης. Εξετάζουμε την προσαρμογή μιας φυσικής κυβικής spline στην αθροιστική συνάρτηση κινδύνου κάτω από κατάλληλους περιορισμούς. Με την προτεινόμενη μέθοδο το πρόβλημα ανάγεται σε ελαχιστοποίηση αθροίσματος τετραγώνων κάτω από γραμμικούς περιορισμούς. Η προτεινόμενη μέθοδος συγκρίνεται μέσω προσομοιώσεων με άλλες γνωστές μεθόδους όπως η μη παραμετρική των Kaplan Meier και η logspline. Στην περίπτωση που υπάρχουν συμμεταβλητές, η εκτίμηση της συνάρτησης επιβίωσης γίνεται μέσω του μοντέλου του Cox. Εφαρμόζουμε τη μέθοδο σε πραγματικά δεδομένα που αφορούν σε καρκίνο του στομάχου και του λάρυγγα.

Λέξεις Κλειδιά: Φυσική Spline, Συνάρτηση Επιβίωσης, Αθροιστική Συνάρτηση Κινδύνου, Μονότονη Spline.

1. ΕΙΣΑΓΩΓΗ

Υπάρχουν πολλές μέθοδοι λείανσης της συνάρτησης επιβίωσης, S , ή ισοδύναμα της αθροιστικής συνάρτησης κατανομής, F . Δύο συχνά χρησιμοποιούμενες μέθοδοι είναι η μη παραμετρική μέθοδος μέγιστης πιθανοφάνειας των Kaplan και Meier (1958) (KM) καθώς και τα ευρέως χρησιμοποιούμενα παραμετρικά μοντέλα. Τα παραμετρικά μοντέλα βασίζονται σε πολύ ισχυρές υποθέσεις ενώ ο μη παραμετρικός εκτιμητής δε μπορεί να παρέχει εκτίμηση πέραν του τελευταίου γεγονότος.

Υπάρχουν κάποιες στρατηγικές που έχουν προταθεί για την αντιμετώπιση αυτού του μειονεκτήματος. Ο Efron (1967) προτείνει να θέτουμε την εκτίμηση της συνάρτησης επιβίωσης πέραν του χρόνου του τελευταίου γεγονότος, t_{max} , ίση με μηδέν, ενώ ο Gill (1980) προτείνει να θέτουμε αντίστοιχα την εκτιμητρία συνάρτηση επιβίωσης $\hat{S}(t) = \hat{S}(t_{max}), \forall t \geq t_{max}$. Μία άλλη προσέγγιση συμπλήρωσης της ουράς της συνάρτησης επιβίωσης ευέλικτων μοντέλων όπως το Weibull προτάθηκε από τους Moeschberger και Klein (1985). Ωστόσο, προσομοιώσεις έχουν δείξει ότι μια λεία εκτίμηση της συνάρτησης επιβίωσης είναι προτιμότερη από τη μη παραμετρική

βηματική εκτίμηση (Pan (2000)).

Οι πιο γνωστές μέθοδοι λείανσης είναι αυτές που βασίζονται στη χρήση πυρήνων, στις splines, και στα συνήθη παραμετρικά μοντέλα. Τα τελευταία δε θα συζητηθούν περαιτέρω σε αυτήν την εργασία μιας και βασίζονται σε πολύ ισχυρές υποθέσεις. Η μεθοδολογία των πυρήνων περιγράφεται λεπτομερώς από τους Wand και Jones (1995) και Bowman και Azzalini (1997) στην περίπτωση που υπάρχουν λογοκριμένα δεδομένα. Ωστόσο οι μέθοδοι που βασίζονται στους πυρήνες επίσης αδυνατούν να παρέχουν εκτίμηση πέραν του τελευταίου γεγονότος.

2. LOGSPLINE

Στην εργασία των Kooperberg και Stone (1992) μελετώνται logpline μοντέλα στην περίπτωση των λογοκριμένων δειγμάτων. Εδώ σύντομα περιγράφουμε αυτή τη μεθοδολογία.

Έστω ότι τα διαθέσιμα προς ανάλυση δεδομένα είναι της μορφής (T_i, D_i) , $i = 1, \dots, n$, όπου $T_i = \min(X_i, C_i)$, X_i είναι ο χρόνος μέχρι το γεγονός, C_i είναι η τ.μ. λογοκρισίας, και Δ_i είναι μια δίτιμη δείκτρια μεταβλητή που παίρνει την τιμή 1 αν έχουμε γεγονός και 0 αν έχουμε λογοκρισία, δηλαδή $\Delta_i = I(X_i < C_i)$. Έστω ο φυσικός αριθμός $K \geq 3$ και οι κόμβοι (knots) τ_1, \dots, τ_K με $-\infty \leq L < \tau_1 < \dots < \tau_K \leq U$ όπου L και U ορίζουν το στήριγμα της κατανομής.

Το μοντέλο logspline ορίζεται ως εξής

$$f(x; \theta) = \exp(\theta_1 B_1(x) + \dots + \theta_p B_p(x) - C(\theta)), L < x < U, \quad (1)$$

όπου $C(\theta) = \log(\int_L^U \exp(\theta_1 B_1(x) + \dots + \theta_p B_p(x)) dx)$ είναι η σταθερά κανονικοποίησης και οι συναρτήσεις $B_1(x), B_2(x), \dots, B_p(x)$ μπορούν να επιλεγθούν ώστε η spline να είναι γραμμική πριν από τον πρώτο και μετά από τον τελευταίο κόμβο, ενώ σε κάθε ένα από τα ενδιάμεσα διαστήματα $[\tau_1, \tau_2], \dots, [\tau_{K-1}, \tau_K]$ να έχουμε ένα κυβικό πολυώνυμο. Σε αυτήν την εργασία θεωρούμε ότι το στήριγμα της κατανομής είναι θετικό ($L = 0, U = \infty$). Κάτω από την υπόθεση ότι οι τ.μ. C και X είναι ανεξάρτητες, τότε η προς μεγιστοποίηση \log -πιθανοφάνεια ως προς το διάνυσμα παραμέτρων είναι κοίλη και το μέγιστο της πιθανοφάνειας είναι μοναδικό. Ωστόσο, οι συγγραφείς αναφέρουν ότι στην περίπτωση που υπάρχουν λογοκρισίες αυτό δεν ισχύει.

3. Μονότονη φυσική spline

3.1 Κατασκευάζοντας το μοντέλο

Σε αυτήν την ενότητα θεωρούμε μια μονότονη φυσική spline για λείανση της αθροιστικής συνάρτησης κινδύνου. Η KM εκτιμήτρια της συνάρτησης κινδύνου δίνεται από

$$\hat{H}^{KM}(x) = -\log(\hat{S}^{KM}(x)).$$

Έστω οι κόμβοι $\tau_1 < \dots < \tau_K$ και η φυσική spline

$$H(x) = \beta_1 W_1(x) + \beta_2 W_2(x) + \dots + \beta_{K-2} W_{K-2}(x), \quad (2)$$

όπου για $j = 1, \dots, K - 2$ έχουμε

$$W_j(x) = (x - \tau_j)_+^3 - \frac{(x - \tau_{K-1})_+^3 (\tau_K - \tau_j)}{\tau_K - \tau_{K-1}} + \frac{(x - \tau_K)_+^3 (\tau_{K-1} - \tau_j)}{\tau_K - \tau_{K-1}},$$

όπου $x_+ = \max(0, x)$. Το μοντέλο (2) μπορεί να γραφεί σαν

$$H(x) = \beta_1 (x - \tau_1)_+^3 + \dots + \beta_{K-1} (x - \tau_{K-1})_+^3 + \beta_K (x - \tau_K)_+^3 \quad (3)$$

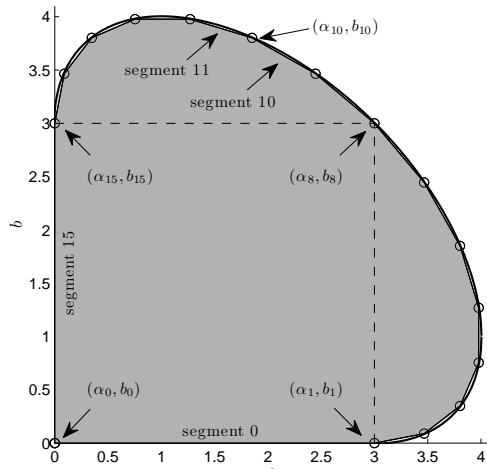
όπου

$$\beta_{K-1} = \frac{\beta_1 (\tau_1 - \tau_K) + \beta_2 (\tau_2 - \tau_K) + \dots + \beta_{K-2} (\tau_{K-2} - \tau_K)}{\tau_K - \tau_{K-1}} \quad (4)$$

$$\beta_K = \frac{\beta_1 (\tau_1 - \tau_{K-1}) + \beta_2 (\tau_2 - \tau_{K-1}) + \dots + \beta_{K-2} (\tau_{K-2} - \tau_{K-1})}{\tau_{K-1} - \tau_K}.$$

Το μοντέλο (3) έχει τις ακόλουθες ιδιότητες: (1) Είναι γραμμικό πριν από τον πρώτο και μετά τον τελευταίο κόμβο, (2) Είναι ίσο με μηδέν πριν από τον πρώτο κόμβο, (3) Η πρώτη και δεύτερη παράγωγος είναι συνεχείς, (4) Η πρώτη παράγωγος είναι μηδέν στον πρώτο κόμβο, (5) Υπάρχουν $K - 2$ παράμετροι προς εκτίμηση.

Για να προσαρμοστεί το παραπάνω μοντέλο στην ΚΜ εκτίμηση της αθροιστικής συνάρτησης κινδύνου χρειάζονται κάποιες συνθήκες μονοτονίας. Ικανές και αναγκαίες συνθήκες μονοτονίας για ένα κυβικό πολυώνυμο δίνονται από τους Fritsch και Carlson (1980). Έστω το κυβικό πολυώνυμο P στο διάστημα $[\tau_j, \tau_{j+1}]$ όπου με P' συμβολίζουμε την πρώτη παράγωγο, και οι ποσότητες $a_j = \frac{\tau_{j+1} - \tau_j}{P(\tau_{j+1}) - P(\tau_j)} P'(\tau_j)$ και $b_j = \frac{\tau_{j+1} - \tau_j}{P(\tau_{j+1}) - P(\tau_j)} P'(\tau_{j+1})$. Έστω επίσης η περιοχή $\mathcal{M} = \mathcal{M}_1 \cup \mathcal{M}_2$, όπου \mathcal{M}_1 είναι το τετράγωνο που ορίζεται από τα σημεία $a = 0, 3$ και $b = 0, 3$ και \mathcal{M}_2 η έλλειψη που ορίζεται από τη σχέση $\phi(a, b) = (a - 1)^2 + (a - 1)(b - 1) + (b - 1)^2 - 3(a + b - 2) = 0$. Το πολυώνυμο είναι μονότονο στο διάστημα $[\tau_j, \tau_{j+1}]$ αν και μόνο αν το σημείο (a_j, b_j) βρίσκεται εντός της περιοχής \mathcal{M} . Κάτω από τον επιπλέον περιορισμό $0 \leq \min(P'(\tau_j), P'(\tau_{j+1}))$ εξασφαλίζεται ότι η συνάρτηση του πολυωνύμου θα είναι μη φθίνουσα. Σημειώνουμε ότι οποιαδήποτε υποπεριοχή της \mathcal{M} παρέχει μια ικανή συνθήκη μονοτονίας. Η περιοχή \mathcal{M} είναι μη γραμμική και το να θεωρήσει κανείς μη γραμμικούς περιορισμούς μπορεί να προκαλέσει υπολογιστικά προβλήματα σε οποιαδήποτε διαδικασία μεγιστοποίησης/ελαχιστοποίησης. Εμείς εξερευνούμε μία γραμμική προσέγγιση, \mathcal{A} , της περιοχής \mathcal{M} , με χρήση 16 ευθειών (segments). Η γραφική απεικόνιση τόσο της περιοχής \mathcal{M} όσο και της γραμμικής προσέγγισης \mathcal{A} που χρησιμοποιούμε φαίνεται και στο Γράφημα 1.



Σχήμα 1: Γραμμική προσέγγιση, \mathcal{A} , με 16 ευθείες (segments) της περιοχής μονοτονίας \mathcal{M} . Το εμβαδόν της \mathcal{A} είναι περίπου το 98.4% του εμβαδού της \mathcal{M} .

Τα σημεία που ορίζουν τη γραμμική προσέγγιση υπολογίστηκαν με αριθμητική μεγιστοποίηση του εμβαδού της \mathcal{A} . Ο Smith (1971) παρουσιάζει έναν αλγόριθμο με τον οποίο μπορεί κανείς να υπολογίσει ακριβώς που πρέπει να τοποθετηθούν τα σημεία στην περίμετρο μιας έλλειψης έτσι ώστε το εμβαδό του πολυγώνου που προκύπτει από αυτά να είναι το μέγιστο δυνατό. Αυτός ο αλγόριθμος δεν λύνει το πρόβλημα για βέλτιστη γραμμική προσέγγιση (ως προς το εσωκλειόμενο εμβαδό) της περιοχής \mathcal{M} μιας και μόνο ένα κομμάτι της αποτελεί μέρος έλλειψης. Ωστόσο, μπορεί να αποδειχθεί ότι με χρήση αυτού του αλγόριθμου μπορεί κανείς να υπολογίσει άμεσα οποιοδήποτε $(8k + 2)$ -γωνο, $k = 1, 2, \dots$, και έτσι να προσεγγίσει την περιοχή μονοτονίας \mathcal{M} με οποιαδήποτε επιθυμητή ακρίβεια. Η απόδειξη βασίζεται στην κατάτμηση του περιγεγραμμένου κύκλου της έλλειψης $\phi(a, b)$ σε ίσες γωνίες και στον υπολογισμό τριγωνομετρικών συναρτήσεων αυτών (παραλείπεται χάριν συντομίας). Οι προσομοιώσεις που παρουσιάζονται έχουν εκτελεστεί και για προσέγγιση της περιοχής μονοτονίας με βέλτιστο 18-γωνο και 26-γωνο χωρίς να παρατηρήσουμε καμία διαφορά στα αποτελέσματα.

3.2 Εύρεση των γραμμικών περιορισμών και εκτίμηση με ελάχιστα τετράγωνα

Θεωρούμε ότι τα σημεία $(0,0)$, $(0,3)$ και $(3,0)$ περιλαμβάνονται στη γραμμική προσέγγιση της περιοχής \mathcal{M} . Έστω τα σημεία που ορίζουν τη γραμμική προσέγγιση: $(a_0, b_0) = (a_{Q+2}, b_{Q+2}) = (0,0)$, $(a_1, b_1) = (3,0)$, $(a_2, b_2), \dots, (a_Q, b_Q)$, και $(a_{Q+1}, b_{Q+1}) = (0,3)$. Συμβολίζουμε το i -στό ευθύγραμμο τμήμα που ενώνει το (a_i, b_i) και το (a_{i+1}, b_{i+1}) με g_i , και η αντίστοιχη εξίσωσή του να δίνεται από

$g_i(a, b) = b + \gamma_1^{(i)} a + \gamma_0^{(i)} = 0$, όπου $i = 1, \dots, Q$ (βλέπε και Γράφημα 1 όπου $Q = 14$).

Πρώτα, χρειάζεται να ορίσουμε μια ποσότητα που θα μας βοηθά να βρισκόμαστε στο σωστό ημιεπίπεδο για κάθε ένα από τα Q ευθύγραμμα τμήματα (συνολικά έχουμε $Q + 2$ ευθύγραμμα τμήματα αλλά τα 2 συμπίπτουν με τον κάθετο και οριζόντιο άξονα μιας και τα σημεία $(0,0)$, $(0,3)$ και $(3,0)$ συμπεριλαμβάνονται στην προσέγγιση). Έτσι έχουμε για το i -οστό ευθύγραμμο τμήμα:

1. Αν $b_{i+1} \leq 1$ τότε απαιτούμε $g_i(a, b) \geq 0$ (και θέτουμε $v_i = -1$)
2. Αν $b_i \geq 1$ τότε απαιτούμε $g_i(a, b) \leq 0$ (και θέτουμε $v_i = 1$)
3. Αν $b_i < 1$ και $b_{i+1} > 1$ τότε
 - αν $a_i < a_{i+1}$ τότε απαιτούμε $g_i(a, b) \geq 0$ (και θέτουμε $v_i = -1$)
 - αν $a_i \geq a_{i+1}$ τότε απαιτούμε $g_i(a, b) \leq 0$ (και θέτουμε $v_i = 1$)

Έτσι για τους περιορισμούς που απαιτούνται για την προσέγγιση της περιοχής \mathcal{A} , ύστερα από λίγη άλγεβρα έχουμε τους ακόλουθους $Q(K - 1)$ περιορισμούς

$$v_i \left\{ \beta_J \left[3(\tau_{J+1} - \tau_J)^2 + \gamma_0^{(i)} (\tau_{J+1} - \tau_J)^2 \right] + I_{(J \geq 2)} \sum_{j=1}^{J-1} \beta_j \left[3(\tau_{J+1} - \tau_j)^2 + 3\gamma_1^{(i)} (\tau_J - \tau_j)^2 + \gamma_0^{(i)} \frac{(\tau_{J+1} - \tau_j)^3 - (\tau_J - \tau_j)^3}{\tau_{J+1} - \tau_J} \right] \right\} \leq 0, \quad (5)$$

όπου $i = 1, 2, \dots, Q$ και $I_{(J \geq 2)} = 1$ αν $J \geq 2$ και 0 αλλιώς.

Επιπλέον απαιτούμε οι παράγωγοι πριν από τον πρώτο κόμβο και μετά τον τελευταίο να είναι μη αρνητικές. Ισχύει ότι $H'(\tau_1) = 0$. Έτσι απαιτούμε $H'(\tau_j) \geq 0$, για $j = 2, \dots, K - 1$ που μας οδηγούν στους ακόλουθους $K - 2$ περιορισμούς

$$- \sum_{j=1}^{J-1} 3\beta_j (\tau_J - \tau_j)^2 \leq 0, \quad J = 2, \dots, K - 1. \quad (6)$$

Επίσης η αυστηρή ανίσωση ότι η παράγωγος στο τελευταίο κόμβο είναι μεγαλύτερη του μηδέν δίνει τον ακόλουθο περιορισμό

$$- \sum_{j=1}^{K-1} 3\beta_j (\tau_K - \tau_j)^2 < 0. \quad (7)$$

Σημειώνουμε ότι οι ανισώσεις $a_j \geq 0$ και $b_j \geq 0$, $\forall j = 1, \dots, K - 1$ είναι τετραμμένες λόγω του ότι $\tau_1 < \tau_2 < \dots < \tau_K$ και λόγω των περιορισμών (6) και (7).

Τέλος, χρειαζόμαστε και έναν περιορισμό ισότητας για την παράμετρο β_{K-1} που δίνεται από τη σχέση (4). Έτσι συνολικά υπάρχουν $Q(K - 1) + (K - 2) + 1 + 1 = Q(K - 1) + K$ περιορισμοί. Όλοι οι περιορισμοί είναι δυνατόν να γραφούν και σε μορφή πινάκων ως εξής:

$$\mathbf{A}[\beta_1, \beta_2, \dots, \beta_{K-1}]' \leq \mathbf{0} \quad (8)$$

όπου ο πίνακας \mathbf{A} έχει $n_c = K - 1$ στήλες και $n_r = Q(K - 1) + K + 1$ γραμμές.

Η συνάρτηση προς ελαχιστοποίηση είναι η

$$\Psi(\beta) = \sum_i \left(H(T_i|D_i = 1) - \hat{H}^{KM}(T_i|D_i = 1) \right)^2,$$

κάτω από τους περιορισμούς (8). Έτσι έχουμε ένα πρόβλημα ελαχιστοποίησης αθροίσματος τετραγώνων κάτω από γραμμικούς περιορισμούς. Η συνάρτηση προς ελαχιστοποίηση είναι πάντα κυρτή και για την επίλυση τέτοιων προβλημάτων υπάρχουν έτοιμες συναρτήσεις σε αρκετά υπολογιστικά πακέτα.

Σε αυτήν την εργασία χρησιμοποιούμε 6 κόμβους και θεωρούμε τρεις στρατηγικές τοποθέτησής τους: 1. Ισαπέχοντες κόμβους στο διάστημα $\min(T_i|D_i = 1)$ έως $\max(T_i|D_i = 1)$, 2. Κόμβοι στα $\min(T_i|D_i = 1)$, 5ο, 25ο, 50ο, 75ο ποσοστημόρια και $\max(T_i|D_i = 1)$, 3. Κόμβοι στα 0, 5ο, 25ο, 50ο, 75ο ποσοστημόρια και στο $\max(T_i|D_i = 1)$. Η καλύτερη τοποθέτηση επιλέγεται με βάση τη μικρότερη απόσταση από τον εκτιμητή ΚΜ. Τα παραπάνω ποσοστημόρια υπολογίζονται μόνο από τα πλήρη δεδομένα. Έτσι, για ένα σετ δεδομένων, δοκιμάζονται και οι τρεις στρατηγικές τοποθέτησης των κόμβων και επιλέγεται η καλύτερη βάσει της ποσότητας $\Psi(\hat{\beta})$. Στην περίπτωση που υπάρχουν διαθέσιμες και άλλες συμμεταβλητές τότε μπορεί να χρησιμοποιηθεί το μοντέλο του Cox και να εφαρμοστεί η παραπάνω μεθοδολογία με σκοπό τη λείανση της αναφορικής αθροιστικής συνάρτησης κινδύνου. Έτσι μπορεί κανείς να λάβει μια λεία εκτίμηση της συνάρτησης επιβίωσης για οποιοδήποτε προφίλ κάποιου ασθενή.

4. ΠΡΟΣΟΜΟΙΩΣΕΙΣ

Διενεργήσαμε προσομοιώσεις για να αξιολογήσουμε την προτεινόμενη μέθοδο HCNS (Hazard Constrained Natural Spline). Θεωρήσαμε δύο πυκνότητες, τη $Weibull(2, 3)$ και τη δίκροφη μίξη $0.5Weibull(5, 4) + 0.5Weibull(4, 2)$. Το κριτήριο με το οποίο αξιολογούνται οι μέθοδοι είναι το Mean Integrated Squared Error ($MISE$) που υπολογίζεται με βάση 1000 επαναλήψεις. Για κάθε επανάληψη υπολογίζεται το $ISE = \int_{X_{10}}^{X_{90}} (\hat{S} - S)^2$, όπου X_{10} και X_{90} είναι το 10ο και 90ο ποσοστημόριο της πραγματικής κατανομής. Αυτό που παρατηρούμε γενικά είναι ότι η προτεινόμενη μέθοδος παρέχει σχεδόν σε όλες τις περιπτώσεις μικρότερο $MISE$ από τον ΚΜ εκτιμητή ενώ γενικά η logspline τεχνική παρέχει στις περισσότερες περιπτώσεις μικρότερο $MISE$ και από τις τρεις μεθόδους. Εξάιρεση αποτελεί η περίπτωση της μικτής κατανομής όπου η προτεινόμενη μεθοδολογία φαίνεται να είναι καλύτερη για μικρότερα δείγματα και από τη logspline. Ωστόσο, παρατηρούμε ότι σε αντίθεση με την προτεινόμενη μέθοδο, η logspline τεχνική έχει προβλήματα σύγκλισης (βλέπε και Πίνακα 1). Σαν μέτρο σύγκρισης, εκτός από το $MISE$, χρησιμοποιήσαμε και την Anderson Darling (AD) απόσταση $MISE_{(AD)} = E(\int_{X_{10}}^{X_{90}} w(x)(\hat{S} - S)^2 dx)$, όπου $w(x) = 1/(S(x)(1 - S(x)))$. Τα αποτελέσματα είναι αντίστοιχα και παρουσιάζονται στο κάτω μέρος του Πίνακα 1.

Πίνακας 1: Αποτελέσματα προσομοίωσης από 1000 επαναλήψεις για τις μεθόδους KM, HCNS και logspline. Μεγέθη δειγμάτων $n = 300$, $n = 100$ και $n = 60$ με 70%, 50% και 30% λογοκρισία. Το κριτήριο σύγκρισης είναι το MISE. Αριστερά είναι τα αποτελέσματα για τη Weib(2,3) και δεξιά για τη μικτή κατανομή. Το πρώτο (άνω) κομμάτι του πίνακα η σύγκριση γίνεται με το κριτήριο MISE ενώ στο δεύτερο (κάτω) με το $(MISE_{(AD)})$.

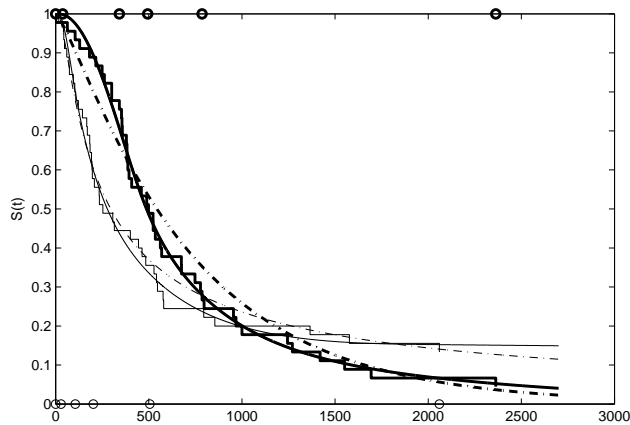
n	Λογοκρισία	Weibull(2,3)			Mixture		
		KM	HCNS	logspline	KM	HCNS	logspline
<i>MISE</i>							
300	30%	0.0030	0.0038	0.0027	0.0027	0.0030	0.0026
	50%	0.0040	0.0044	0.0034	0.0038	0.0040	0.0035
	70%	0.0069	0.0064	0.0049	0.0068	0.0064	— ⁽³⁾
100	30%	0.0089	0.0090	0.0069	0.0079	0.0075	0.0081
	50%	0.0121	0.0115	0.0090	0.0110	0.0102	0.0119 ⁽²⁾
	70%	0.0210	0.0176	0.0132 ⁽²⁾	0.0178	0.0161	— ⁽³⁾
60	30%	0.0153	0.0148	0.0119	0.0144	0.0133	0.0149
	50%	0.0207	0.0186	0.0149	0.0193	0.0170	0.0191 ⁽¹⁾
	70%	0.0365	0.0294	0.0229	0.0350	0.0284	— ⁽³⁾
<i>MISE_(AD)</i>							
300	30%	0.0163	0.0219	0.0141	0.0139	0.0158	0.0129
	50%	0.0223	0.0255	0.0182	0.0194	0.0210	0.0180
	70%	0.0392	0.0372	0.0271	0.0359	0.0338	— ⁽³⁾
100	30%	0.0482	0.0504	0.0369	0.0398	0.0385	0.0406
	50%	0.0670	0.0643	0.0049	0.0567	0.0519	0.0524 ⁽²⁾
	70%	0.1214	0.0975	0.0688 ⁽²⁾	0.1106	0.0900	— ⁽³⁾
60	30%	0.0829	0.0812	0.0627	0.0737	0.0674	0.0743
	50%	0.1142	0.1020	0.0789	0.0999	0.0869	0.0976 ⁽¹⁾
	70%	0.2086	0.1591	0.1231	0.1859	0.1467	— ⁽³⁾

(1) 2.0% των επαναλήψεων είχαν προβλήματα σύγκλισης
(2) 3.0% των επαναλήψεων είχαν προβλήματα σύγκλισης
(3) Περισσότερα προβλήματα σύγκλισης ($\geq 10\%$)

5. ΕΦΑΡΜΟΓΕΣ

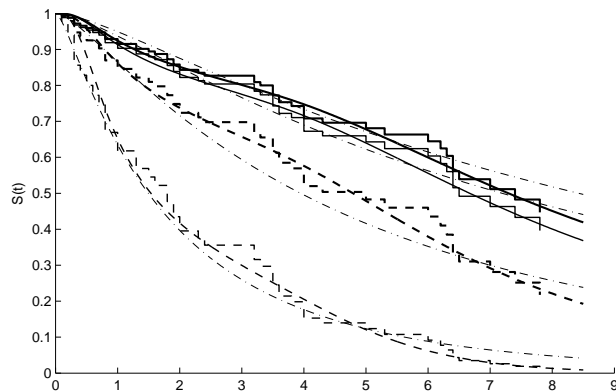
Το 1982 διεξήχθη μια κλινική μελέτη από το Gastrointestinal Tumor Study Group που αφορούσε στην αντιμετώπιση του καρκίνου του στομάχου. Οι υπό μελέτη θεραπείες ήταν δύο, η χημειοθεραπεία και η χημειοθεραπεία σε συνδυασμό με τη ραδιοθεραπεία. Συμμετείχαν 90 ασθενείς και τυχαιοποιήθηκαν στις δύο θεραπείες (βλέπε και Stablein και Koutrouvelis(1985)). Χρησιμοποιήσαμε τις μεθόδους εκτίμησης Kaplan Meier, logspline, HCNS και το συχνά χρησιμοποιούμενο μοντέλο της Generalized Gamma (GG) κατανομής για εκτίμηση των συνάρτησεων επιβίωσης (Γράφημα 2). Παρατηρούμε ότι τις πρώτες 1000 μέρες περίπου η χημειοθεραπεία φαίνεται να είναι προτιμότερη. Σημειώνουμε ότι οι καμπύλες επιβίωσης της μεθόδου logspline σχεδόν ταυτίζονται με τις αντίστοιχες της HCNS και δεν παρουσιάζονται για να είναι πιο ευδιάκριτο το γράφημα.

Θεωρήσαμε επίσης τα δεδομένα που παρουσιάζονται από τον Kardaun (1983) και αφορούν σε 90 άντρες με καρκίνο του λάρυγγα. Υπάρχουν τέσσερα στάδια του καρκίνου, (1 έως 4) από το λιγότερο σοβαρό ως το σοβαρότερο. Εκτός από το χρόνο μέχρι το θάνατο ή τη λογοκρισία, για κάθε ασθενή είναι διαθέσιμο το στάδιο του καρκίνου καθώς και η ηλικία του. Αφού προσαρμόσαμε το μοντέλο του Cox, προσαρμόσαμε το προτεινόμενο μοντέλο στην αναφορική συνάρτηση κινδύνου και παραθέτουμε το Γράφημα 3 όπου φαίνονται οι τέσσερις καμπύλες επιβίωσης (μια για κάθε στάδιο) για τη μέση ηλικία (64.6111). Αυτό που παρατηρούμε είναι αναμενόμενο,



Σχήμα 2: Έντονες γραμμές: Θεραπεία χημειοθεραπείας, Λεπτές γραμμές: Συνδυασμένη θεραπεία, Βηματικές Συναρτήσεις: εκτίμηση της επιβίωσης με Kaplan Meier, Συνεχείς γραμμές: εκτίμηση επιβίωσης με την HCNS, Διακεκομένες γραμμές: εκτίμηση επιβίωσης με GG, Έντονοι κύκλοι: κόμβοι για την ομάδα της χημειοθεραπείας, Λεπτοί κύκλοι: κόμβοι για τη συνδυασμένη θεραπεία.

όσο πιο προχωρημένο είναι το στάδιο του καρκίνου τόσο χαμηλότερες εκτιμημένες καμπύλες επιβίωσης λαμβάνουμε.



Σχήμα 3: Οι καμπύλες επιβίωσης για κάθε στάδιο καρκίνου για τη μέση ηλικία (64.6111). Έντονες συνεχείς, λεπτές συνεχείς, έντονες διακεκομένες, λεπτές διακεκομένες γραμμές είναι οι εκτιμήσεις επιβίωσης για τα στάδια 1-4 αντίστοιχα με την προτεινόμενη μέθοδο. Οι γραμμές που διακόπτονται με παύλες και τελείες είναι από το GG μοντέλο. Οι βηματικές συναρτήσεις είναι του μοντέλου του Cox.

6. ΣΥΜΠΕΡΑΣΜΑΤΑ

Σε αυτήν την εργασία προτείνουμε μια μέθοδο εκτίμησης της συνάρτησης επιβίωσης με χρήση μιας μονότονης κυβικής φυσικής spline. Η μέθοδος είναι υπολογιστικά απλή και δεν έχει προβλήματα σύγκλισης, σε αντίθεση με τη logspline τεχνική, μιας και η συνάρτηση που ελαχιστοποιεί είναι πάντα κυρτή. Η γενίκευση της μεθόδου για την περίπτωση που υπάρχουν διαθέσιμες συμμεταβλητές είναι άμεση κάτω από την υπόθεση των ανάλογων κινδύνων. Ένα επιπλέον ελκυστικό χαρακτηριστικό της μεθόδου είναι ότι μπορεί να παρέχει εκτίμηση της συνάρτησης επιβίωσης και πέραν του τελευταίου γεγονότος, κάτι που αδυνατούν να κάνουν τόσο ο *KM* εκτιμητής όσο και οι τεχνικές που βασίζονται στους πυρήνες. Παραιτέρω έρευνα θα μπορούσε να επικεντρωθεί στην επιλογή του πλήθους των κόμβων. Προσομοιώσεις (που δεν παρουσιάζονται εδώ) έχουν δείξει ότι 6 κόμβοι παρέχουν καλύτερα αποτελέσματα σε αρκετές περιπτώσεις σε σχέση με αυτά που λαμβάνουμε με χρήση λιγότερων κόμβων. Ωστόσο θα είχε ενδιαφέρον η κατασκευή ενός κριτηρίου βάσει του οποίου θα μπορούσε να γίνει η επιλογή του πλήθους των κόμβων.

Ευχαριστίες: Θα θέλαμε να ευχαριστήσουμε τον ανώνυμο κριτή για τα χρήσιμα σχόλιά του.

ABSTRACT

In this paper we study survival estimation in the presence of censoring. We explore fitting a natural cubic spline to the cumulative hazard step function under appropriate constraints. Under the proposed method the problem reduces to a restricted least squares one, with linear restrictions. The proposed method is compared via simulations to other known methods such as the non parametric Kaplan Meier and the logspline technique. In the presence of covariates, estimation is made under the proportional hazards assumption. We apply our method to real data sets regarding gastric and larynx cancer.

ΑΝΑΦΟΡΕΣ

- Bowman A. W. and Azzalini, A. (1997). *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-plus Illustrations*. Oxford University Press.
- Fritsch F. N., Carlson, R. E. (1980). Monotone Piecewise Cubic Interpolation *SIAM Journal of Numerical Analysis* **2**, 238–246.
- Gill, R. D. (1980). *Censoring and Stochastic Integrals*. Mathematisch Centrum, Amsterdam. **124**
- Kaplan, E. L. and Meier, P. (1958). Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association* **53**, 457–481.

- Efron, B. (1967). The Two Sample Problem with Censored Data. *In Proceedings of the Fifth Berkeley Symposium, New York: Prentice-Hall* **4**, 831–853.
- Kardaun, O. (1983). Statistical Analysis of Male Larynx Cancer Patients, A Case Study. *Statistica Neerlandica* **37**, 103–125.
- Kooperberg, C., Stone, J. S. (1992). Log-spline density estimation for censored data. *Journal of Computational and Graphical Statistics* **1**, 301–328.
- Moeschberger, M. L. and Klein, J. P. (1985). Comparison of Several Methods of Estimating the Survival Function When There is Extreme Censoring. *Biometrics* **41**, 253–259.
- Pan, W. (2000). Smooth Estimation of the Survival Function for interval censored data. *Statistics in Medicine* **19**, 2611–2624.
- Smith, L. B. (1971). Drawing ellipses, hyperbolas or parabolas with a fixed number of points and maximum inscribed area. *The Computer Journal*. **14**, 81–86.
- Stablein, D. M. and Koutrouvelis, I. A. A (1985). A Two-Sample Test Sensitive to Crossing Hazards in Uncensored and Singly Censored Data. *Biometrics* **41**, 643–652.
- Wand, M. P., and Jones, M. C. (1995). *Kernel Smoothing*. Monographs on Statistics and Applied Probability London: Chapman and Hall.

ΝΕΕΣ ΚΛΑΣΕΙΣ ΕΚΤΙΜΗΤΩΝ ΓΙΑ ΤΟΝ ΠΙΝΑΚΑ ΔΙΑΣΠΟΡΩΝ-ΣΥΝΔΙΑΣΠΟΡΩΝ ΠΟΛΥΜΕΤΑΒΛΗΤΗΣ ΚΑΝΟΝΙΚΗΣ ΚΑΤΑΝΟΜΗΣ

Παναγιώτης Μπομποτάς, Σταύρος Κουρούκλης
Τμήμα Μαθηματικών, Πανεπιστήμιο Πατρών
{pb, stavros}@math.upatras.gr

ΠΕΡΙΛΗΨΗ

Στην εργασία αυτή μελετάται το πρόβλημα της εκτίμησης του πίνακα διασπορών-συνδιασπορών, Σ , πολυμεταβλητής κανονικής κατανομής από την πλευρά της Στατιστικής Θεωρίας Αποφάσεων ως προς τη συνάρτηση ζημίας εντροπίας και την τετραγωνική συνάρτηση ζημίας. Αποδεικνύεται ότι η κατασκευή βελτιωμένων εκτιμητών του Σ ανάγεται στην κατασκευή βελτιωμένων εκτιμητών της διασποράς (μονομεταβλητής) κανονικής κατανομής. Η αναγωγή αυτή οδηγεί αμέσως στην εύρεση εκτιμητών τύπου Stein (1964), τύπου Brewster and Zidek (1974) και τύπου Strawderman (1974) για τον πίνακα Σ . Η εργασία ενοποιεί και επεκτείνει υπάρχοντα αποτελέσματα για την εκτίμηση του Σ .

1. ΕΙΣΑΓΩΓΗ

Έστω X_1, \dots, X_n τυχαίο δείγμα από την πολυμεταβλητή κανονική κατανομή $N_p(\theta, \Sigma)$, όπου θ, Σ είναι άγνωστες παράμετροι. Η (ελάχιστη) επαρκής στατιστική συνάρτηση είναι (\bar{X}, S) όπου για $\bar{X} = n^{-1} \sum_{i=1}^n X_i$ έχουμε $\sqrt{n}\bar{X} \sim N_p(\sqrt{n}\theta, \Sigma)$ και $S = \sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})' \sim W_p(n-1, \Sigma)$ (Wishart). Αντικείμενο της εργασίας είναι η εκτίμηση του πίνακα διασπορών-συνδιασπορών, Σ , ως προς τη συνάρτηση ζημίας εντροπίας $L_1(\delta, \Sigma) = \text{tr}(\delta \Sigma^{-1}) - \ln |\delta \Sigma^{-1}| - p$ και την τετραγωνική συνάρτηση ζημίας $L_2(\delta, \Sigma) = \text{tr}(\delta \Sigma^{-1} - 1)^2$. Δύο εκτιμητές συγκρίνονται με κριτήριο τη μέση ζημία τους (risk function) και καλύτερος είναι αυτός που έχει μικρότερη μέση ζημία για κάθε τιμή των άγνωστων παραμέτρων. Ο εκτιμητής με μεγαλύτερη μέση ζημία για κάθε τιμή των άγνωστων παραμέτρων λέγεται μη αποδεκτός (inadmissible), ενώ αυτός με μικρότερη αναφέρεται ως βελτιωμένος (improved) σε σχέση με τον μη αποδεκτό. Ο εκτιμητής με τη μικρότερη μέση ζημία σε μία κλάση εκτιμητών αναφέρεται ως βέλτιστος εκτιμητής της κλάσης.

Ο βέλτιστος αναλλοίωτος κατά θέση-κλίμακα εκτιμητής του Σ ως προς την L_1 είναι ο αμερόληπτος εκτιμητής ελαχίστης διασποράς $\delta_0 = (n-1)^{-1}S$, ο οποίος όμως είναι μη αποδεκτός. Βασιζόμενοι στην ιδέα του Stein (1964), οι Sinha and Ghosh (1987) κατασκεύασαν τον βελτιωμένο εκτιμητή

$$\delta_{SG} = \begin{cases} (n-1)^{-1}S & \text{εάν } n\bar{X}'S^{-1}\bar{X} > (n-1)^{-1} \\ n^{-1}(S+n\bar{X}\bar{X}') & \text{διαφορετικά} \end{cases},$$

ενώ ο Perron (1990) τον βελτιωμένο εκτιμητή

$$\delta_{PR} = \begin{cases} n^{-1}(S+(n-1)^{-1}(\bar{X}'S^{-1}\bar{X})^{-1}\bar{X}\bar{X}') & \text{εάν } n\bar{X}'S^{-1}\bar{X} > (n-1)^{-1} \\ n^{-1}(S+n\bar{X}\bar{X}') & \text{διαφορετικά} \end{cases}.$$

Οι Kubokawa et al. (1992) κατασκεύασαν τον εμπειρικό εκτιμητή Bayes

$$\delta_S = \begin{cases} n^{-1}(S+p(n-p)^{-1}(\bar{X}'S^{-1}\bar{X})^{-1}\bar{X}\bar{X}') & \text{εάν } n\bar{X}'S^{-1}\bar{X} > p(n-p)^{-1} \\ n^{-1}(S+n\bar{X}\bar{X}') & \text{διαφορετικά} \end{cases}. \quad (1)$$

Σημειώνεται ότι οι εκτιμητές δ_{PR} και δ_S είναι καλύτεροι και από τον δ_{SG} , πέραν του δ_0 (Kubokawa et al., 1992).

Παρακινούμενοι από τον Perron (1990), οι Kubokawa et al. (1993) θεώρησαν την κλάση των εκτιμητών του Σ ,

$$\delta^* = aS + b(\bar{X}'S^{-1}\bar{X})^{-1}\bar{X}\bar{X}', \quad a > 0, \quad a+b > 0, \quad (2)$$

και έδειξαν ότι ο βέλτιστος εκτιμητής σε αυτή την κλάση ως προς την L_1 , είναι ο

$$\delta_0^* = n^{-1}S + p(n(n-p))^{-1}(\bar{X}'S^{-1}\bar{X})^{-1}\bar{X}\bar{X}'. \quad (3)$$

Άρα, ο δ_0^* είναι καλύτερος από τον $\delta_0 = (n-1)^{-1}S$, για $p \geq 2$ (ενώ για $p=1$ ταυτίζονται).

Από τα παραπάνω είναι φανερό ότι περισσότερο ενδιαφέρον παρουσιάζει το πρόβλημα βελτίωσης του δ_0^* παρά εκείνο του δ_0 , αφού κάθε εκτιμητής που βελτιώνει τον δ_0^* είναι καλύτερος και από τον δ_0 . Προς την κατεύθυνση της βελτίωσης του δ_0^* , οι Kubokawa et al. (1993) θεώρησαν εκτιμητές της μορφής

$$\delta_\psi = \frac{1}{n}S + \frac{\psi(W)}{\bar{X}'S^{-1}\bar{X}}\bar{X}\bar{X}', \quad (4)$$

όπου $\psi(\cdot)$ μία συνάρτηση και $W = n\bar{X}'S^{-1}\bar{X}$, και έδειξαν ότι ο δ_S , που ανήκει στην κλάση (4) με $\psi(w) = \min\{p(n-p)^{-1}w^{-1}, 1\}$, προκύπτει χρησιμοποιώντας την τεχνική

Stein (1964) και είναι καλύτερος από τον δ_0^* . Επί πλέον, βασιζόμενοι στην ιδέα των Brewster and Zidek (1974) κατασκεύασαν ακόμη έναν καλύτερο εκτιμητή από τον δ_0^* , που επίσης ανήκει στην κλάση (4), τον

$$\delta_{BZ} = \frac{1}{n}S + \frac{\int_0^1 \lambda^{p/2} (1 + \lambda n \bar{X}' S^{-1} \bar{X})^{-n/2-1} d\lambda}{\int_0^1 \lambda^{p/2-1} (1 + \lambda n \bar{X}' S^{-1} \bar{X})^{-n/2-1} d\lambda} \bar{X} \bar{X}'. \quad (5)$$

Ο δ_{BZ} είναι πιο «ομαλός» (smooth) από τον δ_S και γενικευμένος εκτιμητής Bayes (Kubokawa et al., 1993).

Πριν προχωρήσουμε στην παρουσίαση των αποτελεσμάτων μας για την εκτίμηση του Σ , παραθέτουμε ορισμένα γνωστά αποτελέσματα για την εκτίμηση της διασποράς (μονομεταβλητής) κανονικής κατανομής, τα οποία είναι χρήσιμα για τη σύνδεση των δύο προβλημάτων. Έστω X, S_0 ανεξάρτητες στατιστικές συναρτήσεις τέτοιες ώστε $X \sim N_p(\mu, \sigma^2 I_p)$, $S_0 \sim \sigma^2 \chi_m^2$, όπου μ, σ^2 είναι άγνωστες παράμετροι. Ένα παράδειγμα αυτού του μοντέλου είναι η Ανάλυση Διασποράς κατά ένα παράγοντα, όπου θεωρούμε $X_{ij} = \theta_i + \varepsilon_{ij}$, $i = 1, \dots, p$, $j = 1, \dots, n_i$, με $\varepsilon_{ij} \sim N(0, \sigma^2)$. Τότε, για $\bar{X}_i = n_i^{-1} \sum_{j=1}^{n_i} X_{ij}$, $\mu = (\sqrt{n_1} \theta_1, \dots, \sqrt{n_p} \theta_p)'$, $\sum_{i=1}^p n_i = n$, έχουμε $X = (\sqrt{n_1} \bar{X}_1, \dots, \sqrt{n_p} \bar{X}_p)' \sim N_p(\mu, \sigma^2 I_p)$ και $S_0 = \sum_{i=1}^p \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2 \sim \sigma^2 \chi_{n-p}^2$. Ο βέλτιστος αναλλοίωτος κατά θέση-κλίμακα εκτιμητής του σ^2 ως προς τη συνάρτηση ζημίας εντροπίας $L(\delta, \sigma^2) = \delta / \sigma^2 - \ln(\delta / \sigma^2) - 1$ είναι ο $\delta_0^{\sigma^2} = S_0 / m$. Στην βιβλιογραφία έχουν δοθεί βελτιωμένοι εκτιμητές του σ^2 της μορφής

$$\delta_\varphi^{\sigma^2} = \varphi(W_0) S_0, \quad (6)$$

όπου $\varphi(\cdot)$ μία θετική συνάρτηση και $W_0 = \|X\|^2 / S_0$, όπως ο εκτιμητής Stein (1964) (βλ. Kubokawa, 1994)

$$\delta_S^{\sigma^2} = \min \left\{ \frac{1}{m}, \frac{1 + W_0}{p + m} \right\} S_0, \quad (7)$$

ο εκτιμητής Brewster and Zidek (1974) (βλ. Kubokawa, 1994)

$$\delta_{BZ}^{\sigma^2} = \frac{1}{m + p} \frac{\int_0^1 \lambda^{p/2-1} (1 + \lambda W_0)^{-(m+p)/2} d\lambda}{\int_0^1 \lambda^{p/2-1} (1 + \lambda W_0)^{-(m+p)/2-1} d\lambda} S_0, \quad (8)$$

οι εκτιμητές Maruyama (1998)

$$\delta_M^{\sigma^2} = \frac{1}{m+p} \frac{\int_0^1 \lambda^{ap/2-1} (1+\lambda W_0)^{-a(m+p)/2} d\lambda}{\int_0^1 \lambda^{ap/2-1} (1+\lambda W_0)^{-a(m+p)/2-1} d\lambda} S_0, \quad a > 1, \quad (9)$$

και οι εκτιμητές τύπου Strawderman (1974) (από τους Maruyama and Strawderman, 2006)

$$\delta_{ST}^{\sigma^2} = \frac{1}{m} \frac{1+W_0}{r+1+W_0} S_0, \quad 0 < r \leq \max_{\varepsilon} \min \left\{ \frac{B(\varepsilon)}{1-B(\varepsilon)}, \frac{1}{\varepsilon} - 1 \right\}, \quad (10)$$

όπου

$$B(\varepsilon) = \min \left\{ B_*, \left(1 + \frac{m}{2p\varepsilon} \frac{\Gamma(m/2+2\varepsilon)\Gamma((m+p)/2+\varepsilon+1)}{\Gamma(m/2+\varepsilon)\Gamma((m+p)/2+2\varepsilon)} \right)^{-1} \right\}$$

και B_* η μοναδική λύση της εξίσωσης $x^{-1} \ln(1-x) = (m+p)/m$.

Στην εργασία αυτή, αποδεικνύουμε ότι οποιοσδήποτε βελτιωμένος εκτιμητής του σ^2 (υπάρχων ή μελλοντικός) της μορφής (6) παράγει αμέσως έναν αντίστοιχο βελτιωμένο εκτιμητή του Σ της μορφής (4), και αντιστρόφως. Έτσι, από τις (7), (8), (9), (10) προκύπτουν αμέσως βελτιωμένοι εκτιμητές του Σ , τύπου Stein (1964), τύπου Brewster and Zidek (1974), τύπου Maruyama (1998) και τύπου Strawderman (1974). Οι δύο πρώτοι εξ αυτών, συμπίπτουν με τους δ_S και δ_{BZ} των Kubokawa et al. (1993), ενώ οι δύο τελευταίοι είναι νέοι βελτιωμένοι εκτιμητές. Είναι πρόδηλο ότι το αποτέλεσμα αυτό ενοποιεί και επεκτείνει τα αποτελέσματα των Kubokawa et al. (1993) και γενικά ανάγει την εκτίμηση του πίνακα διασπορών-συνδιασπορών πολυμεταβλητής κανονικής κατανομής στην εκτίμηση της διασποράς μονομεταβλητής κανονικής κατανομής. Το αποτέλεσμα προκύπτει από μία πολύ απλή ταυτότητα που συνδέει τις συναρτήσεις κινδύνου των δ_ψ και $\delta_\varphi^{\sigma^2}$. Ανάλογα αποτελέσματα ισχύουν και για το τετραγωνικό σφάλμα.

2. ΚΥΡΙΑ ΑΠΟΤΕΛΕΣΜΑΤΑ

Σε αυτή την ενότητα θα δούμε αρχικά πώς συνδέονται τα δύο προβλήματα, αυτό της βελτιωμένης εκτίμησης του πίνακα διασπορών-συνδιασπορών, Σ , πολυμεταβλητής κανονικής κατανομής και της βελτιωμένης εκτίμησης της διασποράς, σ^2 , κανονικής κατανομής. Η παράθεση των αποτελεσμάτων γίνεται με το διαχωρισμό τους, αρχικώς σε αποτελέσματα που αφορούν στη συνάρτηση ζημίας εντροπίας και έπειτα στην τετραγωνική συνάρτηση ζημίας.

2.1 Συνάρτηση ζημίας εντροπίας

Για τη βελτίωση του εκτιμητή δ_0^* που δίνεται από την (3) ως προς τη συνάρτηση ζημίας εντροπίας L_1 , θεωρούμε εκτιμητές δ_ψ της μορφής (4).

Ένα βασικό στοιχείο για τη σύνδεση των δύο προβλημάτων είναι ότι η στατιστική συνάρτηση W μπορεί να γραφεί ως πηλίκο μιας τυχαίας μεταβλητής με μη κεντρική χ -τετράγωνο κατανομή και μιας ανεξάρτητης προς αυτή τυχαίας μεταβλητής με κεντρική χ -τετράγωνο κατανομή, συγκεκριμένα, $W = U/V$ όπου $U \sim \chi_p^2(\tau)$, $\tau = n\theta'\Sigma^{-1}\theta$, $V \sim \chi_{n-p}^2$ (βλ. Shorrock and Zidek, 1976 και Kubokawa et al., 1993). Τότε, εάν $T \sim W_p(n-1, I_p)$, για τις μέσες ζημιές των δ_0^* και δ_ψ , $R(\delta_0^*)$ και $R(\delta_\psi; p, n, \tau)$, αντίστοιχα, μπορούν να αποδειχθούν οι σχέσεις

$$R(\delta_0^*) = EL_1(\delta_0^*, \Sigma)$$

$$= \frac{1}{n} E \operatorname{tr} T + E \frac{p}{n(n-p)} V - E \ln |T| + (p-1) \ln n - \ln \left(\frac{1}{n} + \frac{p}{n(n-p)} \right) - p,$$

$$R(\delta_\psi; p, n, \tau) = EL_1(\delta_\psi, \Sigma)$$

$$= \frac{1}{n} E \operatorname{tr} T + E \psi(W) V - E \ln |T| + (p-1) \ln n - E \ln \left(\frac{1}{n} + \psi(W) \right) - p.$$

Επομένως, αφού $EV = n - p$, η διαφορά των μέσων ζημιών είναι

$$R(\delta_0^*) - R(\delta_\psi; p, n, \tau) = \frac{p}{n} - E \psi(W) V - \ln \frac{1}{n-p} + E \ln \left(\frac{1}{n} + \psi(W) \right)$$

$$= 1 - E \left(\frac{1}{n} + \psi(W) \right) V - \ln \frac{1}{n-p} + E \ln \left(\frac{1}{n} + \psi(W) \right). \quad (11)$$

Ανάλογα, η στατιστική συνάρτηση W_0 στην (6) μπορεί να γραφεί ως πηλίκο μιας τυχαίας μεταβλητής με μη κεντρική χ -τετράγωνο κατανομή και μιας ανεξάρτητης προς αυτή τυχαίας μεταβλητής με κεντρική χ -τετράγωνο κατανομή, συγκεκριμένα, $W_0 = U_0/V_0$ όπου $U_0 = \|X\|^2/\sigma^2 \sim \chi_p^2(\lambda)$, $\lambda = \|\mu\|^2/\sigma^2$, $V_0 = S_0/\sigma^2 \sim \chi_m^2$. Επίσης, οι μέσες ζημιές (ως προς τη συνάρτηση ζημίας εντροπίας) των $\delta_0^{\sigma^2}$ και $\delta_\phi^{\sigma^2}$ είναι αντίστοιχα, $R(\delta_0^{\sigma^2}) = EV_0/m - \ln(1/m) - E \ln V_0 - 1$ και $R(\delta_\phi^{\sigma^2}; p, m, \lambda) = E\phi(W_0)V_0 - E \ln \phi(W_0) - E \ln V_0 - 1$, οπότε η διαφορά των μέσων ζημιών δίνεται από τη σχέση

$$R(\delta_0^{\sigma^2}) - R(\delta_\phi^{\sigma^2}; p, m, \lambda) = 1 - E\phi(W_0)V_0 - \ln(1/m) + E \ln \phi(W_0), \quad (12)$$

επειδή $EV_0 = m$.

Εάν, τώρα, θέσουμε $\psi(t) = \phi(t) - 1/n$, για $\phi(t) > 0$, $t > 0$, αλλά και επειδή για $m = n - p$ οι τυχαίες μεταβλητές W , W_0 και V , V_0 συμπίπτουν ως προς την κατανομή, από τις (11) και (12) έχουμε τη σχέση

$$R(\delta_\psi; p, n, \tau) - R(\delta_0^*) = R(\delta_\phi^{\sigma^2}; p, n - p, \tau) - R(\delta_0^{\sigma^2}), \quad (13)$$

η οποία αποτελεί τον συνδυαστικό κρίκο των δύο προβλημάτων.

Το συμπέρασμα που προκύπτει από τη (13) είναι άμεσο και περιλαμβάνεται στην πρόταση που ακολουθεί.

Πρόταση 1. Έστω $m = n - p$, $\varphi(t) > 0$ και $\psi(t) = \varphi(t) - 1/n$, $t > 0$. Ο $\delta_\psi = (1/n)S + \psi(n\bar{X}'S^{-1}\bar{X})(\bar{X}'S^{-1}\bar{X})^{-1}\bar{X}\bar{X}'$ βελτιώνει τον $\delta_0^* = (1/n)S + p(n(n-p))^{-1}(\bar{X}'S^{-1}\bar{X})^{-1}\bar{X}\bar{X}'$ αν και μόνον αν ο $\delta_\phi^{\sigma^2} = \varphi(|X|^2/S_0)S_0$ βελτιώνει τον $\delta_0^{\sigma^2} = S_0/m$.

Από την Πρόταση 1 γίνεται σαφές ότι όλοι οι βελτιωμένοι εκτιμητές του σ^2 της μορφής (6), είτε οι ήδη υπάρχοντες στη βιβλιογραφία, είτε όσοι προκύπτουν στο μέλλον, επεκτείνονται άμεσα για την εκτίμηση του Σ , και αντιστρόφως.

Ειδικά, με εφαρμογή της Πρότασης 1, από τους εκτιμητές Stein (1964) (βλ. (7)) και Brewster and Zidek (1974) (βλ. (8)) για την εκτίμηση του σ^2 προκύπτουν αντίστοιχοι εκτιμητές για την εκτίμηση του Σ που συμπίπτουν με τους δ_S και δ_{BZ} στις σχέσεις (1) και (5). Επίσης, από τους εκτιμητές Maruyama (1998) (βλ. (9)) και τύπου Strawderman (1974) (βλ. (10)) για την εκτίμηση του σ^2 προκύπτουν νέοι βελτιωμένοι εκτιμητές του Σ που δίνονται από τις ακόλουθες σχέσεις

$$\delta_M = \frac{1}{n}S + \frac{\int_0^1 \lambda^{ap/2} (1 + \lambda n \bar{X}' S^{-1} \bar{X})^{-an/2-1} d\lambda}{\int_0^1 \lambda^{ap/2-1} (1 + \lambda n \bar{X}' S^{-1} \bar{X})^{-an/2-1} d\lambda} \bar{X} \bar{X}', \quad a > 1,$$

$$\delta_{ST} = \frac{1}{n}S + \frac{1}{n\bar{X}'S^{-1}\bar{X}} \frac{p(1+n\bar{X}'S^{-1}\bar{X}) - (n-p)r}{(n-p)(r+1+n\bar{X}'S^{-1}\bar{X})} \bar{X}\bar{X}',$$

$$0 < r \leq \max_\varepsilon \min \left\{ \frac{B(\varepsilon)}{1 - B(\varepsilon)}, \frac{1}{\varepsilon} - 1 \right\},$$

όπου $B(\varepsilon) = \min \left\{ B_*, \left(1 + \frac{n-p}{2p\varepsilon} \frac{\Gamma((n-p)/2 + 2\varepsilon)\Gamma(n/2 + \varepsilon + 1)}{\Gamma((n-p)/2 + \varepsilon)\Gamma(n/2 + 2\varepsilon)} \right)^{-1} \right\}$ και B_* η μοναδική

λύση της εξίσωσης $x^{-1} \ln(1-x) = n/(n-p)$.

2.2 Τετραγωνική συνάρτηση ζημίας

Ο βέλτιστος αναλλοίωτος κατά θέση-κλίμακα εκτιμητής του Σ ως προς την τετραγωνική συνάρτηση ζημίας L_2 είναι ο $\delta_0 = (n+p)^{-1}S$, ενώ ο βέλτιστος εκτιμητής στην κλάση των Kubokawa et al. (1993) στην (2) είναι ο

$$\delta_0^* = (n+p+1)^{-1}S + p((n+p+1)(n-p+2))^{-1}(\bar{X}'S^{-1}\bar{X})^{-1}\bar{X}\bar{X}'.$$

Για τη βελτίωση του δ_0^* θεωρούμε εκτιμητές της μορφής

$$\delta_\psi = \frac{1}{n+p+1}S + \frac{\psi(W)}{\bar{X}'S^{-1}\bar{X}}\bar{X}\bar{X}', \quad W = n\bar{X}'S^{-1}\bar{X}.$$

Αντίστοιχα, ο βέλτιστος αναλλοίωτος κατά θέση-κλίμακα εκτιμητής του σ^2 ως προς την τετραγωνική συνάρτηση ζημίας $L(\delta, \sigma^2) = (\delta/\sigma^2 - 1)^2$ είναι ο $\delta_0^{\sigma^2} = S_0/(m+2)$, για τη βελτίωση του οποίου θεωρούμε εκτιμητές της μορφής $\delta_\varphi^{\sigma^2} = \varphi(W_0)S_0$, $W_0 = \|\mathcal{X}\|^2/S_0$.

Στη συνέχεια, εργαζόμενοι στο ίδιο πνεύμα όπως και στην περίπτωση της συνάρτησης ζημίας εντροπίας προκύπτει ότι η διαφορά των μέσων ζημιών των δ_ψ και δ_0^* είναι

$$R(\delta_\psi; p, n, \tau) - R(\delta_0^*) = \left(\frac{n+2}{n+p+1} \right)^2 \left(R(\delta_\varphi^{\sigma^2}; p, n-p, \tau) - R(\delta_0^{\sigma^2}) \right),$$

όπου $\psi(t) = (n+2)(n+p+1)^{-1}(\varphi(t) - 1/(n+2))$, για $\varphi(t) > 0$, $t > 0$.

Από την τελευταία σχέση προκύπτει το ακόλουθο αποτέλεσμα.

Πρόταση 2. Έστω $m = n - p$, $\varphi(t) > 0$ και $\psi(t) = (n+2)(n+p+1)^{-1} \times (\varphi(t) - 1/(n+2))$, $t > 0$. Ο $\delta_\psi = (n+p+1)^{-1}S + \psi(n\bar{X}'S^{-1}\bar{X})(\bar{X}'S^{-1}\bar{X})^{-1}\bar{X}\bar{X}'$ βελτιώνει τον $\delta_0^* = (n+p+1)^{-1}S + p((n+p+1)(n-p+2))^{-1}(\bar{X}'S^{-1}\bar{X})^{-1}\bar{X}\bar{X}'$ αν και μόνον αν ο $\delta_\varphi^{\sigma^2} = \varphi(\|\mathcal{X}\|^2/S_0)S_0$ βελτιώνει τον $\delta_0^{\sigma^2} = S_0/(m+2)$.

ABSTRACT

The problem of estimation of the variance-covariance matrix, Σ , of a multivariate normal distribution is studied under both entropy and quadratic losses. It is shown that the construction of improved estimators of Σ is reduced to the construction of improved estimators of the variance σ^2 of a (univariate) normal distribution. This reduction leads immediately to the derivation of Stein-type, Brewster and Zidek-type and Strawderman-type estimators for the matrix Σ . This work unifies and extends existing results on the estimation of Σ .

ΑΝΑΦΟΡΕΣ

- Brewster J.F, Zidek J.V. (1974). Improving on equivariant estimators. *Ann. Statist.*, **2**, 21-38.
- Kubokawa , T. (1994). A unified approach to improving equivariant estimators. *Ann. Statist.*, **22**, 290-299.
- Kubokawa , T., Roberts C., Saleh A.K.Md. E. (1992). Empirical Bayes estimation of the covariance matrix of a normal distribution with unknown mean under an entropy loss. *Sankhya A*, **54**, 402-410.
- Kubokawa , T., Honda T., Morita K. Saleh A.K.Md. E. (1993). Estimating a covariance matrix of a normal distribution with unknown mean. *J. Japan. Statist. Soc.*, **23**, 131-144.
- Maruyama Y. (1998). Minimax estimators of a normal variance. *Metrika*, **48**, 209-214.
- Maruyama Y., Strawderman W. (2006). A new class of minimax generalized Bayes estimators of a normal variance. *J. Statist. Plan. Infer.*, **136**, 3822-3836.
- Perron F. (1990). Equivariant estimators of the covariance matrix. *Canadian J. Statist.*, **18**, 179-182.
- Shorrock R.B., Zidek J.V. (1976). An improved estimator of the generalized variance. *Ann. Statist.*, **4**, 629-638.
- Sinha B.K., Ghosh M. (1987). Inadmissibility of the best equivariant estimators of the variance-covariance matrix, the precision matrix and the generalized variance under entropy loss. *Statist. Decisions*, **5**, 201-227.
- Stein C. (1964). Inadmissibility of the usual estimator for the variance of a normal distribution with unknown mean. *Ann. Inst. Statist. Math.*, **16**, 155-160.
- Strawderman W. (1974). Minimax estimation of powers of the variance of a normal population under squared error loss. *Ann. Statist.*, **2**(1), 190-198.

ΣΤΡΑΤΗΓΙΚΕΣ ΑΝΑΛΗΨΗΣ ΡΙΣΚΟΥ ΑΠΟ ΠΑΙΔΙΑ ΚΑΙ ΕΝΗΛΙΚΕΣ

Ζωή Νικηφορίδου¹, Τζένη Παγγέ¹, Θεόδωρος Χατζηπαντελής²

¹ Παιδαγωγικό Τμήμα Νηπιαγωγών, Πανεπιστήμιο Ιωαννίνων,
{znikifor.jpagge}@cc.uoi.gr

² Τμήμα Πολιτικών Επιστημών, Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης,
chadji@polsci.auth.gr

ΠΕΡΙΛΗΨΗ

Τα τελευταία χρόνια έχει παρατηρηθεί έντονη ερευνητική δραστηριότητα σχετικά με την ανάληψη ρίσκου από μικρά παιδιά. Έχει αποδειχθεί ότι ακόμα και από την ηλικία των 4 χρόνων, τα παιδιά αναπτύσσουν και εκδηλώνουν στοιχεία της πιθανολογικής σκέψης καθώς και στρατηγικές εκτίμησης και επεξεργασίας πληροφοριών μέσα σε περιβάλλοντα αβεβαιότητας. Μεθοδολογικά, συνήθως τα παιδιά εμπλέκονται σε πειράματα, αναπροσαρμοσμένα από αυτά που απευθύνονται σε ενήλικες και εφήβους. Συμμετέχουν σε παιχνίδια με πόντους με στόχο την επίτευξη κέρδους ή/και την αποστροφή της ζημιάς.

Στην παρούσα έρευνα εξετάζεται αν τα παιδιά συγκριτικά με τους ενήλικες ακολουθούν κάποια συγκεκριμένη στρατηγική σε ένα ειδικά σχεδιασμένο παιχνίδι ρίσκου στον υπολογιστή. Ο στόχος του παιχνιδιού σχετίζεται με την εκτίμηση και επιλογή ενδεχομένων, μέσα από πολλαπλές δοκιμασίες, ώστε να κερδίσει κανείς τους περισσότερους πόντους και να χάσει τους λιγότερους. Το δείγμα απαρτίζεται από δύο ομάδες. Η πρώτη ομάδα αποτελείται από παιδιά προσχολικής ηλικίας 4-5 χρονών ($N=40$) και η δεύτερη ομάδα αποτελείται από ενήλικες ηλικίας 28-40 χρονών ($N=20$).

Διαφορές εντοπίστηκαν στις στρατηγικές των συμμετεχόντων σχετικά με την περίπτωση του κέρδους και της ζημιάς, το πραγματικό αποτέλεσμα και την ηλικία. Εκπαιδευτικές και μεθοδολογικές προεκτάσεις αναλύονται.

Λέξεις Κλειδιά: στρατηγικές ανάληψης ρίσκου, γραμματισμός του ρίσκου, παιδιά.

1. ΕΙΣΑΓΩΓΗ

Το ρίσκο διακατέχει σχεδόν κάθε ανθρώπινη δραστηριότητα και απόφαση και ως έννοια ενέχει πολλές και ποικίλες διαστάσεις και προσεγγίσεις. Σύμφωνα με τον Adams (2006), η ανάληψη ρίσκου αποτελεί μια διαδικασία εξισορρόπησης ανάμεσα σε γνωστικές και διαισθητικές διεργασίες, στο κέρδος και τη ζημιά, στο πιθανό και το απίθανο ενδεχόμενο και τέλος, στο τυχαίο και το μη τυχαίο. Η αντίληψη του ρίσκου, η διαχείρισή του, ο υπολογισμός του, η επιδίωξή του ή η αποστροφή του αποτελούν αποτέλεσμα υπολογισμών και εκτιμήσεων, με ορθολογικό ή υποκειμενικό τρόπο, είτε σε προσωπικό, είτε σε κοινωνικό επίπεδο. Το ρίσκο είναι μια αβεβαιότητα η οποία

ποσοτικοποιείται με βάση την εμπειρία ή/και τη θεωρία. Ανάλογα με την κατάσταση προβληματικής, το πεδίο εφαρμογής, τα δεδομένα και την προσέγγιση, το ρίσκο αντιμετωπίζεται, συνήθως, ως η πιθανότητα κάποιου ‘ανεπιθύμητου’ ενδεχομένου (Webb, 2006).

Τα τελευταία χρόνια υπάρχει προσανατολισμός στην εκπαιδευτική αξιοποίηση και καλλιέργεια συμπεριφορών και στάσεων λήψης αποφάσεων και εκτίμησης καταστάσεων ρίσκου με ορθολογικό τρόπο με στόχο την πιο επωφελή επιλογή. Ο γραμματισμός του ρίσκου (risk literacy), όπως ορίζεται, αναφέρεται στην ικανότητα να αντιμετωπίζει κανείς την αβεβαιότητα με επίγνωση (Gigerenzer, 2007). Περιλαμβάνει δεξιότητες που σχετίζονται με τη στατιστική σκέψη, την εκτίμηση των πιθανοτήτων, την επεξεργασία πληροφοριών, τη λήψη αποφάσεων, την αλληλεπίδραση ευρετικών μεθόδων και γνωστικών μηχανισμών. Βοηθάει τους μαθητές, ως μελλοντικούς πολίτες, να αναλογίζονται τα πλεονεκτήματα και τα μειονεκτήματα καταστάσεων προκειμένου να παίρνουν λογικές αποφάσεις ζωής σε όλα τα επίπεδα (Boyer, 2006; Rolfe, 2010).

Ήδη σε πολλές χώρες οι Πιθανότητες και η Στατιστική αποτελούν μέρος του αναλυτικού προγράμματος Μαθηματικών στην Προσχολική Ηλικία. Για παράδειγμα, στις ΗΠΑ (National Council of Teachers of Mathematics, 2005) αναφέρεται η ενότητα Ανάλυση Δεδομένων και Πιθανότητες (Data Analysis and Probabilities), στην Αυστραλία (Department of Education and Training Western Australia, 2007) η ενότητα Τυχαίο και Δεδομένα (Chance and Data), στη Βρετανία (Department of Education and Skills, 2001) η ενότητα Διαχείριση Δεδομένων (Data Handling). Ιδιαίτερη έμφαση δίνεται στην ένταξη, τη μεθοδολογία της διδασκαλίας, την ανάπτυξη εκπαιδευτικών σχεδιασμών και την αντιμετώπιση των δυσκολιών ή σφαλμάτων των μαθητών σε στατιστικές έννοιες (Χατζηπαντελής, 2003).

Στην παρούσα εργασία διερευνάται η στάση που κρατούν τα παιδιά προσχολικής ηλικίας απέναντι στο ρίσκο (επιδίωξη ή αποστροφή, risk – seekers vs risk-averters), συγκριτικά με μια ομάδα ενηλίκων, μέσα από ένα παιχνίδι πόντων. Παρουσιάζεται κάποια στρατηγική, όπως για παράδειγμα η *αλλαγή προτίμησης*, όπως έχει προταθεί από τη Θεωρία Προοπτικών των Kahneman and Tversky (1979); Δηλαδή, τα παιδιά ρισκάρουν πιο πολύ όταν είναι να χάσουν παρά όταν είναι να κερδίσουν, όπως συνηθίζεται στους ενήλικες; Τι επιλέγουν μέσα από την επανάληψη του κάθε τεστ 3 φορές; Μπορούν να υπάρξουν εκπαιδευτικές προεκτάσεις στο πλαίσιο του γραμματισμού του ρίσκου;

2. ΜΕΘΟΔΟΛΟΓΙΑ

Το δείγμα αποτέλεσαν 40 παιδιά, ηλικίας 4-6 χρονών και 20 ενήλικες, ηλικίας 28-40 χρονών. Το παιχνίδι – πείραμα ήταν κοινό και στις δυο ομάδες. Υλοποιήθηκαν δυο τεστ, το 1^ο τεστ ‘επιδίωξης κέρδους’ και το 2^ο τεστ ‘αποστροφής ζημιάς’, με επανάληψη 3 φορές.

Χρησιμοποιήθηκαν 4 μεταλλικά κουτιά και αντίστοιχα 4 καρτέλες με χαρούμενα ή θλιμμένα πρόσωπα. Το υλικό του παιχνιδιού παρουσιαζόταν στην οθόνη του υπολογιστή και στη συνέχεια, τα 4 κουτιά χωρίζονταν σε 2 στήλες των 2 κουτιών (Πίνακας 1). Στη μια πλευρά, τη ‘μη ρίσκο’ πλευρά, τοποθετούνταν από 1 καρτέλα

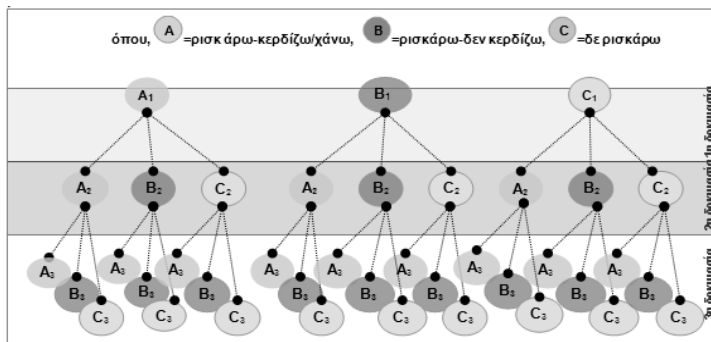
κάτω από κάθε κουτί και στην άλλη πλευρά, τη ‘ρίσκο’ πλευρά, τοποθετούνταν στο ένα κουτί 2 καρτέλες και στο άλλο κουτί καμία καρτέλα. Στο 1^ο τεστ στόχος ήταν η επίτευξη του μεγαλύτερου δυνατού κέρδους, δηλαδή οι 2 πόντοι, ενώ στο 2^ο τεστ στόχος ήταν η αποφυγή της απώλειας, δηλαδή οι 0 πόντοι. Αφού αναποδογυρίζονταν τα κουτιά και ανακατεύονταν στην κάθε πλευρά χωριστά (ρίσκο – μη ρίσκο ή αριστερή και δεξιά στήλη), οι συμμετέχοντες καλούνταν να λάβουν δύο αποφάσεις: σε πρώτο επίπεδο επέλεξαν από ποια πλευρά επιθυμούσαν να παίξουν (την πλευρά ρίσκο ή μη ρίσκο) και σε δεύτερο επίπεδο επέλεξαν ποιο κουτάκι επιθυμούσαν να αναποδογυριστεί προκειμένου να διαπιστώσουν τί και εάν κέρδισαν ή έχασαν.

Πίνακας 1. Μεθοδολογία παιχνιδιού

ΤΕΣΤ	ΕΡΓΑΛΕΙΑ	ΣΧΕΔΙΑΣΜΟΣ
1 (στόχος η επιδίωξη κέρδους)	4 ☐	Μη ρίσκο Ρίσκο
	4 ☺	+1 +2 +1 0
2 (στόχος η αποστροφή ζημιάς)	4 ☐	Ρίσκο Μη Ρίσκο
	4 ☒	0 -1 -2 -1

Στο 1^ο τεστ η πετυχημένη επιλογή θα ήταν το κουτί από την πλευρά ρίσκο που θα έδινε το μεγαλύτερο δυνατό κέρδος, το +2, με πιθανότητα 0.5 (το πάνω δεξιά κουτί στο παράδειγμα του Πίνακα 1). Στο 2^ο τεστ η πετυχημένη επιλογή θα ήταν το κουτί με μηδενικό κόστος, δηλαδή με 0, με πιθανότητα 0.5 (το κάτω αριστερά κουτί στο παράδειγμα του Πίνακα 1). Κάθε τεστ επαναλαμβανόταν 3 φορές και υπήρχε εναλλαγή στις πλευρές ‘ρίσκο’ – ‘μη ρίσκο’. Επομένως η επιλογή του + 1 ή -1 αντίστοιχα τις 3 φορές, θα έδινε συνολικά 3 πόντους με πιθανότητα 1. Η επιδιωκόμενη επιλογή του +2 στο τεστ 1 και του 0 στο τεστ 2 θα έδινε 6 πόντους ή 0 πόντους αντίστοιχα με $(0.5)^3 = 0.125$ πιθανότητα.

Πίνακας 2. Καταγραφή απαντήσεων



Οι απαντήσεις ήταν: A = ρισκάρω – κερδίζω, B = ρισκάρω – δεν κερδίζω, C = δεν ρισκάρω. Καταγράφηκαν δέντρα αποφάσεων, για κάθε φορά επανάληψης των τεστ (Πίνακας 2). Επίσης έγινε ο υπολογισμός του ποσοστού ρίσκου (AUB) σε κάθε

δοκιμασία χωριστά, με βάση το Θεώρημα Bayes, δηλαδή δεδομένου ότι η επιλογή των παιδιών την προηγούμενη φορά ήταν ‘ρискάρω’ ή ‘δεν ρискάρω’.

Τα παιδιά, προκειμένου να κινητοποιηθούν και να εμπλακούν ενεργά στο παιχνίδι, στο 1^ο τεστ κέρδιζαν σφραγίδες για κάθε καρτέλα που τύχαιναν και στο 2^ο τεστ έσβηναν πρόσωπα από μια μπάνκα προσώπων ανάλογα με το αποτέλεσμα της απόφασής τους. Οι ενήλικες απλά σημείωναν τους πόντους που κέρδιζαν ή έχαναν.

3. ΑΠΟΤΕΛΕΣΜΑΤΑ

Στο 1^ο τεστ ‘επιδίωξης κέρδους’, την 1^η φορά, το 62% των παιδιών και το 55% των ενηλίκων επέλεξαν την εκδοχή ‘ρίσκο’.

Ποσοστό 90% των παιδιών και 83% των ενηλίκων που ρίσκαραν και κέρδισαν την 1^η φορά ξαναρίσκαρε τη 2^η φορά,. Ποσοστό 100% των παιδιών και αντίστοιχα 100% των ενηλίκων που ρίσκαρε και έχασε την 1^η φορά ξαναρίσκαρε τη 2^η φορά. τέλος, ποσοστό 40% των παιδιών και 44% των ενηλίκων που δεν ρίσκαρε την 1^η φορά επέλεξε την εκδοχή ‘ρίσκο’ τη 2^η φορά.

Τα αποτελέσματα της 3^{ης} φορές καταγράφονται στον Πίνακα 3, όπου φαίνονται οι αναλογίες των παιδιών και των ενηλίκων που ρίσκαραν, με βάση την επιλογή και το αποτέλεσμα των 2 προηγούμενων φορών (1.1 και 1.2). Παρατηρείται ότι τα παιδιά ρισκάρουν πιο πολύ συγκριτικά με τους ενήλικες. Επίσης, οι ενήλικες επιλέγουν το ρίσκο την 3^η φορά δοθέντος ότι την αμέσως προηγούμενη φορά (1.2), ρίσκαραν και έχασαν (B₂).

Πίνακας 3. Καταγραφή αποτελεσμάτων την 3^η φορά στο 1^ο τεστ

	1.1	A ₁	B ₁	C ₁	A ₁	B ₁	C ₁	A ₁	B ₁	C ₁
	1.2	A ₂	A ₂	A ₂	B ₂	B ₂	B ₂	C ₂	C ₂	C ₂
ΠΑΙΔΙΑ	A ₃ UB ₃	0.88	1	1	0.88	1	1	1	0	0.77
ΕΝΗΛΙΚΕΣ	A ₃ UB ₃	0	0	0	1	1	1	1	0	1

Στο 2^ο τεστ ‘αποστροφής ζημιάς’, την 1^η φορά, το 90% των παιδιών και το 80% των ενηλίκων επέλεξαν την εκδοχή ‘ρίσκο’.

Ποσοστό 96% των παιδιών και 100% των ενηλίκων που ρίσκαραν και κέρδισαν την 1^η φορά ξαναρίσκαρε τη 2^η φορά. Ποσοστό 100% των παιδιών και αντίστοιχα των ενηλίκων που ρίσκαρε και έχασε την 1^η φορά ξαναρίσκαρε τη 2^η φορά. Τέλος, ποσοστό 75% των παιδιών και 100% των ενηλίκων που δεν ρίσκαρε την 1^η φορά, επέλεξε την εκδοχή ‘ρίσκο’ τη 2^η φορά.

Την 3^η φορά (Πίνακας 4), παρουσιάζονται οι αναλογίες των παιδιών και των ενηλίκων που επέλεξαν το ενδεχόμενο ‘ρίσκου’, με βάση την επιλογή και το αποτέλεσμα των 2 προηγούμενων φορών (1.1 και 1.2). Παρατηρείται ότι τα παιδιά

ρискάρουν με υψηλότερα ποσοστά συγκριτικά με τους ενήλικες. Επίσης, τα παιδιά επέλεξαν το ρίσκο με ποσοστό 100%, την 3^η φορά δοθέντος ότι την αμέσως προηγούμενη φορά (1.2), ρίσκαραν και έχασαν (B₂). Παρομοίως, οι ενήλικες επέλεξαν το ρίσκο την 3^η φορά (A₃∪B₃) με υψηλότερα ποσοστά εάν την προηγούμενη φορά είχαν ρισκάρει και δεν κέρδισαν (B₂), παρά εάν είχαν ρισκάρει και κέρδισαν (A₂).

Πίνακας 4. Καταγραφή αποτελεσμάτων την 3^η φορά στο 2^ο τεστ

	1.1	A ₁	B ₁	C ₁	A ₁	B ₁	C ₁	A ₁	B ₁	C ₁
	1.2	A ₂	A ₂	A ₂	B ₂	B ₂	B ₂	C ₂	C ₂	C ₂
ΠΑΙΔΙΑ	A ₃ ∪B ₃	0.88	1	0	1	1	1	1	0	0
ΕΝΗΛΙΚΕΣ	A ₃ ∪B ₃	0	0.75	1	0.71	0.66	1	0	0	1

4. ΣΥΜΠΕΡΑΣΜΑΤΑ – ΣΥΖΗΤΗΣΗ

Στο συγκεκριμένο παιχνίδι πόντων τα παιδιά είχαν μεγαλύτερα ποσοστά ρίσκου από ότι οι ενήλικες. Δεν επέλεξαν την ασφαλή επιλογή, αλλά επεδίωκαν τις σφραγίδες στο 1^ο τεστ και τη μηδενική απώλεια στο 2^ο τεστ. Πρόκειται για ένα ελεγχόμενο παιχνίδι πόντων που δίνει τα δυνατότητα στους χρήστες να κάνουν επιλογές με άμεσο ‘επάθλο’.

Οι συμμετέχοντες και από τις δυο ηλικιακές ομάδες, επέλεξαν την εκδοχή ‘ρίσκο’ με υψηλότερα ποσοστά στην περίπτωση ‘αποστροφής ζημιάς’, παρά στην περίπτωση ‘επιδίωξης κέρδους’. Επίσης, το αποτέλεσμα της προηγούμενης δοκιμασίας φαίνεται να επιδρούσε στην επιλογή της επόμενης δοκιμασίας, κυρίως όταν το αποτέλεσμα ήταν ‘δεν κερδίζω’. Αυτά τα στοιχεία υποστηρίζονται από το φαινόμενο *αλλαγής προτίμησης*, όπως ορίζεται από τη Θεωρία Προοπτικών (Tversky and Kahneman, 1981). Σύμφωνα με αυτή τη Θεωρία, διαπιστώνονται αποκλίσεις από την ορθολογική λήψη αποφάσεων σε καταστάσεις αβεβαιότητας, επειδή υπάρχει υποκειμενική εκτίμηση των κερδών και των ζημιών που σχετίζεται με ευρετικές μεθόδους. Παιδιά και ενήλικες έτειναν να επιλέγουν το ρίσκο ιδιαίτερα εάν μόλις είχαν χάσει, καθώς οι ζημιές X ποσού έχουν μεγαλύτερο συναισθηματικό αντίκτυπο συγκριτικά με την ικανοποίηση κέρδους X ποσού. Στην παρούσα περίπτωση το X ποσό ήταν οι +2 πόντοι στο 1^ο τεστ και οι -2 πόντοι στο 2^ο τεστ.

Η αντιμετώπιση και η διαχείριση του ρίσκου μπορεί να καλλιεργηθεί από την βασική εκπαίδευση με μια δια βίου προοπτική (Gigerenzer, 2007). Μέσα από καταστάσεις προβληματικής, όπως το απλό παιχνίδι της παρούσας έρευνας, που δεν είναι απαραίτητο να άπτονται μόνο των μαθηματικών ή των οικονομικών, αλλά και άλλων γνωστικών πεδίων όπως η αγωγή υγείας, η κυκλοφοριακή αγωγή, η μελέτη περιβάλλοντος, μπορεί να καλλιεργηθεί ο γραμματισμός του ρίσκου. Η ικανότητα ερμηνείας ενός ρίσκου και αποστροφής του (με βάση την εκτίμηση πιθανών

ενδεχομένων) αποτελούν μια από τις πιο σημαντικές δεξιότητες που μπορεί να αναπτύξει κανείς (Boyer, 2006).

Η ένταξη του ρίσκου στο σχολείο βοηθάει τα παιδιά να το αντιμετωπίζουν ως κομμάτι της μάθησης και της προόδου τους. Μελλοντικές προεκτάσεις της παρούσας μελέτης θα μπορούσαν να προσανατολιστούν προς αυτήν την κατεύθυνση. Μέσα από αναπτυξιακά κατάλληλες δραστηριότητες οι μαθητές μπορούν να αναπτύξουν χαρακτηριστικά όπως εμπιστοσύνη, φιλοδοξία, κριτική σκέψη, συνειδητοποίηση, εκτίμηση της πιο επωφελούς απόφασης και ανάπτυξη στρατηγικών ανάληψης ρίσκου. Μπορούν να μάθουν να επιδιώκουν την επιτυχία, καθώς και να εξασκηθούν στην αποτυχία (Rolfé, 2010). Άλλωστε οι εξελίξεις και η σύγχρονη κοινωνία έχουν δώσει στον πολυγραμματισμό πολλές διαστάσεις, όπου ενώ τον 20^ο αιώνα κυρίαρχο ρόλο είχαν η ανάγνωση και γραφή, σήμερα έχουν το ρίσκο, οι πληροφορίες, η τεχνολογία και όχι μόνο.

ABSTRACT

Risk-taking in young children is a field of great interest during the last decades. Even at the age of 4, children have been found to develop probabilistic abilities and strategies of estimating and processing information in situations under uncertainty. Methodologically, children get involved in lab-driven tasks, adjusted from experiments addressed to adults and older participants. They participate in games with points in order to achieve gain or avoid loss.

The current study tests whether children express strategies, if any, compared to adults, while interacting with a computer-based risk-taking game. The aim of the game is to evaluate and choose options, through multiple tasks, in order to gain as many points as possible and respectively lose the minimum possible points. The sample was consisted of two groups, a group of preschoolers aged 4-5 ($N=40$) and a group of adults, aged 28-40 ($N=20$).

Differences were found concerning the domains of gain and loss, the outcome, as well as age-based decisions. Further educational and methodological implications are discussed.

ΑΝΑΦΟΡΕΣ

- Adams J. (2006). *Risk*. Taylor & Francis: Routledge.
- Boyer, T. (2006). The development of risk-taking: A multi-perspective review. *Developmental Review*, 26(3), 291-345.
- Department of Education and Training Western Australia (2007). *First Steps in Mathematics: Chance and Data*. Rigby Harcourt Education.
- Gigerenzer, G. (2007). *Gut Feelings: The Intelligence of the Unconscious*. Penguin/Viking. 132.
- National Council of Teachers of Mathematics (NCTM). (2005). *Principles and Standards for School Mathematics*. Reston, VA: Author
- Rolfé H. (2010). *Learning to take risks, learning to succeed*. National Endowment for Science, Technology and the Arts, UK. \

- Kahneman, D. and Tversky, A. (1979). Prospect theory: an analysis of decision under risk, *Econometrica*. Vol. 47, 263–91.
- Tversky, A. and Kahneman, D. (1981). The Framing of Decisions and the Psychology of Choice. *Science*. Vol 211(4481) pp.453-458
- Webb, S. A. (2006). *Social Work in a Risk Society: Social and Political Perspectives*. London: Palgrave Macmillan.
- Χατζηπαντελής Θ (2003). Η διαδρομή του παιδιού στη στατιστική και οι πιθανότητες από την προσχολική ηλικία έως την ενηλικίωση, 20^ο Πανελλήνιο Συνέδριο Μαθηματικής Παιδείας, Βέροια.



ΕΝΤΟΠΙΣΜΟΣ ΑΜΕΣΩΝ ΣΥΖΕΥΞΕΩΝ ΣΕ ΠΟΛΥΜΕΤΑΒΛΗΤΑ ΣΥΣΤΗΜΑΤΑ

Παπάνα Αγγελική, Κουγιουμτζής Δημήτρης
Γενικό Τμήμα Πολυτεχνικής σχολής Α.Π.Θ.
agrapana@gen.auth.gr, dkugiu@gen.auth.gr

ΠΕΡΙΛΗΨΗ

Στην εργασία αυτή εξετάζονται δύο μέτρα άμεσης αιτιότητας σε χρονοσειρές, η δεσμευμένη αιτιότητα Granger (CGC) και η μερική εντροπία μεταφοράς (PTE), την οποία πρόσφατα αναπτύξαμε ως επέκταση της εντροπίας μεταφοράς. Η στατιστική σημαντικότητα των μέτρων εξετάζεται με έλεγχο τυχαιοποίησης, ενώ για το CGC χρησιμοποιείται και παραμετρικός έλεγχος, καθώς η μηδενική κατανομή του είναι γνωστή. Τα αποτελέσματα προσομοιώσεων σε γραμμικά συστήματα έδειξαν πως το CGC αναδεικνύει τις αληθείς άμεσες σχέσεις αιτιότητας με μεγαλύτερη ακρίβεια όταν το μήκος των χρονοσειρών είναι μικρό, ενώ το PTE τείνει προς την ίδια ακρίβεια καθώς το μήκος αυξάνει. Σε μη-γραμμικά συστήματα, το PTE υπερέχει φανερά του CGC, ακόμα και για μικρά μήκη χρονοσειρών. Το βασικό μειονέκτημα του PTE είναι ότι η απαίτηση σε μέγεθος δεδομένων αυξάνει εκθετικά με το πλήθος των μεταβλητών στο σύστημα, αφού βασίζεται στην εκτίμηση της από κοινού συνάρτησης κατανομής όλων των μεταβλητών. Τέλος, εφαρμόσαμε το CGC και PTE σε διεθνείς χρηματιστηριακές αγορές για να διερευνήσουμε κατά πόσο ισχυρές αγορές, όπως των ΗΠΑ, Γερμανίας και Ιαπωνίας, έχουν άμεση επίδραση σε άλλες αγορές.

Λέξεις Κλειδιά: άμεση αιτιότητα κατά Granger, εντροπία μεταφοράς, έλεγχος τυχαιοποίησης

1. ΕΙΣΑΓΩΓΗ

Η εκτίμηση αλληλεπιδράσεων σε πολύ-μεταβλητές χρονοσειρές μας επιτρέπει να κατανοήσουμε τις σχέσεις μεταξύ των μεταβλητών ή τις συζεύξεις (coupling) μεταξύ υπο-συστημάτων και γενικότερα τη δομή ενός πολύ-μεταβλητού και συχνά πολύπλοκου δυναμικού συστήματος. Χαρακτηριστικά παραδείγματα τέτοιων συστημάτων είναι η χρηματιστηριακή αγορά (Schweitzer et al, 2009) και το δυναμικό του εγκεφάλου (Dana et al, 2009). Η αλληλεπίδραση παραπέμπει στην κατά Granger αιτιότητα (Granger causality) δύο τυχαίων μεταβλητών X_1 και X_2 (Granger 1969). Αν υπάρχει μια τρίτη μεταβλητής X_3 , η αιτιώδης σχέση των X_1 και X_2 μπορεί να είναι είτε άμεση ή έμμεση, δηλαδή μέσω της σχέσης της κάθε μιας με τη X_3 .

Στην εργασία αυτή εξετάζουμε δύο μέτρα άμεσης αιτιότητας, ένα γραμμικό μέτρο που βασίζεται στην κατά Granger αιτιότητα, και ένα μη γραμμικό μέτρο, το οποίο είναι επέκταση της εντροπίας μεταφοράς (Schreiber, 2000).

2. ΜΕΤΡΑ ΑΙΤΙΟΤΗΤΑΣ ΚΑΙ ΕΛΕΓΧΟΙ ΣΗΜΑΝΤΙΚΟΤΗΤΑΣ

Η «αιτιότητα Granger» βασίζεται στην ιδέα ότι αν η πρόβλεψη των τιμών μιας χρονοσειράς X_1 βελτιώνεται με την χρήση των τιμών της X_2 , τότε λέμε ότι η X_2 «οδηγεί» την X_1 . Έστω δύο χρονοσειρές $\{x_1(t)\}$ και $\{x_2(t)\}$, $t=1, \dots, n$, στις οποίες προσαρμόζεται ένα διμεταβλητό αυτοπαλινδρομούμενο μοντέλο τάξης P

$$x_1(t+1) = \sum_{j=0}^{P-1} a_{1,j} x_1(t-j) + \sum_{j=0}^{P-1} a_{2,j} x_2(t-j) + \varepsilon_1(t+1) \quad (1)$$

όπου $a_{1,j}, a_{2,j}$ είναι οι συντελεστές του μοντέλου και ε_1 τα υπόλοιπα της προσαρμογής του μοντέλου. Εάν η διασπορά των υπολοίπων του «μη-περιορισμένου» (unrestricted) μοντέλου στην (1), s_{1U}^2 , είναι μικρότερη από αυτήν για το «περιορισμένο» (restricted) μοντέλο που δε συμπεριλαμβάνει τη X_2 (το δεύτερο άθροισμα στην (1)), s_{1R}^2 , τότε η μεταβλητή X_2 «οδηγεί κατά Granger» την X_1 (Granger, 1969). Το μέγεθος της επίδρασης της X_2 στη X_1 δίνεται από το δείκτη της αιτιότητας κατά Granger που ορίζεται ως $GC_{X_2 \rightarrow X_1} = \ln(s_{1R}^2 / s_{1U}^2)$. Θεωρώντας επιπλέον τη μεταβλητή X_3 στο μοντέλο της (1) (προσθέτοντας έναν επιπλέον όρο αθροίσματος ως προς X_3) και ορίζοντας το μη-περιορισμένο και περιορισμένο μοντέλο κατά τον ίδιο τρόπο, ο δείκτης της δεσμευμένης αιτιότητας κατά Granger είναι αντίστοιχα $CGC_{X_2 \rightarrow X_1 | X_3} = \ln(s_{1R}^2 / s_{1U}^2)$. Με τον ίδιο τρόπο ο δείκτης CGC ορίζεται για περισσότερες μεταβλητές (Geweke, 1982). Θεωρώντας τον F-έλεγχο σημαντικότητας για κάθε έναν από τους P συντελεστές $a_{2,j}$ και με την απαραίτητη διόρθωση λόγω πολλαπλών ελέγχων (π.χ. Bonferroni), έχουμε τον παραμετρικό έλεγχο σημαντικότητας για τα GC και CGC, ώστε να ελέγξουμε την μηδενική υπόθεση ότι η μεταβλητή X_2 δεν «οδηγεί» την X_1 (Granger, 1969).

Η εντροπία μεταφοράς (transfer entropy) ορίζει την αιτιότητα κατά Granger με εντροπίες αντί γραμμικών μοντέλων. Θεωρώντας το διάνυσμα των τελευταίων P τιμών για τη X_1 , $\mathbf{x}_{1,t} = [x(t), x(t-1), \dots, x(t-P+1)]'$, και αντίστοιχα για τη X_2 , η εντροπία μεταφοράς από την μεταβλητή X_2 στη X_1 ορίζεται ως (Schreiber, 2000)

$$\begin{aligned} TE_{X_2 \rightarrow X_1} &= H(x_{1,t+1} | \mathbf{x}_{1,t}) - H(x_{1,t+1} | \mathbf{x}_{1,t}, \mathbf{x}_{2,t}) \\ &= H(\mathbf{x}_{1,t}, \mathbf{x}_{2,t}) - H(x_{1,t+1}, \mathbf{x}_{1,t}, \mathbf{x}_{2,t}) + H(x_{1,t+1}, \mathbf{x}_{1,t}) - H(\mathbf{x}_{1,t}) \end{aligned} \quad (2)$$

όπου $H(x)$ είναι η εντροπία Shannon της μεταβλητής X , $H(x) = -\sum p(x) \ln p(x)$.

Πρόσφατα επεκτείναμε την εντροπία μεταφοράς για πολλές μεταβλητές στη λεγόμενη μερική εντροπία μεταφοράς (partial transfer entropy) που μετρά την άμεση επίδραση της X_2 στη X_1 παρουσία όλων των υπολοίπων μεταβλητών (Parana and Kugiumtzis, 2011). Στην περίπτωση τριών μεταβλητών X_1, X_2, X_3 η μερική εντροπία μεταφοράς ορίζεται ως (Parana and Kugiumtzis, 2011)

$$PTE_{X_1 \rightarrow X_1 | X_3} = H(x_{1,t+1} | \mathbf{x}_{1,t}, \mathbf{x}_{3,t}) - H(x_{1,t+1} | \mathbf{x}_{1,t}, \mathbf{x}_{2,t}, \mathbf{x}_{3,t}). \quad (3)$$

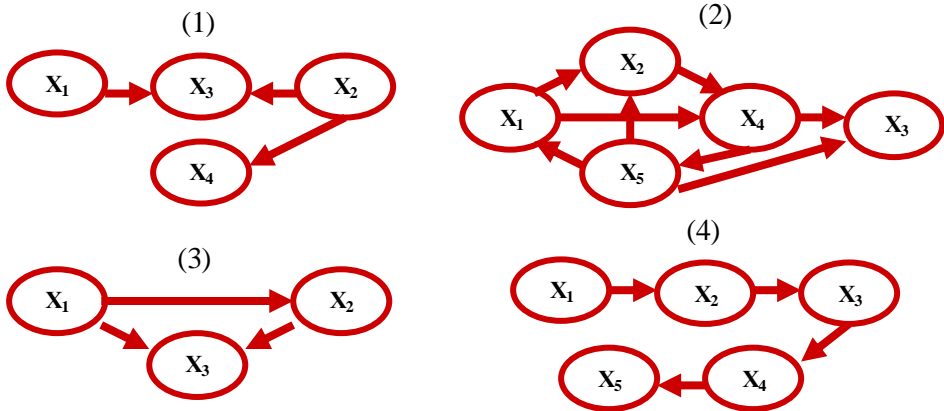
Γενικά τα στοιχεία των διανυσμάτων $\mathbf{x}_{1,t}, \mathbf{x}_{2,t}, \mathbf{x}_{3,t}$ μπορούν να διαφέρουν σε πλήθος και να επιλεγθούν με μη-ομοίμορφο τρόπο, αλλά για τις ανάγκες της εργασίας θα περιοριστούμε μόνο στην παράμετρο P όπως και για το CGC.

Η στατιστική σημαντικότητα της PTE εξετάστηκε με έλεγχο τυχαιοποίησης, καθώς ο εκτιμητής της δεν ακολουθεί γνωστή κατανομή. Δημιουργούμε M υποκατάστατες χρονοσειρές μήκους n ώστε να είναι συνεπείς ως προς την H_0 : η X_2 δεν επιδρά στη X_1 . Για να αποτρέψουμε τυχόν επίδραση της X_2 στη X_1 , χωρίς όμως να αλλάξουμε τη δυναμική στην κάθε μια, επιλέγουμε τυχαία έναν αριθμό $d < n$ και οι d πρώτες τιμές της χρονοσειράς $\{x_2(t)\}$ μεταφέρονται στο τέλος της. Οι άλλες χρονοσειρές μένουν ως έχουν. Υπολογίζουμε την PTE στην αρχική πολυ-μεταβλητή χρονοσειρά, έστω q_0 , και σε κάθε μια από τις υποκατάστατες πολυ-μεταβλητές χρονοσειρές, έστω q_1, q_2, \dots, q_M . Αν το q_0 βρίσκεται στην ουρά της εμπειρικής κατανομής των q_1, \dots, q_M , τότε η H_0 απορρίπτεται. Για δίπλευρο έλεγχο, αν r_0 είναι η τάξη του q_0 στην κατάταξη κατά αύξουσα σειρά των q_0, q_1, \dots, q_M η p -τιμή του ελέγχου είναι $2(r_0 - 0.326)/(M + 1 + 0.348)$ αν $r_0 < (M + 1)/2$ και $2[1 - (r_0 - 0.326)/(M + 1 + 0.348)]$ αν $r_0 \geq (M + 1)/2$, εφαρμόζοντας τη διόρθωση των Yu and Huang (2001).

3. ΠΡΟΣΟΜΟΙΩΣΕΙΣ ΣΕ ΓΝΩΣΤΑ ΣΥΣΤΗΜΑΤΑ

Η αξιολόγηση των μέτρων CGC, PTE γίνεται μέσω Monte Carlo προσομοιώσεων σε πολυ-μεταβλητά συζευγμένα συστήματα (coupled systems): (1) σύστημα τεσσάρων γραμμικά συζευγμένων αυτοπαλινδρομούμενων διαδικασιών τάξης 5, VAR(5) (Winterhalder, 2005, Eq.12), (2) σύστημα πέντε γραμμικά συζευγμένων αυτοπαλινδρομούμενων διαδικασιών (Schelter, 2006), (3) σύστημα τριών διαδικασιών με γραμμικές και μη-γραμμικές συζεύξεις (Gourevitch, 2006, Model 7), (4) σύστημα πέντε συζευγμένων απεικονίσεων Henon, για $c=0.3$ (Papana and Kugiumtzis, 2011, όπως το σύστημα S2). Οι συζεύξεις των 4 συστημάτων δίνονται στο Σχήμα 1.

Σχήμα 1. Συζεύξεις μεταξύ μεταβλητών για τα συστήματα προσομοιώσεων.

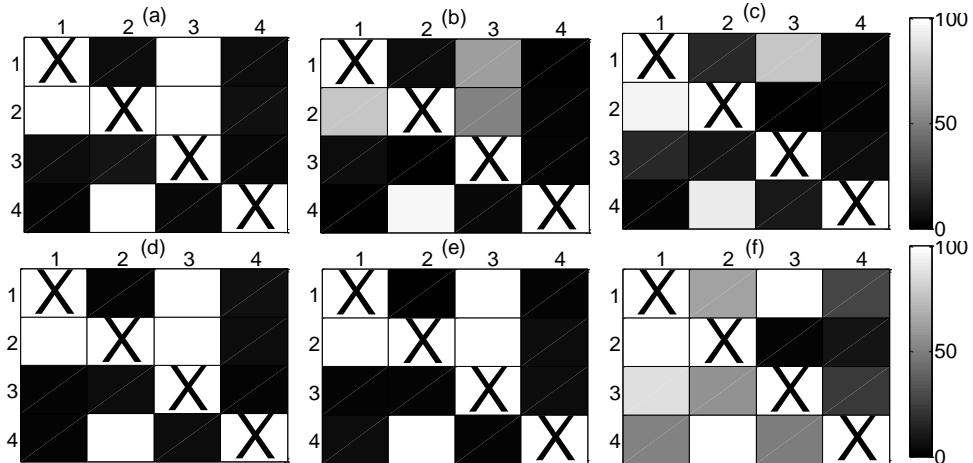


Για κάθε σύστημα δημιουργούμε 100 πραγματοποιήσεις για μήκη χρονοσειρών $n = 512$ και 2048 και για κάθε πραγματοποίηση υπολογίζονται τα μέτρα για τιμές της παραμέτρου P που ορίζονται με βάση την πολυπλοκότητα των συστημάτων. Επίσης γίνεται έλεγχος σημαντικότητας για κάθε μέτρο και τιμή του P . Για το PTE γίνεται έλεγχος τυχαιοποίησης για $M=100$, ενώ για το CGC γίνεται έλεγχος τυχαιοποίησης αλλά και ο παραμετρικός έλεγχος που αναφέρθηκε στην αρχή του εδαφίου 2 (για

λόγους σύγκρισης). Για να εξετάσουμε τη στατιστική σημαντικότητα κάθε μέτρου για κάθε συζευγμένο σύστημα προσομοιώσεων, υπολογίσαμε τα ποσοστά απόρριψης της μηδενικής υπόθεσης H_0 της μη ύπαρξης αιτιότητας (p -τιμή < 0.05), από τις τιμές των δύο μέτρων για όλα τα ζεύγη μεταβλητών.

Για το σύστημα 1 και θέτοντας $P=5$, όση και η πραγματική τάξη του συστήματος, το CGC εντοπίζει σωστά όλες τις άμεσες σχέσεις αιτιότητας με παραμετρικό έλεγχο και για μικρό μήκος χρονοσειρών, όπως δίνεται στο Σχήμα 2a για $n=512$. Η απόρριψη της H_0 είναι στο 100% όταν υπάρχουν άμεσες συζεύξεις (λευκά κελιά) και στο 0% όταν δεν υπάρχουν (μαύρα κελιά). Το CGC δεν αποδίδει το ίδιο καλά σε έλεγχο τυχαιοποίησης, δίνοντας μικρά ποσοστά απόρριψης στις 3 από τις 4 άμεσες συζεύξεις, ενώ το PTE δίνει ακόμα μικρότερα ποσοστά απόρριψης (Σχήμα 2b και c). Αυξάνοντας το n , ο έλεγχος τυχαιοποίησης με CGC βελτιώνεται στα επίπεδα του παραμετρικού ελέγχου, ενώ με PTE εντοπίζει σωστά σε ποσοστό 100% τις 3 από τις 4 άμεσες συζεύξεις (αποτυγχάνει στην $X_2 \rightarrow X_3$), αλλά εντοπίζει σε σημαντικά ποσοστά και έμμεσες συζεύξεις (Σχήματα 2d-f για $n=2048$).

Σχήμα 2. Ποσοστά απόρριψης για έλεγχο σημαντικότητας άμεσης αιτιότητας για το σύστημα 1: (a) CGC και F-έλεγχος, (b) και (c) CGC και PTE αντίστοιχα και έλεγχος τυχαιοποίησης, $P=5$ και $n=512$. Ομοίως στα (d) – (f) αλλά για $n = 2048$.

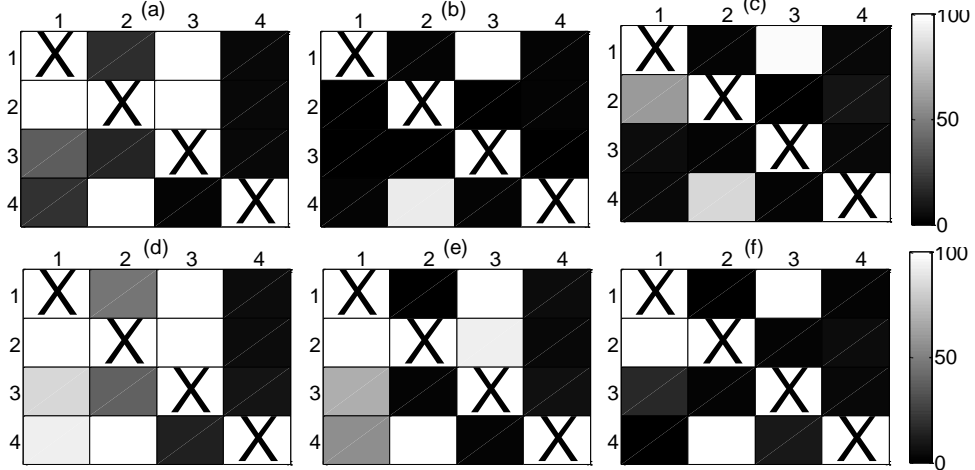


Η αδυναμία του PTE για μεγάλη τάξη P και ιδιαίτερα για μικρά n αιτιολογείται από την αδυναμία εκτίμησης της από κοινού συνάρτησης κατανομής σε μεγάλες διαστάσεις με μικρά δείγματα. Μειώνοντας την τάξη σε $P=2$, το CGC εξακολουθεί να είναι πιο κατάλληλο από το PTE μόνο για μικρό μήκος χρονοσειρών ($n=512$), ενώ για $n=2048$ δίνει χειρότερα αποτελέσματα από το PTE, το οποίο εξακολουθεί να αποτυγχάνει να εντοπίσει τη σύζευξη $X_2 \rightarrow X_3$ (δες Σχήμα 3).

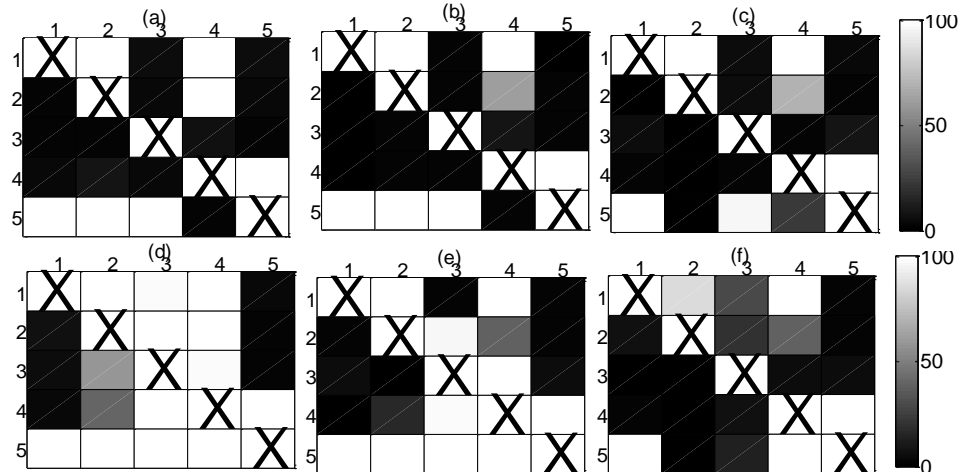
Για το σύστημα 2, όπως δίνεται στο Σχήμα 4 για $n=2048$, το CGC εντοπίζει σωστά τις συζεύξεις για $P=4$ και μόνο για έλεγχο τυχαιοποίησης το ποσοστό της άμεσης σύζευξης $X_2 \rightarrow X_4$ είναι μικρότερο του 100%, ενώ το PTE δεν εντοπίζει επίσης την άμεση σχέση $X_5 \rightarrow X_2$. Για $P=2$, όλοι έλεγχοι δεν αποδίδουν σωστά, με το CGC να εντοπίζει επιπρόσθετα με τις σωστές και ψευδείς άμεσες συζεύξεις (και με τους δύο

ελέγχους), ενώ το PTE να μην προσθέτει ψευδείς συζεύξεις αλλά να αδυνατεί να εντοπίσει τις πραγματικές άμεσες συζεύξεις.

Σχήμα 3. Ομοίως με Σχήμα 2, αλλά για $P=2$.



Σχήμα 4. Ομοίως με Σχήμα 2, αλλά για το σύστημα 2 για $n=2048$ (α) $P=4$ και (β) $P=2$.

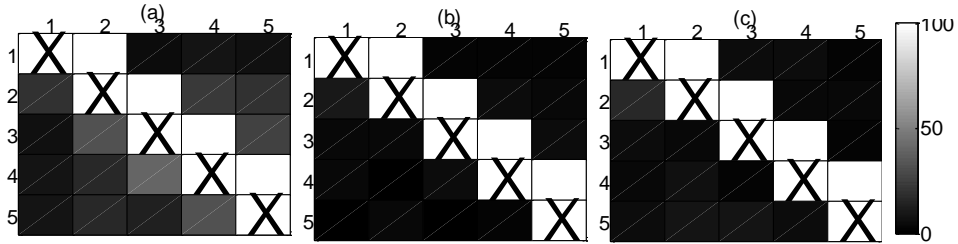


Το σύστημα 3 έχει γραμμικές και μη-γραμμικές σχέσεις αιτιότητας. Για την πραγματική τάξη του συστήματος, $P=1$, αλλά και για $P=2$, το CGC και με τους δύο ελέγχους εντοπίζει την γραμμική συζεύξη $X_2 \rightarrow X_3$, αλλά αποτυγχάνει να εντοπίσει τις άμεσες συζεύξεις $X_1 \rightarrow X_2$ και $X_1 \rightarrow X_3$, ακόμα και για μεγάλο n . Αντίθετα το PTE εντοπίζει όλες τις άμεσες συζεύξεις.

Το PTE υπερέχει επίσης στο σύστημα 4, που αποτελείται από συζευγμένες χαοτικές απεικονίσεις. Για μεγάλα n , το PTE εντοπίζει σε ποσοστό 100% όλες τις άμεσες συζεύξεις και καθόλου τις έμμεσες, ενώ το CGC με τον F-έλεγχο εντοπίζει και έμμεσες συζεύξεις, το οποίο διορθώνει ο έλεγχος τυχαιοποίησης, όπως δίνεται στο Σχήμα 5 για $n=2048$. Για μικρότερα n , το ποσοστό απόρριψης για τις σωστές

άμεσες συζεύξεις μειώνεται με το PTE, ενώ το CGC με τον F-έλεγχο συνεχίζει να εντοπίζει και έμμεσες συζεύξεις.

Σχήμα 5. Ομοίως με Σχήμα 2, αλλά για το σύστημα 4, $P=2$ και $n=2048$.



4. ΕΦΑΡΜΟΓΗ ΣΤΙΣ ΣΧΕΣΕΙΣ ΔΙΕΘΝΩΝ ΑΓΟΡΩΝ

Για να μελετήσουμε άμεσες και έμμεσες σχέσεις στις διεθνείς χρηματιστηριακές αγορές, χρησιμοποιήσαμε σταθμισμένους ημερήσιους δείκτες αγοράς κεφαλαίων σε 23 αναπτυγμένες αγορές της Ευρώπης, Βόρειας Αμερικής και Ασίας για την περίοδο 5/3/2004 ως 5/3/2009 (πηγή: Morgan Stanley Capital International market). Οι 23 αγορές δίνονται στον Πίνακα 1.

Πίνακας 1. Οι 23 αγορές που χρησιμοποιήθηκαν στην μελέτη.

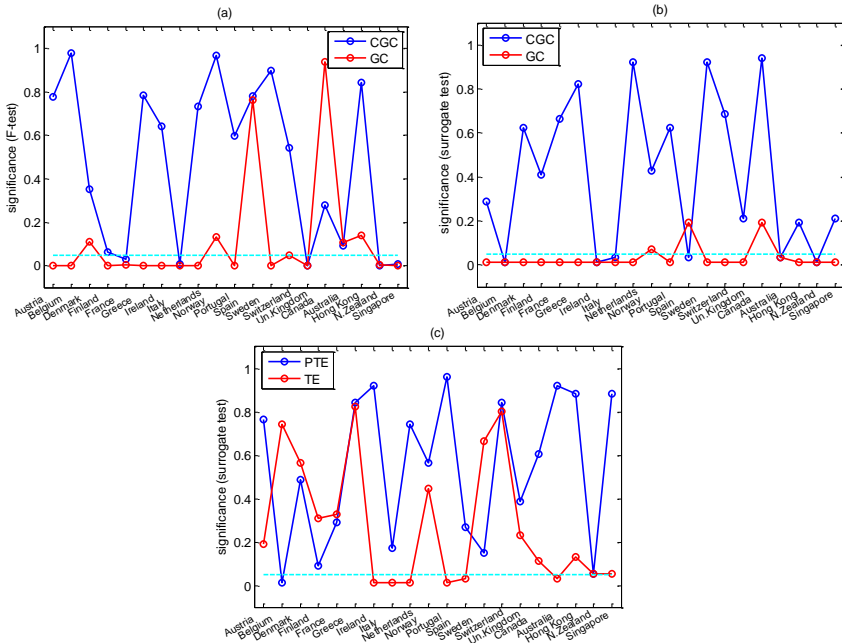
Ευρώπη				Β. Αμερική	Ασία
Αγγλία	Γερμανία	Ελβετία	Δανία	Η.Π.Α	Αυστραλία
Αυστρία	Ολλανδία	Ελλάδα	Ιταλία	Καναδάς	Χονγκ Κονγκ
Βέλγιο	Νορβηγία	Ιρλανδία	Σουηδία		Ιαπωνία
Γαλλία	Πορτογαλία	Ισπανία	Φιλανδία		Ν. Ζηλανδία
					Σιγκαπούρη

Θεωρώντας 3 ισχυρές αγορές, τη Γερμανία στην Ευρώπη, την Η.Π.Α. στην Β. Αμερική και την Ιαπωνία στην Ασία, εξετάσαμε αν πράγματι αυτές οι τρεις αγορές έχουν άμεση επίδραση στις υπόλοιπες. Κάνουμε τους ελέγχους σημαντικότητας με τα στατιστικά CGC και PTE, θέτοντας ως «οδηγό» κάθε μια από τις 3 ισχυρές αγορές, «οδηγούμενη» μια από τις υπόλοιπες 20 αγορές, αφαιρώντας πάντα την επίδραση των άλλων δύο ισχυρών αγορών. Για να αξιολογήσουμε την αναγκαιότητα μέτρων άμεσης αιτιότητας, κάνουμε τους ίδιους ελέγχους με στατιστικά άμεσης και έμμεσης αιτιότητας, το δείκτη αιτιότητας Granger (GC) και την εντροπία μεταφοράς (TE).

Τα διμεταβλητά μέτρα GC και TE εντοπίζουν συζεύξεις μεταξύ των περισσότερων ζευγών αγορών, με το TE να είναι πιο συντηρητικό, σε αντίθεση με τα μέτρα άμεσης αιτιότητας CGC και PTE τα οποία εντοπίζουν λίγες άμεσες συζεύξεις. Για παράδειγμα, για την επίδραση της Γερμανίας στις άλλες αγορές στο Σχήμα 6 (αφαιρώντας την επίδραση των Η.Π.Α. και Ιαπωνίας στον υπολογισμό των CGC και PTE), παρατηρούμε πως σύμφωνα με το GC και με παραμετρικό και μη-παραμετρικό έλεγχο η Γερμανία έχει στατιστικά σημαντική επίδραση σε 14 και 17 αγορές, αντίστοιχα, ενώ σύμφωνα με το CGC έχει στατιστικά σημαντική άμεση επίδραση σε μόλις 5 και 6 αγορές αντίστοιχα. Το TE εντοπίζει 5 σημαντικές σχέσεις αιτιότητας ενώ το PTE μόνο 2 σημαντικές σχέσεις άμεσης αιτιότητας. Έχοντας ως «οδηγό» την

αγορά των Η.Π.Α. προκύπτουν περισσότερες στατιστικά σημαντικές σχέσεις αιτιότητας με αντίστοιχη όμως μείωση στις άμεσες σχέσεις. Αντίθετα η αγορά της Ιαπωνίας έχει πολύ λιγότερες στατιστικά σημαντικές σχέσεις αιτιότητας που επικεντρώνονται στις αγορές της Ασίας.

Σχήμα 6. Στατιστική σημαντικότητα (p -τιμή) για την επίδραση της Γερμανίας σε 20 αγορές (με αφαίρεση επίδρασης της Η.Π.Α. και Ιαπωνίας για CGC και PTE) για το GC και CGC από F -έλεγχο στο (a) και έλεγχο τυχαιοποίησης στο (b) και για το TE και PTE από έλεγχο τυχαιοποίησης στο (c).



7. ΣΥΜΠΕΡΑΣΜΑΤΑ

Στην εργασία αυτή μελετήθηκαν δύο μέτρα άμεσης αιτιότητας και προτάθηκε ένας έλεγχος σημαντικότητας με τυχαιοποιημένες χρονοσειρές. Το γραμμικό μέτρο της δεσμευμένης αιτιότητας Granger (CGC) είδαμε ότι υπερέχει σε γραμμικά συζευγμένα συστήματα, ενώ το μη-γραμμικό μέτρο της μερικής εντροπίας μεταφοράς (PTE) σε μη γραμμικά συστήματα. Και τα δύο μέτρα παρουσιάζουν ευαισθησία στην επιλογή της παραμέτρου τάξης P , με τον έλεγχο με CGC να εντοπίζει και έμμεσες σχέσεις αιτιότητας όταν το P είναι ακατάλληλα μικρό. Το σημαντικό μειονέκτημα του PTE είναι οι απαιτήσεις σε δεδομένα που μεγαλώνουν όταν εμπλέκονται περισσότερες μεταβλητές. Από την άλλη, η απόδοση του ελέγχου με CGC σε μη-γραμμικά συστήματα δεν βελτιώνεται με το μέγεθος των δεδομένων. Για το CGC, ο τυχαιοποιημένος έλεγχος είναι πιο συντηρητικός από τον F -έλεγχο, και συνήθως δεν εντοπίζει συζεύξεις αν δεν υπάρχουν. Η εφαρμογή στις διεθνείς χρηματιστηριακές αγορές έδειξε πως τα μέτρα άμεσης αιτιότητας είναι απαραίτητα στον εντοπισμό άμεσης επίδρασης της μιας αγοράς σε μια άλλη.

ABSTRACT

In this work, we examine two causality measures that detect direct causal effects, conditional Granger causality (CGC) and partial transfer entropy (PTE), which we recently developed as an extension to transfer entropy. The statistical significance of the measures is evaluated by randomization test, while for CGC the F-test can also be used to test the hypotheses of no causal effects. The results of simulations on linear systems showed that CGC detects the correct direct causalities with high accuracy when the time series length is small, while PTE obtains this accuracy only for larger time series. For nonlinear systems, PTE outperforms CGC even for small time series length. The main down-side of PTE is that data requirement increases exponentially with the number of variables in the system because it relies on the estimation of joint probabilities. Finally, we applied CGC and PTE to world finance markets in order to investigate whether USA, Germany and Japan, have direct effects on other markets.

ΑΝΑΦΟΡΕΣ

- Dana S.K., Roy P.K. and Kurths J. (Eds.) (2009). *Complex Dynamics in Physiological Systems: From Heart to Brain, Series: Understanding Complex Systems*, Springer, Berlin / Heidelberg.
- Geweke J. (1982). Measurement of linear dependence and feedback between multiple time series. *Journal of the American Statistical Association*, **77** (378), 304–313.
- Granger C.W.J. (1969). Investigating causal relations by econometric models and cross-spectral methods, *Econometrica*, **37** (3), 424–438.
- Gourevitch B., Jeannès R.B. and Faucon G. (2006). Linear and nonlinear causality between signals: methods, examples and neurophysiological applications. *Biological Cybernetics* **95**, 349–369.
- Papana A. and Kugiumtzis D. (2011). Detection of direct causal effects and application to epileptic electroencephalograms. *International Journal of Bifurcation and Chaos*, accepted.
- Seth A.K. (2010). A MATLAB toolbox for Granger causal connectivity analysis. *Journal of Neuroscience Methods*, **186**, 262–273.
- Schelter B., Winterhalder M., Hellwig B., Guschlbauer B., Lucking C.H. and J. Timmer (2006). Direct or indirect? Graphical models for neural oscillators. *Journal of Physiology* **99**, 37–46.
- Schreiber T. (2000). Measuring information transfer. *Physical Review Letters*, **85** (2), 461–464.
- Schweitzer F., Fagiolo G., Sornette D., Vega-Redondo F., Vespignani A. and White D.R. (2009). Economic networks: the new challenges, *Science*, **325**, 422–425.
- Winterhalder M., Schelter B., Hesse W., Schwab K., Leistriz L., Klan D, Bauer R., Timmer J. and Witte H. (2005) Comparison of linear signal processing techniques to infer directed interactions in multivariate neural systems. *Signal Processing*, **85**, 2137–2160.

ΕΚΤΙΜΗΣΗ ΤΑΞΗΣ ΤΗΣ ΜΑΡΚΟΒΙΑΝΗΣ ΑΛΥΣΙΔΑΣ ΜΕΣΩ ΤΗΣ ΔΕΣΜΕΥΜΕΝΗΣ ΑΜΟΙΒΑΙΑΣ ΠΛΗΡΟΦΟΡΙΑΣ

*M. Παπαπέτρου*¹, *Δ. Κουγιουμτζής*²

Γενικό Τμήμα, Πολυτεχνική Σχολή, Α.Π.Θ

¹maria.pap@gen.auth.gr, ²dkugiu@gen.auth.gr

ΠΕΡΙΛΗΨΗ

Η παρούσα εργασία αναφέρεται στην εκτίμηση της τάξης μιας Μαρκοβιανής αλυσίδας μέσω της δεσμευμένης αμοιβαίας πληροφορίας (Conditional Mutual Information (CMI)). Για μια Μαρκοβιανή αλυσίδα K συμβόλων $\{x_i\}$ ορίζεται η CMI τάξης m , $I_c(m)$, ως η αμοιβαία πληροφορία των x_i και x_{i+m} δεσμευμένη ως προς τις ενδιάμεσες μεταβλητές της αλυσίδας, $x_{i+1}, \dots, x_{i+m-1}$. Εκτιμάται το $I_c(m)$ σε μια συμβολοσειρά από τις από κοινού και δεσμευμένες συναρτήσεις μάζας πιθανότητας, υπολογίζοντας τις αντίστοιχες σχετικές συχνότητες. Ορίζεται ο έλεγχος τυχαιοποίησης για τη σημαντικότητα του $I_c(m)$, δημιουργώντας τυχαιοποιημένες συμβολοσειρές με τυχαία μετάθεση της αρχικής συμβολοσειράς. Ο έλεγχος σημαντικότητας εφαρμόζεται για αυξανόμενες τάξεις m και η τάξη L της Μαρκοβιανής αλυσίδας εκτιμάται από το μεγαλύτερο m για το οποίο το $I_c(m)$ είναι στατιστικά σημαντικό. Αναπτύσσεται επίσης προσεγγιστικά αναλυτικό όριο σημαντικότητας με βάση τη μεροληψία της εκτίμησης της CMI και συγκρίνονται η επιλογή τάξης με τον έλεγχο τυχαιοποίησης και με αυτό το όριο σε Monte Carlo προσομοιώσεις Μαρκοβιανών αλυσίδων για διαφορετική τάξη L , διαφορετικό μήκος συμβολοσειράς N , καθώς και διαφορετικό πλήθος συμβόλων K . Τα αποτελέσματα έδειξαν ότι για ικανά μεγάλο N , ο έλεγχος τυχαιοποίησης με τη CMI εντοπίζει με ακρίβεια τη σωστή τάξη L και σημαντικά καλύτερα από το αναλυτικό όριο.

Λέξεις Κλειδιά: εκτίμηση τάξης, Μαρκοβιανές Αλυσίδες, Δεσμευμένη Αμοιβαία Πληροφορία (CMI), έλεγχος τυχαιοποίησης.

1. ΕΙΣΑΓΩΓΗ

Έστω μια συμβολοσειρά, δηλαδή μια ακολουθία συμβόλων, $\{x_t\}_{t=1}^N$, που παράγεται από μια Μαρκοβιανή αλυσίδα $\{X_t\}$, $t \geq 1$, τάξης $L \geq 1$, η οποία παίρνει τιμές από ένα διακριτό χώρο K καταστάσεων $A = \{a_1, \dots, a_K\}$. Για $n \geq L$ ισχύει

$$p(x_1, \dots, x_n) = p(x_1, \dots, x_L) \prod_{i=L+1, \dots, n} p(x_i | x_{i-L}, \dots, x_{i-1}),$$

όπου $p(\cdot)$ και $p(\cdot | \cdot)$ δηλώνουν πιθανότητα και δεσμευμένη πιθανότητα (εδώ εννοείται η πιθανότητα μετάβασης) αντίστοιχα και γίνεται η παραδοχή $p(x) = \Pr(X = x)$ για

$x \in A$. Το πρόβλημα της εύρεσης της τάξης L Μαρκοβιανής αλυσίδας ανάγεται στη χρήση των δεδομένων της συμβολοσειράς για την εκτίμηση του L .

Στη βιβλιογραφία έχουν προταθεί διάφορα κριτήρια εκτίμησης της τάξης L . Τα πιο γνωστά είναι αυτά που βασίζονται στη μέγιστη πιθανοφάνεια, όπως το κριτήριο πληροφορίας του Akaike (Akaike Information Criterion, AIC) και το Μπείσιανό κριτήριο πληροφορίας (Bayesian Information Criterion, BIC). Το AIC τείνει να υπερεκτιμά την τάξη και το BIC αδυνατεί να εκτιμήσει σωστά μεγάλες τάξεις για μικρό N (Katz, 1981; Csiszár and Shields, 2000). Ένα πιο πρόσφατο κριτήριο τέτοιου τύπου είναι αυτό του αποτελεσματικού προσδιορισμού (Efficient Determination Criterion, EDC) (Zhao et al., 2001). Οι Peres and Shields (2005) πρότειναν την ονομαζόμενη μέθοδο μέγιστης διακύμανσης που εκτιμά το L από την πιθανότητα μετάβασης σε όρους αυξανόμενου μήκους. Αυτή η μέθοδος απαιτεί τον υπολογισμό πιθανοτήτων μετάβασης ως κάποια μέγιστη τάξη L_{max} και την αναζήτηση κατάλληλου κατωφλίου για κάθε τάξη L , και για αυτό έχουν προταθεί βελτιώσεις (Daveli et al., 2006; Heath and Pati, 2007). Άλλες πιο πρόσφατες μέθοδοι είναι η μέθοδος της σχετικής εντροπίας που μετρά την απόκλιση μεταξύ πιθανοτήτων για αυξανόμενες τάξεις (Baigorri et al., 2009), καθώς και η μέθοδος των Menendez et al. (2011) που κάνει έλεγχοι λόγου πιθανοφάνειας για αυξανόμενες τάξεις με χρήση του στατιστικού της ϕ -απόκλισης (ϕ -divergence).

Σε μια διαφορετική προσέγγιση από τις παραπάνω, βασιζόμενοι στη θεωρία της πληροφορίας και ειδικότερα στην εντροπία και αμοιβαία πληροφορία, προτείνεται η εκτίμηση του L με βάση τη δεσμευμένη αμοιβαία πληροφορία (Conditional Mutual Information (CMI)). Αρχικά, στην ενότητα 2 ορίζεται η CMI, η εκτίμηση της σε συμβολοσειρά, καθώς και το αναλυτικό όριο σημαντικότητας της CMI με βάση τη μεροληψία της εκτίμησης. Για την εκτίμηση του L προτείνεται ο έλεγχος τυχαιοποίησης για αυξανόμενες τάξεις. Στην ενότητα 3 παρουσιάζεται η απόδοση του ελέγχου τυχαιοποίησης με το στατιστικό CMI σε προσομοιώσεις με Μαρκοβιανές αλυσίδες διαφόρων τάξεων και αριθμού διακριτών καταστάσεων, καθώς και για διαφορετικά μήκη συμβολοσειρών. Στην ενότητα 4 παρουσιάζονται τα συμπεράσματα αυτής της μελέτης.

2. ΕΚΤΙΜΗΣΗ ΤΑΞΗΣ ΜΑΡΚΟΒΙΑΝΗΣ ΑΛΥΣΙΔΑΣ

Για την τυχαία μεταβλητή X_t σε χρόνο t μιας Μαρκοβιανής αλυσίδας θεωρούμε την εντροπία του Shannon (Shannon entropy) που εκφράζει την αβεβαιότητα της X_t

$$H(X_t) = - \sum_{x_t} p(x_t) \ln p(x_t),$$

όπου το άθροισμα ορίζεται για όλα τα δυνατά σύμβολα (διακριτές τιμές) $x_t \in A$. Κατ' επέκταση η εντροπία Shannon μιας λέξης μήκους m , $\mathbf{X}_t = \{X_t, X_{t-1}, \dots, X_{t-m+1}\}$, είναι

$$H(\mathbf{X}_t) = - \sum_{x_t, \dots, x_{t-m+1}} p(\mathbf{x}_t) \ln p(\mathbf{x}_t),$$

όπου $\mathbf{x}_t = \{x_t, x_{t-1}, \dots, x_{t-m+1}\} \in A^m$.

2.1 Δεσμευμένη αμοιβαία πληροφορία

Η αμοιβαία πληροφορία $I(m) = I(X_t, X_{t-m})$ δύο τυχαίων μεταβλητών που απέχουν m θέσεις στην αλυσίδα, ορίζεται μέσω της εντροπίας ως (Cover and Thomas, 2006)

$$I(m) = H(X_t) + H(X_{t-m}) - H(X_t, X_{t-m}) = \sum_{x_t, x_{t-m}} p(x_t, x_{t-m}) \ln \frac{p(x_t, x_{t-m})}{p(x_t)p(x_{t-m})}.$$

Αν θέλουμε να εκφράσουμε την πληροφορία για τη X_t που λαμβάνουμε γνωρίζοντας τη X_{t-m} (και αντίστροφα), δηλαδή να μην υπάρχει πληροφορία στις ενδιάμεσες μεταβλητές $X_{t-1}, \dots, X_{t-m+1}$, θα πρέπει να θεωρήσουμε τη δεσμευμένη αμοιβαία πληροφορία (Conditional Mutual Information, CMI), που ορίζεται ως

$$\begin{aligned} I_c(m) &= I(X_t, X_{t-m} | X_{t-1}, \dots, X_{t-m+1}) \\ &= -H(X_t, \dots, X_{t-m}) + H(X_{t-1}, \dots, X_{t-m}) \\ &\quad + H(X_t, \dots, X_{t-m+1}) - H(X_{t-1}, \dots, X_{t-m+1}) \\ &= \sum_{x_t, \dots, x_{t-m}} p(x_t, \dots, x_{t-m}) \ln \frac{p(x_t | x_{t-1}, \dots, x_{t-m})}{p(x_t | x_{t-1}, \dots, x_{t-m+1})}. \end{aligned} \quad (1)$$

Για διαδοχικές τυχαίες μεταβλητές στην αλυσίδα, δηλαδή για $m = 1$, η CMI συμπίπτει με την αμοιβαία πληροφορία, $I_c(1) = I(1)$.

2.2 Εκτίμηση δεσμευμένης αμοιβαίας πληροφορίας

Η εκτίμηση της CMI δίνεται μέσω της εκτίμησης της από κοινού πιθανότητας και των δεσμευμένων πιθανοτήτων στην (1) από τις αντίστοιχες σχετικές συχνότητες. Ειδικότερα, για την πιθανότητα εμφάνισης λέξης $\{i_1, \dots, i_m\} \in A^m$ στη συμβολοσειρά $\{x_t\}_{t=1}^N$, ο εκτιμητής μέγιστης πιθανοφάνειας είναι $\hat{p}(x_t, x_{t-1}, \dots, x_{t-m+1}) = n_{i_1, \dots, i_m} / K^m$, όπου $n_{i_1, \dots, i_m} = \sum_{t=m}^N \mathbb{I}(x_t = i_1, \dots, x_{t-m+1} = i_m)$, και I η δείκτρια συνάρτηση. Αντίστοιχα ο εκτιμητής μέγιστης πιθανοφάνειας για τη δεσμευμένη πιθανότητα $p(x_t | x_{t-1}, \dots, x_{t-m})$ είναι

$$\hat{p}(x_t | x_{t-1}, \dots, x_{t-m}) = \frac{n_{i_1, \dots, i_m, i_{m+1}}}{n_{i_1, \dots, i_m}}.$$

Για μια συμβολοσειρά δεδομένου μήκους N και πλήθους K συμβόλων, η εκτίμηση του CMI γίνεται πιο δύσκολη για μεγάλη τάξη m , καθώς θα πρέπει να εκτιμηθούν οι πιθανότητες σε δειγματοχώρους μεγαλύτερης διάστασης. Επίσης η εκτίμηση του CMI περιορίζεται με τη μείωση του N καθώς και με την αύξηση του K .

Είναι γνωστό πως η εκτίμηση της εντροπίας παρουσιάζει αρνητική μεροληψία, δηλαδή η εκτιμώμενη τιμή είναι μικρότερη της πραγματικής. Αυτό έχει ως συνέπεια

η αμοιβαία πληροφορία να παρουσιάζει θετική μεροληψία, που μεγαλώνει με το N , K και m . Η δεσμευμένη αμοιβαία πληροφορία παρουσιάζει επίσης θετική μεροληψία, αλλά πολύ μικρότερη από αυτήν της αμοιβαίας πληροφορίας (π.χ. βλέπε (Vlachos and Kugiumtzis, 2010)). Βασιζόμενοι στο προσεγγιστικό αποτέλεσμα του Roulston (1999) για τη μεροληψία της εντροπίας, και μετά από υπολογισμούς καταλήγουμε στην παρακάτω προσέγγιση για τη μεροληψία της CMI

$$\hat{I}_c(m) - I_c(m) = \begin{cases} \frac{(K-1)^2}{2N} & \text{αν } m = 1 \\ \frac{K^{m-1}(K-1)^2}{2N} & \text{αν } m > 1, \end{cases} \quad (2)$$

όπου $I_c(m)$ είναι η πραγματική τιμή της CMI και $\hat{I}_c(m)$ η εκτίμηση της. Παρατηρούμε ότι η υπερεκτίμηση της CMI αυξάνει με την αύξηση οποιουδήποτε από τα N , K και m .

2.3 Έλεγχος τυχαιοποίησης για τη σημαντικότητα της CMI

Χρησιμοποιείται η CMI για να διερευνηθεί η τάξη L της Μαρκοβιανής αλυσίδας. Για κάποια τάξη $m \leq L$ θα πρέπει να ισχύει $I_c(m) > 0$, δηλαδή να υπάρχει εξάρτηση μεταξύ δύο συμβόλων που απέχουν m σύμβολα, δεδομένων των $m - 1$ ενδιάμεσων συμβόλων, ενώ για $m > L$ θα είναι $I_c(m) = 0$. Άρα αυξάνοντας την τάξη m , μπορούμε να εντοπίσουμε την τάξη L από το m για το οποίο ισχύει $I_c(m) > 0$ και $I_c(m + 1) = 0$. Λόγω της μεροληψίας, για την εκτίμηση του $I_c(m)$ έχουμε πάντα $\hat{I}_c(m) > 0$. Για αυτό καταφεύγουμε σε έλεγχο σημαντικότητας για το $I_c(m)$ με στατιστικό το $\hat{I}_c(m)$ (για κάθε m που δοκιμάζουμε) και επειδή η κατανομή του $\hat{I}_c(m)$ κάτω από τη μηδενική υπόθεση, $H_0: I_c(m) = 0$, δεν είναι γνωστή κάνουμε έλεγχο τυχαιοποίησης, που συνίσταται στα παρακάτω βήματα:

1. Δημιουργούμε M τυχαιοποιημένες συμβολοσειρές $\{x_t^{*1}\}_{t=1}^N, \dots, \{x_t^{*M}\}_{t=1}^N$, όπου η κάθε μια προκύπτει από τυχαία αντιμετάθεση των συμβόλων της αρχικής συμβολοσειράς $\{x_t\}_{t=1}^N$.
2. Υπολογίζουμε τη δεσμευμένη αμοιβαία πληροφορία $\hat{I}_c(m)$ της πραγματικής αλλά και των τυχαιοποιημένων συμβολοσειρών και τις συμβολίζουμε

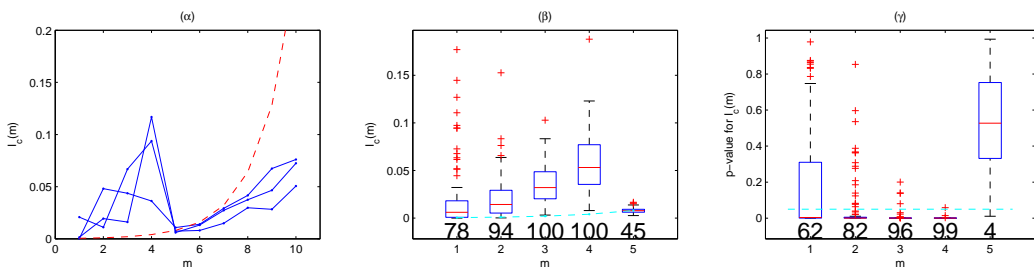
$$\hat{I}_c^0(m), \hat{I}_c^{*1}(m), \dots, \hat{I}_c^{*M}(m).$$

3. Επειδή η εναλλακτική υπόθεση είναι $H_1: I_c(m) > 0$, απορρίπτεται η H_0 αν η $\hat{I}_c^0(m)$ βρίσκεται στη δεξιά ουρά της εμπειρικής μηδενικής κατανομής που σχηματίζεται από τις τιμές $\hat{I}_c^{*1}(m), \dots, \hat{I}_c^{*M}(m)$. Για το λόγο αυτό χρησιμοποιείται η τάξη r^0 , δηλαδή η θέση της $\hat{I}_c^0(m)$ στην κατάταξη κατά αύξουσα σειρά των $M + 1$ τιμών. Η p -τιμή του δεξιά μονόπλευρου ελέγχου είναι $1 - (r^0 - 0.326)/(M + 1 + 0.348)$ χρησιμοποιώντας τη διόρθωση που προτείνεται στην εργασία των Yu and Huang (2001).

3. MONTE CARLO ΠΡΟΣΟΜΟΙΩΣΕΙΣ

Για την αξιολόγηση του κριτηρίου $I_c(m) > 0$ και $I_c(m+1) = 0$ για την εκτίμηση της τάξης L Μαρκοβιανής αλυσίδας, χρησιμοποιώντας την αναλυτική προσέγγιση της (2) και τον έλεγχο τυχαιοποίησης, κάνουμε προσομοιώσεις Μαρκοβιανών αλυσίδων για διάφορες τιμές των παραμέτρων L , K (αριθμός συμβόλων) και N (μήκος της αλυσίδας). Για κάθε περίπτωση γίνονται 100 πραγματοποιήσεις συμβολοσειρών και για κάθε πραγματοποίηση δημιουργούνται $M = 1000$ τυχαιοποιημένες συμβολοσειρές για τον έλεγχο σημαντικότητας του $I_c(m)$, όπου $m = 1, \dots, L+1$. Οι πραγματοποιήσεις γίνονται για τυχαίους πίνακες μετάβασης που ορίζουν τη Μαρκοβιανή αλυσίδα.

Μια πρώτη παρατήρηση είναι πως η μεροληψία στην εκτίμηση του $I_c(m)$ αυξάνει με το m , όπως φαίνεται στο Σχήμα 1α για τρεις πραγματοποιήσεις ($K = 2$, $L = 4$, $N = 1000$). Παρατηρούμε αύξηση του $\hat{I}_c(m)$ για $m = 1, \dots, L$, που ακολουθείται **Σχήμα 1: Monte Carlo προσομοιώσεις για Μαρκοβιανή αλυσίδα $K = 2$ συμβόλων, τάξης $L = 4$ και μήκους $N = 1000$: (α) CMI για $m = 1, \dots, 10$ από 3 τυχαίες πραγματοποιήσεις. (β) Θηκογράμματα της CMI για $m = 1, \dots, L+1$ από 100 πραγματοποιήσεις. (γ) Θηκογράμματα των p -τιμών ελέγχου τυχαιοποίησης για τη σημαντικότητα της CMI για τις ίδιες 100 πραγματοποιήσεις. Στο (α) και (β) η διακεκομμένη γραμμή παριστάνει το προσεγγιστικό όριο σημαντικότητας, ενώ στο (γ) το επίπεδο σημαντικότητας $\alpha = 0.05$. Στο (β) και (γ) οι αριθμοί κάτω από τα θηκογράμματα δηλώνουν το πλήθος των στατιστικά σημαντικών CMI όπως χαρακτηρίζονται από την αντίστοιχη ανάλυση.**

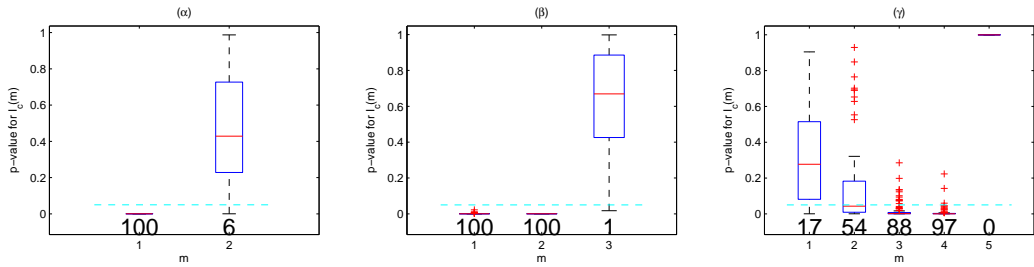


από απότομη πτώση κοντά στο 0 για $m = L+1$ και πάλι αύξηση για μεγαλύτερα m , με αποτέλεσμα για $m \leq L$ οι τιμές του $\hat{I}_c(m)$ να είναι μεγαλύτερες από τα προσεγγιστικά αναλυτικά όρια σημαντικότητας που δίνονται από την (2) για $I_c(m) = 0$, ενώ για $m > L$ να τείνουν να είναι μικρότερες. Η πρώτη αυτή ένδειξη καλής εκτίμησης της τάξης L με τη χρήση του ορίου σημαντικότητας από την (2) δεν επιβεβαιώνεται όταν συγκεντρώνουμε τα αποτελέσματα από 100 πραγματοποιήσεις. Όπως δείχνουν τα θηκογράμματα στο Σχήμα 1β, ενώ για $m = L$ το $\hat{I}_c(m)$ είναι πάνω από το όριο σημαντικότητας της (2) για όλες τις πραγματοποιήσεις, για $m = L+1$ μόνο σε 55 πραγματοποιήσεις το $\hat{I}_c(m)$ είναι κάτω από το αντίστοιχο όριο σημαντικότητας της (2). Με τον έλεγχο τυχαιοποίησης τα αποτελέσματα είναι πολύ καλύτερα και όπως δίνεται με τα θηκογράμματα των p -τιμών του ελέγχου στο Σχήμα 1γ, σε επίπεδο σημαντικότητας $\alpha = 0.05$, σε 99 από τις 100 πραγματοποιήσεις απορρίπτουμε την $H_0: I_c(m) = 0$ για $m = L$ και μόνο σε 4 πραγματοποιήσεις για $m = L+1$.

Στη συνέχεια επικεντρωθήκαμε στον έλεγχο τυχαιοποίησης για τη σημαντικότητα του $I_c(m)$ για $m \leq L$ και $m > L$. Πρώτα παρατηρούμε πως για τάξη L σχετικά μεγάλη, ο έλεγχος για $m = 1$ έχει μικρή ισχύ που βελτιώνεται καθώς το m αυξάνει και πλησιάζει το L (π.χ. δεξ Σχήμα 1γ). Αυτό είναι γενικό γνώρισμα ανεξάρτητα από τις παραμέτρους K και N και πιθανώς οφείλεται στο ότι οι Μαρκοβιανές αλυσίδες δημιουργούνται από τυχαίους πίνακες μετάβασης σε κάθε πραγματοποίηση. Για παράδειγμα για $L = 4$ η επίδραση του x_{t-1} στο x_t περιορίζεται λόγω της επίδρασης που έχουν επίσης οι x_{t-2} , x_{t-3} , x_{t-4} στο x_t και επειδή οι επιδράσεις ορίζονται τυχαία, μπορεί σε κάποιες πραγματοποιήσεις το $I_c(1)$ που μετρά την επίδραση του x_{t-1} στο x_t να μην είναι στατιστικά σημαντικό. Αυτό το φαινόμενο τείνει να εξαφανίζεται για μεγαλύτερα m έτσι ώστε για $m = L$ το $I_c(m)$ να είναι στατιστικά σημαντικό και να επιτρέπει τη σωστή εκτίμηση του L , αφού για $m = L + 1$ το $I_c(m)$ δεν είναι στατιστικά σημαντικό.

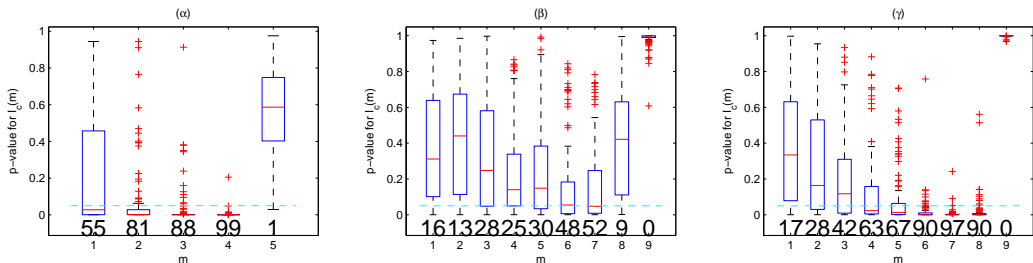
Η εκτίμηση του L γίνεται με την ίδια επιτυχία για μικρή αύξηση των συμβόλων K . Όπως δίνεται στο Σχήμα 2 για $K = 4$ και $N = 1000$, εκτιμάται σωστά η τάξη $L = 1$ σε 94% των πραγματοποιήσεων, η τάξη $L = 2$ σε 99% των πραγματοποιήσεων και η τάξη $L = 4$ σε 97% των πραγματοποιήσεων.

Σχήμα 2: Τα διαγράμματα (α), (β) και (γ) είναι όπως στο Σχήμα 1γ αλλά για $K = 4$, $N = 1000$ και για $L = 1$ στο (α), $L = 2$ στο (β) και $L = 4$ στο (γ).



Οι προσομοιώσεις σε μικρές συμβολοσειρές έδειξαν ότι ο έλεγχος τυχαιοποίησης για τη σημαντικότητα του $I_c(m)$ αποδίδει καλά όταν το K και L είναι μικρά. Για παράδειγμα για $K = 2$ και $N = 500$, εκτιμάται η σωστή τάξη $L = 4$ σε ποσοστό 98% των πραγματοποιήσεων, όπως δίνεται στο Σχήμα 3α. Όταν όμως για το ίδιο

Σχήμα 3: Τα διαγράμματα (α), (β) και (γ) είναι όπως στο Σχήμα 1γ ($K = 2$) και οι άλλες παράμετροι είναι: (α) $L = 4$, $N = 500$, (β) $L = 8$, $N = 500$ και (γ) $L = 8$, $N = 1000$.



μικρό N η τάξη L αυξάνει, αυξάνει και η διασπορά του εκτιμητή της δεσμευμένης και από κοινού πιθανότητας, ιδιαίτερα για διάσταση m κοντά στο L , με αποτέλεσμα η εκτίμηση του $I_c(m)$ για μεγάλα m κοντά στο L να είναι προβληματική. Σε αντίθεση με τη σωστή εκτίμηση του $L = 4$ (Σχήμα 3α), η εκτίμηση του $L = 8$ αποτυγχάνει, όπως δίνεται στο Σχήμα 3β. Ενώ η ισχύς του ελέγχου σημαντικότητας τείνει να αυξάνει με το m για $m \leq 7$, δε φτάνει σε υψηλά επίπεδα και μάλιστα για $m = 8$ μειώνεται με αποτέλεσμα μόνο σε ποσοστό 9% των πραγματοποιήσεων να εντοπίζεται η αλλαγή από στατιστικό σημαντικό $I_c(m)$ για $m = L$ σε μη-σημαντικό για $m = L + 1$. Η κατάσταση βελτιώνεται αν αυξήσουμε το μήκος της συμβολοσειράς και, όπως δίνεται στο Σχήμα 3γ, διπλασιάζοντας το N πετυχαίνουμε σωστή εκτίμηση του $L = 4$ στο 90% των πραγματοποιήσεων.

4. ΣΥΜΠΕΡΑΣΜΑΤΑ

Η χρήση του μέτρου της δεσμευμένης αμοιβαίας πληροφορίας (CMI) που προτείνεται σε αυτήν την εργασία για την εκτίμηση της τάξης L Μαρκοβιανής αλυσίδας είναι παρόμοια με τη χρήση της μερικής αυτοσυσχέτισης για την εκτίμηση της τάξης αυτοπαλινδρομούμενου μοντέλου σε χρονοσειρές (Box et al., 1994). Ενώ για τη μερική αυτοσυσχέτιση προσδιορίζονται παραμετρικά τα όρια σημαντικότητας (με κάποιες παραδοχές), για τη CMI μόνο προσεγγιστικά, ορίστηκαν όρια σημαντικότητας. Οι προσομοιώσεις έδειξαν ότι τα όρια αυτά δεν είναι αρκετά ακριβή για να μας οδηγήσουν σε ασφαλή εκτίμηση του L . Έτσι προτείνεται έλεγχος τυχαιοποίησης για τη σημαντικότητα της CMI και ορίζεται η εκτίμηση του L από τη μεγαλύτερη τάξη m για την οποία η CMI είναι στατιστικά σημαντική. Το κριτήριο εξετάστηκε με Monte Carlo προσομοιώσεις ως προς την τάξη L , το πλήθος συμβόλων και το μήκος της συμβολοσειράς. Τα αποτελέσματα έδειξαν πως το προτεινόμενο κριτήριο αποδίδει καλά ακόμη και για μικρά μήκη συμβολοσειρών. Ωστόσο, για μεγάλες τάξεις ή πλήθος συμβόλων της αλυσίδας, για την εκτίμηση απαιτείται αντίστοιχα μεγάλο μήκος της συμβολοσειράς. Στη συνέχεια, θα πρέπει να συγκριθεί η απόδοση του προτεινόμενου κριτηρίου με άλλα κριτήρια υπάρχοντα στη βιβλιογραφία.

Ένα πλεονέκτημα του προτεινόμενου κριτηρίου είναι ότι ο έλεγχος σημαντικότητας εφαρμόζεται ανεξάρτητα για κάθε τάξη m και δεν εμπλέκει κάποια μέγιστη τάξη, όπως για παράδειγμα τα κριτήρια που βασίζονται σε ελέγχους κλάσματος πιθανοφάνειας. Επίσης μπορεί να εφαρμοσθεί επαναληπτικά, αυξάνοντας κάθε φορά κατά ένα την τάξη και κάνοντας έλεγχο σημαντικότητας και τερματίζοντας όταν δεν απορρίπτεται η μηδενική υπόθεση.

ABSTRACT

The present study introduces the conditional mutual information (CMI) for the estimation of the order L of a Markov chain $\{x_i\}$, $i = 1, \dots, N$ of K symbols. We define CMI of order m , $I_c(m)$, as the mutual information of x_i and x_{i+m}

conditioning on the intermediate variables of the chain, $x_{i+1}, \dots, x_{i+m-1}$. The estimate of $I_c(m)$ is derived by estimating joint and conditioned mass probability functions with the corresponding relative frequencies. We define a randomization significance test for $I_c(m)$, where the randomized symbol sequences are formed by random permutation of the components of the original symbol sequence. The randomization test is applied for increasing orders m , and L is estimated by the largest m for which $I_c(m)$ is statistically significant. We also develop approximate analytic significance limits based on the estimation bias of CMI, and we compare the order selection using these limits and the randomization test in Monte Carlo simulations of Markov chains for varying L , N and K . The results showed that for sufficiently large N , the randomization test with CMI detects accurately the correct order L and significantly better than the approximate significance limits.

ΑΝΑΦΟΡΕΣ

- Baigorri, A.R., Gonçalves, C.R. and Resende, P.A.A. (2009). Markov chain order estimation and relative entropy. *arXiv:math/0506080v1* .
- Box, G. E. P., Jenkins, G. M. and Reinsel, G. C. (1994). *Time Series Analysis: Forecasting and Control*, third edn, Prentice-Hall, New Jersey.
- Cover, T. and Thomas, J. (2006). *Elements of Information Theory*, 2nd edn, Wiley.
- Csiszár, I. and Shields, P.C. (2000). The consistency of the BIC Markov order estimator. *The Annals Statistics*, **28(6)**, 1601–1619.
- Daveli, D., Dubhashi, D. and Hermansson, M. (2006). A new order estimator for fixed and variable length Markov models with applications to DNA sequence similarity. *Statistical Applications in Genetics and Molecular Biology*, **5(8)**.
- Heath, L.S. and Pati, A. (2007). Predicting Markov chain order in genomic sequences. *IEEE International Conference on Bioinformatics and Biomedicine, Proceedings*, 159–164.
- Katz, R.W. (1981). On some criteria for estimating the order of a Markov chain. *Technometrics*, **23(3)**, 243–249.
- Menendez, M.L., Pardo, L., Pardo, M.C. and Zografos, K. (2011). Testing the order of Markov dependence in DNA sequences. *Methodol. Comput. Appl. Probab.*, **13**, 59–74.
- Peres, Y. and Shields, P. (2005). Two new order estimators. *arXiv:math/0506080v1* .
- Roulston, M.S. (1999). Estimating the errors on Measured Entropy and Mutual Information. *Physica D*, **125**, 285–294.
- Vlachos, I. and Kugiumtzis, D. (2010). Non-uniform state space reconstruction and coupling detection. *Physical Review E*, **82**, 016207.

- Yu, G.-H. and Huang, C.-C. (2001). A distribution free plotting position. *Stochastic Environmental Research And Risk Assessment*, **15(6)**, 462–476.
- Zhao, L.C., Dorea, C.C.Y. and Gonçalves, C.R. (2001). On determination of the order of a Markov chain. *Statistical Inference of Stochastic Processes*, **4**, 273–282.



ΜΙΑ ΟΙΚΟΓΕΝΕΙΑ ΚΑΤΑΝΟΜΩΝ ΣΤΗΝ ΑΝΑΛΥΣΗ ΕΠΙΒΙΩΣΗΣ

B. Α. Παππάς¹, Κ. Αδαμίδης², Σ. Λουκάς¹,

¹Τμήμα Μαθηματικών, Πανεπιστήμιο Ιωαννίνων

vapappas@cc.uoi.gr, sloukas@uoi.gr

²Τμήμα Διοίκησης Επιχειρήσεων Αγροτικών Προϊόντων και Τροφίμων, Πανεπιστήμιο Ιωαννίνων

cadamid@cc.uoi.gr

ΠΕΡΙΛΗΨΗ

Μια τετρα-παραμετρική τροποποιημένη γενίκευση της κατανομής Weibull εισάγεται, ως μια ειδική περίπτωση μιας γενικευμένης οικογένειας κατανομών που δημιουργήθηκε σε παραλληλισμό με τους Marshall και Olkin (1997). Μελετώνται διάφορες ιδιότητες της κατανομής και με χρήση των εκτιμητών μέγιστης πιθανοφάνειας ελέγχεται η προσαρμογή της σε πραγματικά δεδομένα.

Λέξεις Κλειδιά: Κυπελλοειδής καμπύλη, Επεκταμένη κατανομή Weibull, Συνάρτηση κινδύνου, Λογαριθμική κατανομή, Εκτίμηση μέγιστης πιθανοφάνειας, Τροποποιημένη κυρτή καμπύλη, Συνάρτηση επιβίωσης, Εφαρμογές.

1. ΕΙΣΑΓΩΓΗ

Οι κατανομές πιθανότητας χρησιμοποιούνται συχνά στην Ανάλυση Επιβίωσης για τη μοντελοποίηση δεδομένων, επειδή προσφέρουν εικόνα για τη φύση των διαφόρων παραμέτρων και συναρτήσεων, ιδιαίτερα της συνάρτησης κινδύνου. Κατά τη διάρκεια των τελευταίων δεκαετιών, ένα σημαντικό μέρος της έρευνας είναι αφιερωμένο στη δημιουργία κατανομών πιθανότητας με διαφόρων μορφών συνάρτηση κινδύνου. Προφανώς, το κίνητρο για την τάση αυτή, είναι να δοθεί μεγαλύτερη ελευθερία επιλογής και βελτιστοποίηση της περιγραφής δεδομένων που συχνά απορρέουν από σύνθετες πραγματικές καταστάσεις. Ενδεικτικά αναφέρονται οι εργασίες των: Adamidis και Loukas (1998), Adamidis et al. (2005), Dimitrakopoulou et al. (2007), Kus (2007), Barreto-Souza et al. (2008), Marshall και Olkin (2007), Tahmasbi και Rezaei (2008), Barreto-Souza και Cribari-Neto (2009), Chahkandi και Ganjali (2009) και οι αναφορές σε αυτές. Σε αυτή την εργασία εισάγεται μια γενικευμένη οικογένεια κατανομών, με την προσθήκη μιας επιπλέον παραμέτρου σε μια γενική κατανομή πιθανότητας, σε παραλληλισμό με τους Marshall και Olkin (1997). Εν συνεχεία, η οικογένεια κατανομών χρησιμοποιείται για την ανάπτυξη μιας τετρα-παραμετρικής τροποποιημένης επέκτασης της κατανομής Weibull, με συνάρτηση κινδύνου διαφόρων μορφών, που συναγωνίζεται επαρκώς άλλες κατανομές στην προσαρμογή σε πραγματικά δεδομένα.

Η εργασία είναι οργανωμένη ως εξής: Η ενότητα 2 περιλαμβάνει τη γενικευμένη οικογένεια κατανομών και ορισμένες ιδιότητες αυτής. Η προτεινόμενη τετρα-παραμετρική κατανομή εισάγεται στην ενότητα 3, μαζί με τη μελέτη των ιδιοτήτων της και θεμάτων αξιοπιστίας. Οι παράμετροι της κατανομής εκτιμώνται με τη μέθοδο μέγιστης πιθανοφάνειας και υπολογίζεται ο παρατηρούμενος πίνακας πληροφορίας. Τέλος ελέγχεται η προσαρμογή της προτεινόμενης κατανομής έναντι άλλης εναλλακτικής κατανομής, σε δύο πραγματικά σύνολα δεδομένων.

2. Η ΓΕΝΙΚΕΥΜΕΝΗ ΟΙΚΟΓΕΝΕΙΑ ΚΑΤΑΝΟΜΩΝ

Η γενίκευση μιας κατανομής πιθανότητας μπορεί να επιτευχθεί με την προσθήκη μιας παραμέτρου μορφής, σε παραλληλισμό με τους Mashall και Olkin (1997). Έτσι, έχοντας μια κατανομή με συνάρτηση επιβίωσης s_0 , η συνάρτηση επιβίωσης της προτεινόμενης οικογένειας κατανομών με την επιπρόσθετη παράμετρο p θα δίνεται από τον τύπο:

$$s(x) = \frac{\ln \{1 - (1-p)s_0(x)\}}{\ln p}, x, p \in \mathbb{R}_+ \quad (1)$$

όπου όταν $p \rightarrow 1$, τότε $s \rightarrow s_0$. Οι συναρτήσεις πυκνότητας πιθανότητας και κινδύνου υπολογίζονται άμεσα και δίνονται από τους τύπους:

$$\begin{aligned} f(x) &= \frac{(p-1)f_0(x)}{\{1 - (1-p)s_0(x)\} \ln p}, \\ h(x) &= \frac{(p-1)s_0(x)h_0(x)}{\{1 - (1-p)s_0(x)\} \ln \{1 - (1-p)s_0(x)\}}, \end{aligned} \quad (2)$$

όπου f_0 και h_0 οι συναρτήσεις πυκνότητας πιθανότητας και κινδύνου που αντιστοιχούν στην αρχική κατανομή με συνάρτηση επιβίωσης s_0 , και $x, p \in \mathbb{R}_+$. Εφόσον,

$$\lim_{x \rightarrow -\infty} h(x) = \frac{p-1}{p \ln p} \lim_{x \rightarrow -\infty} h_0(x) \text{ και } \lim_{x \rightarrow +\infty} h(x) = \lim_{x \rightarrow +\infty} h_0(x)$$

από την (2) έπεται ότι,

$$\frac{p-1}{p \ln p} h_0(x) \leq h(x) \leq h_0(x), x \in \mathbb{R}_+, p \geq 1,$$

$$h_0(x) \leq h(x) \leq \frac{p-1}{p \ln p} h_0(x), x \in \mathbb{R}_+, p \in (0, 1].$$

Επομένως, η ποσότητα $h(x)/h_0(x)$ με $x \in \mathbb{R}_+$ είναι αύξουσα για $p \geq 1$ και φθίνουσα για $p \in (0, 1]$. Όταν $s_0(0) = 1$, η συνάρτηση κινδύνου στη αρχική της τιμή, $h(0)$, συμπεριφέρεται διαφορετικά σπό ότι οι αντίστοιχες συναρτήσεις των Weibull και Γάμμα κατανομών. Και για τις δύο αυτές οικογένειες, η κατανομή μπορεί να είναι εκθετική με $h(0) = 0$, ή $h(0) = \infty$, έτσι ώστε η $h(0)$ να είναι μη συνεχής στην παράμετρο μορφής. Αυτό δεν ισχύει στην περίπτωση συναρτήσεων κινδύνου της μορφής (2), και επομένως η προτεινόμενη οικογένεια κατανομών μπορεί να φανεί

χρήσιμη στη βελτιστοποίηση της προσαρμογής της κατανομής με συνάρτηση επιβίωσης s_0 .

3. ΜΕΛΕΤΗ ΜΙΑΣ ΕΙΔΙΚΗΣ ΠΕΡΙΠΤΩΣΗΣ

Η συνάρτηση επιβίωσης της τροποποιημένης επέκτασης της κατανομής Weibull, που προτάθηκε από τους Xie *et al.* (2002) και μελετήθηκε επιπρόσθετα από τους Tang *et al.* (2003), δίνεται από τη σχέση:

$$s_0(x) = e^{-\alpha\lambda\left(e^{\left(\frac{x}{\alpha}\right)^\beta} - 1\right)}, \quad (3)$$

για $x, \alpha, \beta, \lambda \in \mathbb{R}_+$. Εφ' εξής, χάρη συντομίας θα αναφέρουμε ως XTG κατανομή, την κατανομή με συνάρτηση επιβίωσης της μορφής (3). Αντικαθιστώντας την (3) στη σχέση (1), προκύπτει η συνάρτηση επιβίωσης της προτεινόμενης τετρα-παραμετρικής κατανομής:

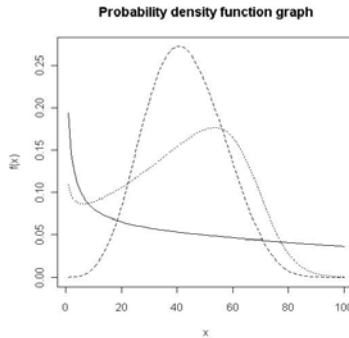
$$s(x; \theta) = \frac{\ln\{1 - (1-p)e^{-\alpha\lambda\left(e^{\left(\frac{x}{\alpha}\right)^\beta} - 1\right)}\}}{\ln p}, \quad (4)$$

όπου $\theta = (\alpha, \beta, \lambda, p)$ και $x, \alpha, \beta, \lambda, p \in \mathbb{R}_+$. Από τη (2) ή την (4), μπορεί άμεσα να υπολογιστεί η σ.π.π., η οποία δίνεται από τον τύπο:

$$f(x; \theta) = \frac{\beta\lambda(p-1)\left(\frac{x}{\alpha}\right)^{\beta-1} e^{\left(\frac{x}{\alpha}\right)^\beta - \alpha\lambda\left(e^{\left(\frac{x}{\alpha}\right)^\beta} - 1\right)}}{\{1 - (1-p)e^{-\alpha\lambda\left(e^{\left(\frac{x}{\alpha}\right)^\beta} - 1\right)}\} \ln p}, \quad (5)$$

όπου α είναι παράμετρος κλίμακας και β, λ, p είναι παράμετροι μορφής. Αποδεικνύεται ότι η σ.π.π. μπορεί να λάβει μονότονα φθίνουσα, μονοκόρυφη ή ακόμη και τροποποιημένη κυρτή μορφή (roller-coster type). Οι διάφορες μορφές της σ.π.π. φαίνονται στο παρακάτω Γράφημα 1, που σχεδιάστηκε για συγκεκριμένες τιμές των παραμέτρων $(\alpha, \beta, \lambda, p)$.

Γράφημα 1. Συναρτήσεις πυκνότητας πιθανότητας της EXTG κατανομής για $(\alpha, \beta, \lambda, p) = (2, 0.5, 0.03, 0.2)$ (συνεχής γραμμή), $(\alpha, \beta, \lambda, p) = (12, 4, 2, 0.2)$ (διακεκομμένη γραμμή) και $(\alpha, \beta, \lambda, p) = (2, 0.7, 0.3, 0.2)$ (σημειωμένη με τελείες γραμμή).



Γίνεται εύκολα αντιληπτό ότι για $p \rightarrow 1$ η προτεινόμενη κατανομή υποβιβάζεται στην XTG κατανομή, επομένως η προτεινόμενη κατανομή μπορεί να θεωρηθεί ως μια επέκταση της XTG (η οποία αποτελεί ασυμπτωτική επέκταση της συνήθους δι-παραμετρικής Weibull κατανομής) και εάν, επιπλέον, $\alpha = 1$ τότε η (5) ορίζει την κατανομή του Chen (2000). Χάρη συντομίας, εφ' εξής θα αναφέρουμε την προτεινόμενη επεκταμένη XTG κατανομή ως EXTG. Επιπρόσθετα, μπορεί να αποδειχθεί ότι για $p \in (0,1)$ η (5) αποτελεί μίξη λογαριθμικής και XTG κατανομής. Για του λόγου το αληθές, συνδυάζοντας τα αποτελέσματα των Barlow και Proschan (1981) και Arnold *et al.* (1992), θεωρούμε τους χρόνους ζωής $X = \min(X_1, X_2, \dots, X_Z)$ ενός συστήματος Z όμοιων εξαρτημάτων σε σειρά, όπου η αποτυχία συμβαίνει εάν τουλάχιστον ένα από τα εν σειρά εξαρτήματα πάψει να λειτουργεί. Υποθέτοντας ότι οι χρόνοι ζωής των εξαρτημάτων είναι ανεξάρτητες και ισόνομες τυχαίες μεταβλητές με συναρτήσεις επιβίωσης που δίνονται από τη σχέση (3) και η κατανομή του αριθμού τους Z είναι η λογαριθμική, ανεξάρτητη των X , με συνάρτηση πιθανότητας:

$$p(z; p) = -\frac{(1-p)^z}{z \ln p},$$

για $z \in \mathbb{N} - \{0\}$, $p \in (0,1)$, τότε η κατανομή του $X | Z$ έχει σ.π.π.:

$$f(x|z; \beta, \lambda) = \beta \lambda z \left(\frac{x}{\alpha}\right)^{\beta-1} e^{\left(\frac{x}{\alpha}\right)^\beta - \alpha \lambda \left(e^{\left(\frac{x}{\alpha}\right)^\beta} - 1\right)},$$

για $x, \alpha, \beta, \lambda \in \mathbb{R}_+$ και η κατανομή του X είναι η EXTG κατανομή με σ.π.π. που δίνεται από τη σχέση (5).

Ο υπολογισμός των r -τάξεως ροπών απαιτεί χρήση τυπικών διαδικασιών αριθμητικής ολοκλήρωσης, που είναι διαθέσιμες στην πλειοψηφία των μαθηματικών υπολογιστικών πακέτων. Για $p \in (0,1)$ ωστόσο, μπορούν να εκφραστούν από τη σχέση:

$$E(X^r) = -\frac{1}{\ln p} \sum_{k=1}^{\infty} \frac{(1-p)^k}{k} \int_0^{\infty} \alpha^r \left\{ \ln\left(\frac{x}{\alpha k \lambda} + 1\right) \right\}^{r/\beta} e^{-x} dx.$$

Επιπλέον, γνωρίζοντας ότι $F(x; \theta) = 1 - S(x)$, με αναστροφή της σχέσης (4) προκύπτει η συνάρτηση υπολογισμού των εκατοστιαίων σημείων, η οποία δίνεται από τη σχέση:

$$F^{-1}(q) = \alpha \left[\ln \left\{ 1 - \frac{1}{\alpha \lambda} \ln \left(\frac{1-p^{1-q}}{1-p} \right) \right\} \right]^{1/\beta},$$

για $p \in (0, 1)$. Επομένως, η διάμεσος θα είναι $M = \alpha \left[\ln \left\{ 1 + \frac{1}{\alpha \lambda} \ln(1 + \sqrt{p}) \right\} \right]^{1/\beta}$.

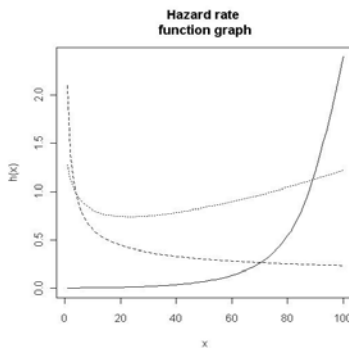
3.1 Συναρτήσεις κινδύνου και μέσου υπολειπόμενου χρόνου ζωής

Από τις σχέσεις (4) και (5) προκύπτει η συνάρτηση κινδύνου της EXTG κατανομής:

$$h(x; \theta) = \frac{\beta \lambda (p-1) \left(\frac{x}{\alpha} \right)^{\beta-1} e^{\left(\frac{x}{\alpha} \right)^{\beta} - \alpha \lambda \left(e^{\left(\frac{x}{\alpha} \right)^{\beta} - 1} \right)}}{\left\{ 1 - (1-p) e^{-\alpha \lambda \left(e^{\left(\frac{x}{\alpha} \right)^{\beta} - 1} \right)} \right\} \ln \left\{ 1 - (1-p) e^{-\alpha \lambda \left(e^{\left(\frac{x}{\alpha} \right)^{\beta} - 1} \right)} \right\}}.$$

Μπορεί να αποδειχθεί ότι για $\beta \geq 1$, η συνάρτηση κινδύνου της κατανομής EXTG είναι γνησίως αύξουσα (IFR) (Glaser, 1980). Σε διαφορετική περίπτωση, δηλαδή για $\beta < 1$, η συνάρτηση κινδύνου μπορεί να είναι γνησίως αύξουσα, φθίνουσα, ή ακόμη να λάβει και κυπελλοειδή μορφή. Οι διάφορες μορφές της συνάρτησης κινδύνου της EXTG κατανομής φαίνονται στο παρακάτω Γράφημα 2, το οποίο σχεδιάστηκε για συγκεκριμένες τιμές των παραμέτρων $(\alpha, \beta, \lambda, p)$.

Γράφημα 2. Συναρτήσεις κινδύνου της EXTG κατανομής για $(\alpha, \beta, \lambda, p) = (0.05, 0.45, 0.002, 20)$ (συνεχής γραμμή), $(\alpha, \beta, \lambda, p) = (0.2, 0.2, 3, 2)$ (διακεκομμένη γραμμή) και $(\alpha, \beta, \lambda, p) = (10, 0.9, 0.5, 0.3)$ (σημειωμένη με τελείες γραμμή).



Δεδομένου ότι δεν υπήρξε αποτυχία προγενέστερη του x_0 , ο υπολοιπόμενος χρόνος ζωής ορίζεται ως το χρονικό διάστημα από το x_0 μέχρι τη στιγμή της αποτυχίας. Η συνάρτηση του μέσου υπολοιπόμενου χρόνου ζωής της κατανομής EXTG, για $p \in (0,1)$, δίνεται από τη σχέση:

$$\begin{aligned}
 m(x_0; \theta) &= E(X - x_0 | X \geq x_0) = \int_{x_0}^{\infty} \frac{(x_0 - x)f(x; \theta)}{1 - F(x_0; \theta)} dx \\
 &= - \frac{1}{\ln \left\{ 1 - (1-p)e^{-\alpha \lambda \left(e^{\left(\frac{x_0}{\alpha} \right)^\beta - 1} \right)} \right\}} \times \\
 &\quad \sum_{k=1}^{\infty} \frac{(1-p)^k}{k} \times \int_{\alpha k \lambda \left(e^{\left(\frac{x_0}{\alpha} \right)^\beta - 1} \right)}^{\infty} \left[\alpha \left\{ \ln \left(\frac{x}{\alpha k \lambda} + 1 \right) \right\}^{1/\beta} - x_0 \right] e^{-x} dx
 \end{aligned}$$

3.2 Στατιστική συμπερασματολογία

Θεωρώντας ένα τυχαίο δείγμα n παρατηρήσεων $y_{obs} = (x_i; i=1, \dots, n)$ από την (5), η ο λογάριθμος της συνάρτησης πιθανοφάνειας θα δίνεται από τη σχέση:

$$\begin{aligned}
 \ell(\theta; y_{obs}) &= n[\ln\{(p-1)\beta\lambda\} - \ln(\ln p) + \alpha\lambda + (1-\beta)\ln\alpha] + (\beta-1) \sum_{i=1}^n \ln x_i + \sum_{i=1}^n \left(\frac{x_i}{\alpha} \right)^\beta \\
 &\quad - \alpha\lambda \sum_{i=1}^n e^{\left(\frac{x_i}{\alpha} \right)^\beta} - \sum_{i=1}^n \ln \left\{ 1 - (1-p)e^{-\alpha\lambda \left(e^{\left(\frac{x_i}{\alpha} \right)^\beta - 1} \right)} \right\},
 \end{aligned}$$

Παραγωγίζοντας ως προς τις παραμέτρους $\theta = (\alpha, \beta, \lambda, p)$ προκύπτουν οι μερικές παράγωγοι $\frac{\partial \ell}{\partial \alpha}$, $\frac{\partial \ell}{\partial \beta}$, $\frac{\partial \ell}{\partial \lambda}$ και $\frac{\partial \ell}{\partial p}$ οι τύποι υπολογισμού των οποίων είναι διαθέσιμοι

στον αναγνώστη από τους συγγραφείς.

Οι εκτιμητές μέγιστης πιθανοφάνειας $\hat{\alpha}$, $\hat{\beta}$, $\hat{\lambda}$ και \hat{p} των παραμέτρων α , β , λ , p προκύπτουν από την επίλυση του συστήματος των τεσσάρων μη-γραμμικών κανονικών εξισώσεων: $\frac{\partial \ell}{\partial \alpha} = 0$, $\frac{\partial \ell}{\partial \beta} = 0$, $\frac{\partial \ell}{\partial \lambda} = 0$ και $\frac{\partial \ell}{\partial p} = 0$. Μιας και το σύστημα των εξισώσεων

δεν έχει καμία προφανή λύση, η επίλυση του μπορεί να γίνει με χρήση μιας επα-

ναληπτικής αριθμητικής μεθόδου, όπως είναι οι μέθοδοι Newton-Raphson, quasi-Newton και Nelder-Mead.

Εφόσον οι εκτιμητές μέγιστης πιθανοφάνειας των αγνώστων παραμέτρων $\theta = (\alpha, \beta, \lambda, p)$ δεν μπορούν να προκύψουν σε κλειστή μορφή, δεν είναι εύκολο να προσδιοριστεί η ακριβής κατανομή των εκτιμητών αυτών. Ωστόσο, εφαρμόζοντας προσέγγιση μεγάλου μεγέθους δείγματος, μπορεί να υποθεθεί ότι οι εκτιμητές των παραμέτρων ακολουθούν προσεγγιστικά τετραδιάστατη κανονική κατανομή με μέση τιμή $\theta = (\alpha, \beta, \lambda, p)$ και πίνακα διακυμάνσεων-συνδιακυμάνσεων τον αντίστροφο του αναμενόμενου πίνακα πληροφορίας $J(\theta) = E(I; \theta)$ (βλ. Lawless 2003), όπου $I(\theta)$ είναι ο παρατηρούμενος πίνακας πληροφορίας με στοιχεία $I_{ij} = -\frac{\partial^2 l}{\partial \theta_i \partial \theta_j}$, όπου $i, j = 1(1)4$.

Παραγωγίζοντας επιμέρους τις σχέσεις των μερικών παραγώγων μπορούν εύκολα να προκύψουν τα άνω τριγωνικά στοιχεία του, δευτέρας τάξεως, συμμετρικού, παρατηρούμενου πίνακα πληροφορίας I , τα οποία είναι διαθέσιμα στον αναγνώστη από τους συγγραφείς.

Ο παραπάνω πίνακας $I(\theta)$ αποτελεί συνεπή εκτιμητή του $J(\theta)$ και μπορεί να χρησιμοποιηθεί για την κατασκευή ασυμπτωτικών διαστημάτων εμπιστοσύνης για τις άγνωστες παραμέτρους. Ωστόσο, αν οποιαδήποτε από τις πραγματικές τιμές των παραμέτρων είναι μηδέν, τότε η ασυμπτωτική κατανομή των εκτιμητών μέγιστης πιθανοφάνειας είναι μια μίξη κατανομών (Self και Liang, 1987) και η δημιουργία ασυμπτωτικών διαστημάτων εμπιστοσύνης δυσχεραίνει αρκετά. Η περίπτωση αυτή δεν πρόκειται να αναπτυχθεί στην παρούσα εργασία.

3.3 Εφαρμογές

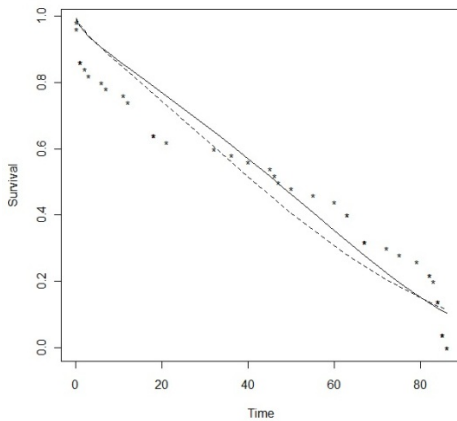
Η προσαρμογή της προτεινόμενης EXTG κατανομής ελέγχεται σε δύο σύνολα πραγματικών δεδομένων. Το πρώτο σύνολο δεδομένων περιλαμβάνει αποτελέσματα από μια δοκιμή σχετικά με τη διάρκεια ζωής 50 ηλεκτρικών εξαρτημάτων (Aarset, 1987). Το δεύτερο σύνολο δεδομένων αποτελείται από 44 παρατηρήσεις που αφορούν το χρόνο ζωής (σε ώρες) όλων των υποσυστημάτων ενός αυτοκινήτου, δηλ. κινητήρα, υδραυλικό υποσύστημα και σύστημα κλιματισμού, φρένα, σύστημα μετάδοσης, λάστιχα, τροχούς και αμάξωμα (Kumar et al., 1989 και Pulcini, 2001). Και στις δύο περιπτώσεις τα δεδομένα ομαδοποιήθηκαν και εκτιμήθηκε η εμπειρική συνάρτηση κινδύνου, επιδεικνύοντας καθαρά κυπελλοειδή μορφή. Εκτός από την προτεινόμενη EXTG κατανομή, και η XTG κατανομή προσαρμόζεται στα δύο σύνολα δεδομένων ούτως ώστε να υπάρξει και μια άμεση αντιπαράθεση των δύο κατανομών μεταξύ τους. Υπολογίζοντας τους εκτιμητές μέγιστης πιθανοφάνειας, χρησιμοποιήθηκε το τεστ καλής προσαρμογής των Kolmogorov-Smirnov (K-S), για τον έλεγχο της προσαρμοστικότητας κάθε μιας από τις δύο κατανομές στα δύο σύνολα δεδομένων. Οι εκτιμητές, ο λογάριθμος της συνάρτησης πιθανοφάνειας (LL), η τιμή του στατιστικού του K-S τεστ καθώς και η αντίστοιχη κρίσιμη τιμή (p-value) περιέχονται στον παρακάτω Πίνακα 1. Επιπρόσθετα, οι τιμές του στατιστικού τεστ πηλίκου πιθανοφάνειας για τον έλεγχο της μηδενικής υπόθεσης $H_0: p=1$, υπολογίστηκαν να είναι 8.7939 ($p=0.003$) και 6.8366 ($p=0.0089$) αντίστοιχα, επιδεικνύοντας ότι η EXTG περιγράφει

τα δύο σύνολα δεδομένων καλύτερα από την εναλλακτική XTG κατανομή. Το συμπέρασμα αυτό ενισχύεται και από τη γραφική παράσταση των εμπειρικών και προσαρμοζόμενων συναρτήσεων επιβίωσης του Γραφήματος 3.

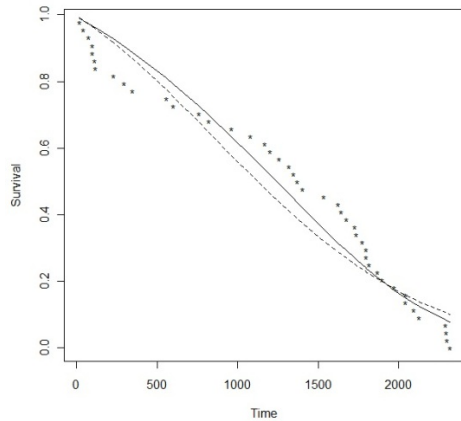
Πίνακας 1. Εκτιμητές παραμέτρων, τιμές λογαρίθμου συνάρτησης πιθανοφάνειας, στατιστικού K-S και στατιστικού λόγου πιθανοφανειών, καθώς και κρίσιμες τιμές από την προσαρμογή των συγκρινόμενων κατανομών πιθανότητας στα δύο διαθέσιμα σύνολα δεδομένων.

Σύνολο Δεδομένων	Κατανομή	$\hat{\theta} = (\hat{\alpha}, \hat{\beta}, \hat{\lambda}, \hat{p})$	LL	K-S (p-value)	LL test
1) (n=50)	EXTG	(14.06, 0.4838, 0.042, 313.57)	227.2496	0.1470 (0.23)	$H_0: p=1$
	XTG	(13.747, 0.588, 0.00876, -----)	231.6466	0.1597 (0.1562)	8.7939 (0.003)
2) (n=44)	EXTG	(0.0831, 0.2007, 0.0211, 20.4087)	696.0237	0.1397 (0.3262)	$H_0: p=1$
	XTG	(0.0806, 0.2086, 0.0057, -----)	702.8603	0.1667 (0.1544)	6.8366 (0.0089)

Γράφημα 3. Γραφικές παραστάσεις της συνάρτησης επιβίωσης της εμπειρικής κατανομής (σημεία *), της EXTG κατανομής (συνεχής γραμμή) και της XTG (διακεκομμένη γραμμή) για τα δύο σύνολα δεδομένων.



(α) Ηλεκτρικά Εξαρτήματα



(β) Υποσυστήματα αυτοκινήτου

ABSTRACT

In this paper, a four-parameter family of Weibull distributions is introduced, as an example of a more general class created along the lines of Marshall and Olkin (1997). Various properties of the distribution are explored and its usefulness in modeling real data is demonstrated using maximum likelihood estimates.

ΑΝΑΦΟΡΕΣ

- Aarset, M. V. (1987). How to identify a bathtub hazard rate, *IEEE Transactions on Reliability*, **36**, 1, 106–108.
- Adamidis, K. and Loukas, S. (1998). A lifetime distribution with decreasing failure rate, *Statistics and Probability Letters*, **39**, 35–42.
- Adamidis, K., Dimitrakopoulou, T. and Loukas S. (2005). On an extension of the exponential-geometric distribution, *Statistics and Probability Letters*, **73**, 259–269.
- Arnold, B. C., Balakrishnan, N. and Nataraja, H. N. (1992). *First Course in Order Statistics*, New York: John Wiley.
- Barlow, R. E. and Proschan, F. (1981). *Statistical Theory of Reliability and Life Testing: Probability Models, To Begin With*, Silver Spring, MD.
- Barreto-Souza, W., Lemos de Moraes, A. and Cordeiro, G. M. (2008). The Weibull-Geometric Distribution, *Preprint: arXiv:0809.2703v1*.
- Barreto-Souza, W. and Cribari-Neto, F. (2009). A generalization of the exponential-Poisson distribution, *Statistics and Probability Letters*, **79**, 2493–2500.
- Chahkandi, M. and Ganjali, M. (2009). On some lifetime distributions with decreasing failure rate, *Computational Statistics and Data Analysis*, **53**, 4433–4440.
- Chen, Z. (2000). A new two-parameter lifetime distribution with bathtub shape or increasing failure rate function, *Statistics and Probability Letters*, **49**, 155–161.
- Dimitrakopoulou, T., Adamidis, K. and Loukas, S. (2007). A lifetime distribution with an upside-down bathtub-shaped hazard function, *IEEE Transactions on Reliability*, **56**, 2, 308–311.
- Glaser, R. E. (1980). Bathtub and related failure rate characterizations, *Journal of the American Statistical Association*, **75**, 667–672.
- Kumar, U., Klefsjo, B. and Granholm, S. (1989). Reliability investigation for a fleet of load haul dump machines in Swedish mine, *Reliability Engineering and System Safety*, **26**, 4, 341–361.
- Kus, C. (2007). A new lifetime distribution, *Computational Statistics and Data Analysis*, **51**, 4497–4509.
- Lawless, J. F. (2003). *Statistical Models and Methods for Life Time Data*, 2nd edition, New York: John Wiley.
- Marshall, A. W. and Olkin, I. (1997). A new method for adding a parameter to a family of distributions with application to the Exponential and Weibull families, *Biometrika*, **84**, 641–652.
- Marshall, A. W. and Olkin, I. (2007). *Life distributions*, New York: Springer.
- Pulcini, G. (2001). Modeling the failure data of a repairable equipment with bathtub type failure intensity, *Reliability Engineering and System Safety*, **71**, 209–218.
- Self, S. G. and Liang, K.-L. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions, *Journal of the American Statistical Association*, **82**, 605–610.
- Tahmasbi, R. and Rezaei, S. (2008). A two-parameter lifetime distribution with decreasing failure rate, *Computational Statistics and Data Analysis*, **52**, 3889–3901.
- Xie, M., Tang, Y. and Goh, T. N. (2002). A modified Weibull extension with bathtub-shaped failure rate function, *Reliability Engineering and System Safety*, **76**, 279–285.
- Tang, Y., Xie, M. and Goh, T. N. (2003). Statistical analysis of a Weibull extension model, *Communications in Statistics - Theory and Methods*, **32**, 913–928.



ΒΕΛΤΙΣΤΟΙ ΣΧΕΔΙΑΣΜΟΙ ΓΡΑΜΜΗΣ ΜΕ ΤΡΕΙΣ ΑΓΩΓΕΣ ΚΑΙ ΕΞΑΡΤΗΜΕΝΕΣ ΠΑΡΑΤΗΡΗΣΕΙΣ

Κατερίνα Περικλέους¹, Στρατής Κουνιάς²

¹Εθνικό Μετσόβιο Πολυτεχνείο, katerina.pericleous@gmail.com

² Πανεπιστήμιο Αθηνών, skounias@math.uoa.gr

ΠΕΡΙΛΗΨΗ

Οι πειραματικές μονάδες (μονάδες) είναι σε διάταξη γραμμής και σε κάθε μονάδα εφαρμόζεται μια από τις τρεις αγωγές. Ο σχεδιασμός γραμμής είναι η αντιστοιχία αγωγών στις μονάδες. Εξετάζεται ο σχεδιασμός που δίνει τη βέλτιστη εκτίμηση των επιδράσεων των αγωγών. Το μοντέλο είναι κύριων επιδράσεων, ομοιογενών μονάδων, όταν οι παρατηρήσεις είναι ανεξάρτητες ή έχουν εξάρτηση αυτοπαλινδρομούμενου μοντέλου πρώτης τάξης. Δίνονται κανόνες (φίλτρα) ώστε να μειωθεί το πλήθος των υπό εξέταση σχεδιασμών. Δίνονται πίνακες με A,D,E,MV-βέλτιστους σχεδιασμούς για 4,5,6 παρατηρήσεις.

Λέξεις κλειδιά: Καθολικά βέλτιστος, φ -βέλτιστος, A,D,E,MV-βέλτιστοι σχεδιασμοί

1. ΕΙΣΑΓΩΓΗ

Στην εργασία αυτή εξετάζουμε την περίπτωση που οι πειραματικές μονάδες (μονάδες) είναι σε διάταξη γραμμής και ασυσχέτιστες ή συσχετισμένες, με εξάρτηση AR(1). Η εργασία αυτή είναι συνέχεια της εργασίας Περικλέους και Κουνιάς (2010), στην οποία αναφερόμαστε ως (Π.Κ.2010).

Δίνουμε τους βέλτιστους σχεδιασμούς για την εκτίμηση των αντιπαραβολών (contrasts) των επιδράσεων τριών αγωγών. Στο 23^ο συνέδριο του ΕΣΙ δώσαμε τους βέλτιστους σχεδιασμούς για την εκτίμηση της διαφοράς των επιδράσεων δύο αγωγών (Π.Κ.2010), ενώ στο 22^ο συνέδριο αναπτύξαμε την έννοια της κυριαρχίας στους πειραματικούς σχεδιασμούς (Περικλέους και Κουνιάς 2009).

Ο ορισμός της συνάρτησης πληροφορίας $\varphi(C) \rightarrow \mathfrak{R} \forall C \in nnd(k)$ δόθηκε στην (Π.Κ.2010), όπου, $nnd(k)$ είναι μη αρνητικά ορισμένος και $pd(k)$: θετικά ορισμένος $k \times k$ πίνακας. Περισσότερα για τη συνάρτηση πληροφορίας και την κυριαρχία πινάκων περιέχονται στο βιβλίο του Pukelsheim, F. (1993).

Ορισμός 1.1 Ο σχεδιασμός d^* με συνάρτηση πληροφορίας $\varphi(C_{d^*})$ είναι καθολικά βέλτιστος, στην κλάση των σχεδιασμών F , αν μεγιστοποιεί την $\varphi(C_d)$, $\forall d \in F$ για κάθε συνάρτηση πληροφορίας $\varphi(C_d)$ •

Ορισμός 1.2 Ο σχεδιασμός d^* του πίνακα πληροφορίας C_{d^*} που έχει ιδιοτιμές $\lambda_1, \lambda_2, \dots, \lambda_v$, είναι φ -βέλτιστος, στην κλάση των σχεδιασμών F , αν ελαχιστοποιεί την $\sum_{i=1}^v \varphi(\lambda_i)$ $\forall d \in F$ για κάθε συνεχή φθίνουσα και κυρτή συνάρτηση $\varphi: \mathbb{R} \rightarrow \mathbb{R}$ •

Ο τελευταίος ορισμός ισχύει αν και μόνον αν το διάνυσμα $\lambda = (\lambda_1, \dots, \lambda_v)$ των ιδιοτιμών του C_{d^*} κυριαρχείται από το διάνυσμα $\mu = (\mu_1, \dots, \mu_v)$ των ιδιοτιμών του πίνακα πληροφορίας C_d , $\forall d \in F$ (Marshall and Olkin p.10).

Ένας καθολικά βέλτιστος σχεδιασμός είναι και φ -βέλτιστος. Το αντίθετο όμως δεν ισχύει. Για παράδειγμα, φ -βέλτιστος σχεδιασμός δεν είναι κατ' ανάγκη MV-βέλτιστος. Όταν ένας σχεδιασμός είναι φ -βέλτιστος είναι και A-, D-, E- βέλτιστος

Επίσης ένας σχεδιασμός μπορεί να είναι βέλτιστος για ένα κριτήριο αλλά όχι και για τα υπόλοιπα κριτήρια.

Επομένως αναζητούμε καθολικά βέλτιστους σχεδιασμούς και αν δεν υπάρχουν καταφεύγουμε στους φ -βέλτιστους. Στην περίπτωση που ούτε και αυτοί υπάρχουν βρίσκουμε A-, D-, E-, MV- κ.λπ., βέλτιστους σχεδιασμούς.

2. ΟΜΟΙΟΓΕΝΕΙΣ ΜΟΝΑΔΕΣ, ΕΞΑΡΤΗΜΕΝΕΣ ΠΑΡΑΤΗΡΗΣΕΙΣ

Με εξαρτημένες παρατηρήσεις σε βέλτιστους σχεδιασμούς έχουν ασχοληθεί πολλοί ερευνητές, όπως οι Morgan J.P. and Uddin, N. (2003). Περικλέους Κ. και Κουινιάς Σ. (2010), όπου υπάρχουν περισσότερες αναφορές..

Όταν έχουμε v αγωγές, το μοντέλο σε διανυσματική μορφή είναι,

$$y = \mu_1 x_1 + \dots + \mu_v x_v + e, \quad E(ee') = \sigma^2 V \quad (2.1)$$

Το $n \times 1$ διάνυσμα $x_j = (x_{1j}, \dots, x_{nj})'$ έχει $x_{ij} = 1$ αν στην μονάδα i εφαρμόζεται

η αγωγή T_j , $j = 1, \dots, v$ και 0 αλλού, οπότε $x'_s x_t = \begin{cases} n_s & \text{αν } s = t \\ 0 & \text{αν } s \neq t \end{cases}$.

Τα σφάλματα ακολουθούν ένα αυτό-παλινδρομούμενο μοντέλο AR(1),

$$e_i - a e_{i-1} = w_i, \quad |a| < 1, \quad i = 1, \dots, n, \quad (2.2)$$

Τα w_i είναι ανεξάρτητες τυχαίες μεταβλητές με μέση τιμή 0 και διασπορά σ^2 , τότε $\text{cov}(e_s, e_{s+t}) = \sigma^2 a^{|t|} / (1 - a^2)$ και $\text{var}(e) = \sigma^2 V$.

Ο πίνακας διασποράς V και ο αντίστροφός του δίνονται στην εργασία (Π.Κ 2010).

Οι παράμετροι που μας ενδιαφέρουν είναι μ_1, \dots, μ_v και ο πίνακας πληροφορίας είναι, $Q = X'V^{-1}X$, $X = (x_1, \dots, x_v)$, οπότε $Q = A + a^2 B - aC$ με

$$\mathbf{A} = \begin{bmatrix} n_1 & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & n_v \end{bmatrix}, \mathbf{B} = \begin{bmatrix} \tilde{n}_1 & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \tilde{n}_v \end{bmatrix}, \mathbf{C} = \begin{bmatrix} 2n_{11} & n_{12}+n_{21} & \dots & n_{1v}+n_{v1} \\ n_{12}+n_{21} & 2n_{22} & \dots & \vdots \\ \vdots & \ddots & \ddots & n_{v-1,v}+n_{v,v-1} \\ n_{1v}+n_{v1} & \dots & n_{v-1,v}+n_{v,v-1} & 2n_{vv} \end{bmatrix}.$$

Όπου $n_i, i=1, \dots, v$ είναι το πλήθος των μονάδων στις οποίες εφαρμόζεται η αγωγή T_i , στις θέσεις $1, 2, \dots, n$, \tilde{n}_i είναι το πλήθος των μονάδων στις οποίες εφαρμόζεται η αγωγή T_i στις εσωτερικές θέσεις $2, \dots, n-1$, $n_{ii}, i=1, \dots, v$ είναι το πλήθος των εμφανίσεων του ζεύγους $T_i T_i$ στις διαδοχικές θέσεις $(1,2), (2,3), \dots, (n-1,n)$, $n_{ij}, i \neq j$ είναι το πλήθος των εμφανίσεων του ζεύγους των αγωγών $T_i T_j, i \neq j$ στις θέσεις $(1,2), (2,3), \dots, (n-1,n)$.

Ορισμός 2.1 Το διάνυσμα \mathbf{x} υπερκυριαρχείται από το διάνυσμα \mathbf{y} και γράφεται $\mathbf{x} <^w \mathbf{y}$, αν: $\sum_{i=1}^k x_{(i)} \geq \sum_{i=1}^k y_{(i)}, k=1, \dots, n, x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}, y_{(1)} \leq y_{(2)} \leq \dots \leq y_{(n)}$. •

Αυτός ο ορισμός είναι ισοδύναμος (Marshall and Olkin. p.10) με: $\mathbf{x} = \mathbf{S}\mathbf{y}$, όπου $\mathbf{S} = (s_{ij}), s_{ij} \geq 0$ είναι $k \times k$ πίνακας και υπάρχει διπλά στοχαστικός πίνακας $\mathbf{D} = (d_{ij}), s_{ij} \leq d_{ij}$ (άθροισμα γραμμών και στηλών ίσο με 1).

Ορισμός 2.2 Αν F είναι μια κλάση σχεδιασμών, ο σχεδιασμός $d \in F$ με πίνακα πληροφορίας Q και ιδιοτιμές $\lambda_1, \dots, \lambda_v, \lambda_{(1)} \leq \lambda_{(2)} \leq \dots \leq \lambda_{(v)}$, θα λέγεται φ -βέλτιστος, στην κλάση F , αν ελαχιστοποιεί το $\sum_{i=1}^v g(\lambda_{(i)})$, για κάθε συνεχή, φθίνουσα και κυρτή

συνάρτηση $g(x): \mathfrak{R} \rightarrow \mathfrak{R}$. •

Για να αποδείξουμε ότι ένας σχεδιασμός είναι φ -βέλτιστος είναι ευκολότερο να χρησιμοποιήσουμε την έννοια της κυριαρχίας ή υπερκυριαρχίας.

Θεώρημα 2.1 (Φίλτρο 1) Αν το μοντέλο (2.1) είναι σωστό και ισχύει η συνθήκη (2.2), τότε:

(i) Αν $0 < a < 1$ στο βέλτιστο σχεδιασμό υπάρχει ροή (δηλαδή συνεχόμενη επανάληψη αγωγής) το πολύ μιας αγωγής $T_i, i=1, \dots, v$ με $n_{ii} > 1$.

(ii) Αν $-1 < a < 0$ στο βέλτιστο σχεδιασμό κάθε ζεύγος αγωγών $T_i T_j, i \neq j$ εμφανίζεται το πολύ μια φορά.

(iii) Αν $a = 0$, στον καθολικά βέλτιστο σχεδιασμό τα $n_i, n_j, i \neq j$ διαφέρουν το πολύ κατά 1, $|n_i - n_j| \leq 1$.

(iv) Η ακολουθία των αγωγών ενός σχεδιασμού έχει τον ίδιο πίνακα πληροφορίας με την ανάστροφη της και δύο τέτοιες ακολουθίες θεωρούνται ισοδύναμες.

(v) Οι δύο σχεδιασμοί $\dots T_i f T_i \dots, \dots \bar{T}_i \bar{f} \bar{T}_i \dots$ έχουν τον ίδιο πίνακα πληροφορίας, όπου f είναι μια ακολουθία αγωγών και \bar{f} είναι η ανάστροφή της.

(vi) Αν κάνουμε μετάθεση δύο ή περισσότερων γραμμών, ο σχεδιασμός που προκύπτει έχει τον ίδιο πίνακα πληροφορίας. •

Για την απόδειξη δεξ (Π.Κ.2010). Στο (vi) η μετάθεση δύο γραμμών ουσιαστικά είναι ή μετάθεση των αντίστοιχων γραμμών και στηλών του πίνακα πληροφορίας \mathbf{Q} .

Φίλτρο 2. Αν d είναι ένας σχεδιασμός με πίνακα πληροφορίας \mathbf{Q} και $\bar{\mathbf{Q}} = c\mathbf{Q} + (1-c)\mathbf{PQP}'$, $0 < c < 1$, τότε σύμφωνα με τον ορισμό της συνάρτησης πληροφορίας θα είναι, $g(\bar{\mathbf{Q}}) \geq g(\mathbf{Q})$ για κάθε συνάρτηση πληροφορίας $g(\cdot)$, όπου \mathbf{P} είναι ένας μεταθετικός πίνακας. •

Φίλτρο 3. Αν $0 < a < 1$, f_1, f_2, f_3 είναι ακολουθίες αγωγών και δίνεται ο σχεδιασμός $d_1 : f_1 S S f_2 T T f_3$, τότε ο σχεδιασμός $d_2 : f_1 S T S f_2 T f_3$ έχει $g(\mathbf{Q}_2) \geq g(\mathbf{Q}_1)$ όπου $g(\cdot)$ είναι συνάρτηση πληροφορίας. Αυτό συμβαίνει διότι $\mathbf{Q}_2 \geq \mathbf{Q}_1$, για την απόδειξη δεξ (Π.Κ. 2010). •

Φίλτρο 4. Αν $-1 < a < 0$, f είναι ακολουθία αγωγών και δίνεται ο σχεδιασμός $d_1 : f_1 S T S f_2 T f_3$, τότε ο σχεδιασμός $d_2 : f_1 S S f_2 T T f_3$ έχει $g(\mathbf{Q}_2) \geq g(\mathbf{Q}_1)$ όπου $g(\cdot)$ είναι συνάρτηση πληροφορίας. Αυτό συμβαίνει διότι $\mathbf{Q}_2 \geq \mathbf{Q}_1$, για την απόδειξη δεξ (Π.Κ. 2010). •

Φίλτρο 5. Αν $\mathbf{Q}_1, \mathbf{Q}_2$ είναι πίνακες $n \times n$ και $\mathbf{Q}_2 \geq \mathbf{Q}_1$, τότε $g(\mathbf{Q}_2) \geq g(\mathbf{Q}_1)$ για κάθε συνάρτηση πληροφορίας $g(\cdot)$, όπως προκύπτει από τον ορισμό της συνάρτησης πληροφορίας.

3. ΤΡΕΙΣ ΑΓΩΓΕΣ S,T,R

Με τρεις αγωγές είναι μεγάλο το πλήθος των προς σύγκριση σχεδιασμών και αυξάνεται όσο αυξάνει το πλήθος n των μονάδων.

Μια διέξοδος είναι να χρησιμοποιήσουμε τα φίλτρα 1, 2, 3, 4 και 5 ώστε να μειωθούν οι σχεδιασμοί. Σε κάθε σχεδιασμό μετέχουν όλες οι αγωγές με μία τουλάχιστον παρατήρηση. Αν μια αγωγή δεν μετέχει στο σχεδιασμό, δεν μπορεί να εκτιμηθεί.

Ονομάζουμε F την κλάση των σχεδιασμών που δεν μπορεί να μειωθεί περισσότερο με τη χρήση των φίλτρων 1,2,3,4,5.

3.1 Περίπτωση 1^η: $0 < a < 1$

Για $n=4, 5, 6$ δίνουμε την κλάση F των σχεδιασμών προς σύγκριση και στο Παράρτημα-Πίνακες δίνονται οι A, D, E, MV -βέλτιστοι σχεδιασμοί για $a=0.1, \dots, 0.9$.

Παίρνουμε την S ως αγωγή που μπορεί να επαναλαμβάνεται.

(i) Όταν $n = 3$, $0 < a < 1$, ο μόνος σχεδιασμός είναι $d : STR$.

(ii) Όταν $n = 4$, $0 < a < 1$, η κλάση F των υπό εξέταση σχεδιασμών είναι,

$d_1 : SSSTR, d_2 : STSR, d_3 : STRS, d_4 : TSSR.$

(ι) Όταν $n = 5, 0 < a < 1$, αρχικά βρίσκουμε 12 σχεδιασμούς,

$d_1 : SSSTR, d_2 : TSSSR, d_3 : SSSTRS, d_4 : SSSTR, d_5 : STSR, d_6 : TSSSR$

$d_7 : TSSSTR, d_8 : SSSTR, d_9 : TSSSTR, d_{10} : STRST, d_{11} : STRTS, d_{12} : STSTR$

Όμως $d_{10} \Leftrightarrow d_{12}$ (φίλτρο 1v), $d_1 \rightarrow d_{12}$ (φίλτρο 2 και 3), $d_3 \Rightarrow d_{12}$ (φίλτρο 2 και 5), $d_4 \rightarrow d_{10}$ (φίλτρο 2, 5, 1(vi)), $d_5 \rightarrow d_{10}$ (φίλτρο 2), $d_6 \rightarrow d_{10}$ (φίλτρο 2, 5), $d_8 \rightarrow d_{10}$ (φίλτρο 2, 4), $d_9 \rightarrow d_{10}$ (φίλτρο 2, 4), $d_9 \rightarrow d_{10}$ (φίλτρο 2), $d_{12} \rightarrow d_{10}$ (φίλτρο 1(v)).

Σημείωση: Το σύμβολο \Leftrightarrow δείχνει ότι οι δύο σχεδιασμοί έχουν τον ίδιο πίνακα πληροφορίας και $d_1 \rightarrow d_2$ σημαίνει ότι ο σχεδιασμός d_2 είναι «καλύτερος» του d_1 ($g(\mathbf{Q}_2) \geq g(\mathbf{Q}_1)$), όπου $g(\cdot)$ είναι συνάρτηση πληροφορίας. Έτσι για $n = 5, 0 < a < 1$ η κλάση F έχει 3 σχεδιασμούς, $d_2 : TSSSR, d_7 : TSSSTR, d_{10} : STRST$.

(ιι) Τέλος, όταν $n = 6, 0 < a < 1$, εργαζόμενοι όπως στη (ι) βρίσκουμε ότι η κλάση F έχει 10 σχεδιασμούς,

$d_1 : TSSSSR, d_2 : SSSTRS, d_3 : TSSSR, d_4 : SSSTR, d_5 : TSSSTR$

$d_6 : STSSR, d_7 : SSSTR, d_8 : STRSTR, d_9 : SSSSTR, d_{10} : TSSSTR$

3.2 Περίπτωση 2^η: $-1 < a < 0$

(ι) Όταν $n = 3, -1 < a < 0$, ο μόνος σχεδιασμός είναι $d : STR$.

(ιι) Όταν $n = 4, -1 < a < 0$, η κλάση F των υπό εξέταση σχεδιασμών είναι,

$d_1 : SSSTR, d_2 : STSR, d_3 : STRS, d_4 : TSSR$

(ιιι) Όταν $n = 5, -1 < a < 0$, βρίσκουμε αρχικά 17 σχεδιασμούς, αλλά με την εφαρμογή των φίλτρων καταλήγουμε σε 4 σχεδιασμούς,

$d_2 : TSSSR, d_7 : TSSSTR, d_{11} : SSSTR, d_{10} : SSSTR$.

(ιιιι) Όταν $n = 6, -1 < a < 0$, όμοια όπως στη (ιι) βρίσκουμε ότι η κλάση F έχει 11 σχεδιασμούς,

$d_1 : TSSSSR, d_2 : SSSTR, d_3 : SSSTRS, d_4 : TSSSTR, d_5 : SSSTR, d_6 : TTSSSR$

$d_7 : SSSSTR, d_8 : SSSTRS, d_9 : SSSTRS, d_{10} : TSSSTR, d_{11} : SSSTR$.

4. ΠΑΡΑΡΤΗΜΑ-ΠΙΝΑΚΕΣ

Για σχεδιασμούς με πλήθος μονάδων $n=4,5,6$ και για τιμές $a = -0.9, \dots, -0.1$ και $a = 0.1, \dots, 0.9$ δίνονται οι E, D, A, MV-βέλτιστοι σχεδιασμοί.

. (i) Πίνακας βέλτιστων σχεδιασμών, $n = 4, u = 3, -1 < a < 0$

$d_1 : SSTR, d_2 : STSR, d_3 : STRS, d_4 : TSSR.$

α	-0,1	-0,2	-0,3	-0,4	-0,5	-0,6	-0,7	-0,8	-0,9
E	d_4	d_4	d_4	d_4	d_4	d_4	d_4	d_3	d_3
D	d_4	d_1, d_4	d_1, d_4	d_1, d_4	d_1, d_4	d_1, d_4	d_1, d_4	d_1, d_4	d_1, d_4
A	d_4	d_4	d_4	d_4	d_4	d_4	d_3	d_3	d_3
MV	d_3	d_3	d_3	d_3	d_3	d_3	d_3	d_3	d_3

. (ii) Πίνακας βέλτιστων σχεδιασμών, $n = 5, u = 3, -1 < a < 0$

$d_2 : TSSSR, d_7 : TSSTR, d_{11} : SSTTR, d_{10} : SSRTT.$

α	-0,1	-0,2	-0,3	-0,4	-0,5	-0,6	-0,7	-0,8	-0,9
E	d_{10}	d_{10}	d_{10}	d_{10}	d_{10}	d_2	d_{11}	d_{11}	d_{11}
D	d_{10}	d_{10}	d_{10}	d_{10}	d_{10}	d_2	d_2	d_{11}	d_{11}
A	d_{10}	d_{10}	d_{10}	d_{10}	d_{10}	d_{11}	d_{11}	d_{11}	d_{11}
MV	d_{10}	d_{10}	d_{10}	d_{10}	d_{10}	d_{11}	d_{11}	d_{11}	d_{11}

. (iii) Πίνακας βέλτιστων σχεδιασμών, $n = 6, u = 3, -1 < a < 0$

$d_1 : TSSSSR, d_2 : SSSTSR, d_3 : SSSTRS, d_4 : TSSSTR, d_5 : SSSTRT, d_6 : TTSSSR$

$d_7 : SSSRTT, d_8 : SSTSRS, d_9 : SSTRTS, d_{10} : TSSRST, d_{11} : SSTTRR$

α	-0,1	-0,2	-0,3	-0,4	-0,5	-0,6	-0,7	-0,8	-0,9
E	d_{11}	d_{11}	d_{11}	d_{11}	d_{11}	d_{11}	d_{11}	d_{11}	d_{11}
D	d_{11}	d_{11}	d_{11}	d_{11}	d_{11}	d_{11}	d_{11}	d_{11}	d_{11}
A	d_{11}	d_{11}	d_{11}	d_{11}	d_{11}	d_{11}	d_{11}	d_{11}	d_{11}
MV	d_{11}	d_{11}	d_{11}	d_{11}	d_{11}	d_{11}	d_{11}	d_{11}	d_{11}

Παρατηρούμε ότι στην περίπτωση αυτή, ο σχεδιασμός SSTTR είναι καθολικά βέλτιστος, για όλες τις τιμές του α .

(iv) Πίνακας βέλτιστων σχεδιασμών, $n = 4, 0 < \alpha < 1$

$d_1 : SSTR, d_2 : STSR, d_3 : STRS, d_4 : TSSR.$

α	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
E	d_4	d_4	d_4	d_4	d_4	d_4	d_4	d_4	d_4
D	d_3	d_3	d_3	d_3	d_3	d_3	d_3	d_3	d_3
A	d_3	d_3	d_3	d_3	d_3	d_3	d_3	d_3	d_3
MV	d_4	d_4	d_4	d_4	d_4	d_4	d_4	d_4	d_4

(v) Πίνακας βέλτιστων σχεδιασμών, $n = 5, 0 < \alpha < 1$

$d_1 : TSSSR, d_2 : TSSTR, d_3 : STRST$

α	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
E	d_2	d_2	d_2	d_2	d_2	d_2	d_2	d_2	d_2
D	d_3	d_3	d_3	d_3	d_3	d_3	d_3	d_3	d_3
A	d_3	d_3	d_3	d_3	d_2	d_2	d_2	d_2	d_2
MV	d_3	d_3	d_3	d_3	d_3	d_2	d_2	d_2	d_2

(vi) Πίνακας βέλτιστων σχεδιασμών, $n = 6, 0 < \alpha < 1$

$d_1 : TSSSSR, d_2 : SSSTRS, d_3 : TSSRS, d_4 : SSSTRT, d_5 : TSSSTR,$

$d_6 : STSSRS, d_7 : SSTRST, d_8 : SSTRST, d_9 : STRSRT, d_{10} : TSSRTR.$

α	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
E	d_9	d_9	d_9	d_9	d_9	d_9	d_9	d_9	d_9
D	d_9	d_9	d_9	d_9	d_9	d_9	d_9	d_9	d_9

A	d_9	d_9	d_9	d_9	d_9	d_9	d_9	d_1, d_9	d_9
MV	d_9	d_9	d_9	d_9	d_9	d_9	d_9	d_1, d_8, d_9	d_9

ABSTRACT

Row designs are examined with three treatments. The population is homogeneous with dependent observations. The dependence follows a first order autoregression with parameter α . A filtering procedure is presented to reduce the number of competing designs. Tables are given for 4,5,6 units with the A,D,E,MV-optimal designs.

ΑΝΑΦΟΡΕΣ

- Marshall, A.A and Olkin, I. (1979). *Inequalities: Theory of majorization and its applications*. Academic Press, New York.
- Morgan J.P. and Uddin, N. (2003). Optimal row-column design for two treatments. *Journal of Statistical Planning and Inference*, 115, 603-622.
- Pukelsheim, F. (1993). *Optimal designs of Experiments*. John Wiley and Sons, Inc.
- Περικλέους Κ. και Κουνιάς Σ. (2009). Η έννοια της κυριαρχίας στους πειραματικούς σχεδιασμούς, *Πρακτικά 22^ο Συνέδριο του ΕΣΙ*, 167-174.
- Περικλέους Κ. και Κουνιάς Σ. (2010). Βέλτιστοι σχεδιασμοί Γραμμής-Στήλης. *Πρακτικά ΕΣΙ 23^ο Συνέδριο*, 239-246.

Κλάσεις βελτιωμένων διαστημάτων εμπιστοσύνης για την παράμετρο κλίμακας από την πολυδιάστατη κατανομή Lomax.

Κωνσταντίνος Πετρόπουλος

Τμήμα Μαθηματικών, Πανεπιστήμιο Πατρών

costas@math.upatras.gr

ΠΕΡΙΛΗΨΗ

Σε αυτήν την εργασία κατασκευάζουμε δύο κλάσεις βελτιωμένων διαστημάτων εμπιστοσύνης για την παράμετρο κλίμακας σ από την κατανομή Lomax, όταν η παράμετρος θέσης μ είναι άγνωστη. Η πρώτη, είναι μια διπαραμετρική κλάση ομαλών διαστημάτων εμπιστοσύνης $I_{a,b}$, με $a \geq 1$ και b καθορισμένο, έτσι ώστε να έχει τον ίδιο λόγο άκρων με το ελαχίστου λόγου άκρων διάστημα εμπιστοσύνης I_0 , αλλά μεγαλύτερη πιθανότητα κάλυψης. Μέσα σε αυτήν την κλάση περιέχεται και το διάστημα εμπιστοσύνης τύπου Brewster and Zidek. Επίσης, υπολογίζουμε και άλλη μία υποκλάση διαστημάτων εμπιστοσύνης που βελτιώνουν την πιθανότητα κάλυψης για όλες τις τιμές των παραμέτρων. Η δεύτερη αυτή κλάση έχει μια πολύ απλή συναρτησιακή μορφή και περιέχει το διάστημα εμπιστοσύνης τύπου Stein.

Λέξεις Κλειδιά: Διαστήματα Εμπιστοσύνης, πολυδιάστατη κατανομή Lomax, παράμετρος κλίμακας.

1. ΕΙΣΑΓΩΓΗ

Η από κοινού κατανομή των X_1, X_2, \dots, X_n λέγεται μείζη εκθετικών κατανομών αν,

$$f_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n) = \int_{-\infty}^{\infty} \frac{\tau^n}{\sigma^n} e^{-\frac{\tau}{\sigma} \sum_{i=1}^n (x_i - \mu)} I_{[\mu, +\infty)}(x_{(1)}) dG(\tau). \quad (1)$$

όπου η παράμετρος θέσης $\mu \in \mathbb{R}$ και η παράμετρος κλίμακας $\sigma > 0$ θεωρούνται άγνωστες, ενώ $x_{(1)} = \min\{x_i : i = 1, \dots, n\}$. Η $G(\cdot)$ είναι γνωστή συνάρτηση κατανομής. Η ερμηνεία του μοντέλου (1) είναι ότι, δοθέντος $\tau > 0$, X_1, \dots, X_n είναι ανεξάρτητες και ισόνομες εκθετικές τυχαίες μεταβλητές $E(\mu, \sigma/\tau)$, και η μεταβλητή μείζης τ έχει κατανομή $G(\cdot)$. Για λόγους απλότητας, συμβολίζουμε με τ είτε την τυχαία μεταβλητή είτε οποιαδήποτε τιμή της. Όταν $\tau \sim \text{Gamma}(\gamma, 1)$, τότε το διάνυσμα (X_1, X_2, \dots, X_n) ακολουθεί την Πολυδιάστατη κατανομή Lomax

$(ML_n(\mu, \sigma))$ με πυκνότητα πιθανότητας,

$$f(x_1, \dots, x_n; \mu, \sigma) = \frac{\Gamma(n + \gamma)}{\Gamma(\gamma)\sigma^n} \frac{1}{\left(1 + \frac{1}{\sigma} \sum_{i=1}^n (x_i - \mu)\right)^{n+\gamma}} I_{[\mu, \infty)}(x_{(1)}). \quad (2)$$

Το μοντέλο (2) είναι μια πολυδιάστατη γενίκευση της κατανομής Lomax (Pareto II), η οποία έχει π.π. $\frac{a}{\sigma} \left(1 + \frac{x - \mu}{\sigma}\right)^{-1-a}$, $x > \mu$ και είναι ένα χρήσιμο μοντέλο για θέματα κοινωνικοοικονομικά, καθώς και για εκείνα που αφορούν χρόνους ζωής. Έχει χρησιμοποιηθεί αρχικά από τον Lomax (1954) σε οικονομικά ζητήματα και έχει προταθεί από τον Bryson (1974) σαν ένα εναλλακτικό μοντέλο της διπαραμετρικής εκθετικής κατανομής. Η επαρκής στατιστική συνάρτηση είναι (X, S) , όπου $X = X_{(1)} = \min\{X_1, X_2, \dots, X_n\}$ και $S = \frac{1}{n} \sum_{i=1}^n (X_i - X_{(1)})$. Δεσμεύοντας στην τ , οι τυχαίες μεταβλητές S και X είναι ανεξάρτητες με κατανομές, αντίστοιχα,

$$S|\tau \sim \text{Gamma}(n - 1, \sigma/(n\tau)) \quad , \quad X|\tau \sim E(\mu, \sigma/(n\tau)).$$

Οι αναλλοίωτοι εκτιμητές του σ , στο μοντέλο (2), είναι της μορφής

$$\delta = cS, \quad c > 0$$

και ανάμεσά τους ο βέλτιστος, ως προς τη συνάρτηση ζημίας του Stein $R(\delta; \sigma) = \frac{\delta}{\sigma} - \ln \frac{\delta}{\sigma} - 1$, είναι ο $\delta_0 = \frac{1}{\gamma-1} \frac{1}{n-1} S$. Οι Petropoulos and Kourouklis (2005) κατασκεύασαν βελτιωμένους εκτιμητές του δ_0 , χρησιμοποιώντας την επιπλέον πληροφορία που υπάρχει στην X , όπως εκτιμητές τύπου Stein (1964) και Brewster and Zidek (1974). Ο Petropoulos (2010) κατασκεύασε μία κλάση βελτιωμένων εκτιμητών για την παράμετρο κλίμακας σε μείζη εκθετικών κατανομών (άρα και για την $ML_n(\mu, \sigma)$). Η μεθοδολογία ήταν ανάλογη με αυτή του Maruyama (1998), ο οποίος ασχολήθηκε με εκτίμηση της διασποράς από μια κανονική κατανομή. Η βασική ιδέα ήταν να κατασκευαστούν γενικευμένοι εκτιμητές Bayes, δ_a , για το $a\sigma$, $a > 0$, και να προταθούν αυτοί οι εκτιμητές

$$\{\delta_a : a \geq 1\}$$

για την εκτίμηση της παραμέτρου κλίμακας σ .

Στην παρούσα εργασία μας ενδιαφέρει το πρόβλημα εκτίμησης της παραμέτρου κλίμακας σ στο μοντέλο (2), μέσω διαστημάτων εμπιστοσύνης. Αρχικά, όπως και στο πρόβλημα της σημειακής εκτίμησης, υπολογίζουμε διαστήματα εμπιστοσύνης τα οποία βασίζονται μόνο στην S . Υπολογίζεται ότι το Διάστημα Εμπιστοσύνης Ελαχίστου Λόγου Άκρων είναι της μορφής $I_0 = \left[\frac{S}{c_2}, \frac{S}{c_1}\right]$, όπου $0 < c_1 < c_2$ τ.ω.

$$\int_0^{+\infty} c_1 \tau g(\tau c_1) dG(\tau) = \int_0^{+\infty} c_2 \tau g(\tau c_2) dG(\tau), \quad P\left(\frac{S}{c_2} < \sigma < \frac{S}{c_1}\right) = 1 - \beta \quad (3)$$

όπου $g(\cdot)$ είναι η π.π. της κατανομής $Gamma(n - 1, 1/\tau)$, $G(\cdot)$ είναι η μεικτική συνάρτηση κατανομής ($Gamma(\gamma, 1)$) και $0 < \beta < 1$.

Ο Kubokawa (1994) θεώρησε Δ.Ε. της μορφής,

$$I_\phi = \begin{cases} \left(\frac{\phi(W)S}{c_2}, \frac{\phi(W)S}{c_1} \right) & , \quad W > 0 \\ I_0 = \left(\frac{S}{c_2}, \frac{S}{c_1} \right) & , \quad \text{διαφορετικά} \end{cases}$$

όπου $W = X/S$. I_ϕ έχει τον ίδιο λόγο άκρων, όπως το I_0 , έχει όμως μεγαλύτερη πιθανότητα κάλυψης από το I_0 εάν

(a) Η $\phi(w)$ είναι μη φθίνουσα και $\lim_{w \rightarrow +\infty} \phi(w) = 1$,

(b) $c_2 g(c_2/\phi(w))H(c_2 w/\phi(w)) - c_1 g(c_1/\phi(w))H(c_1 w/\phi(w)) \geq 0$,

όπου $H(\cdot)$ είναι η σ.κ. της $Gamma(1, 1/n)$. Στην περίπτωση της μείζης εκθετικών κατανομών η συνθήκη (b) του Kubokawa παίρνει την εξής μορφή,

(b') $\int_0^\infty \{ \tau c_2 g(\tau c_2/\phi(w))H(\tau c_2 w/\phi(w)) - \tau c_1 g(\tau c_1/\phi(w))H(\tau c_1 w/\phi(w)) \} dG(\tau) \geq 0$

όπου $G(\cdot)$ είναι η μεικτική συνάρτηση κατανομής ($Gamma(\gamma, 1)$).

Χρησιμοποιώντας τις συνθήκες (a) και (b') μπορούμε να βρούμε βελτιωμένα Δ.Ε. του I_0 , τύπου Stein, I_S , και τύπου Brewster and Zidek, I_{BZ} , αντίστοιχα.

$$I_S = \begin{cases} \left(\frac{\phi_S(W)S}{c_2}, \frac{\phi_S(W)S}{c_1} \right) & , \quad W > 0 \\ I_0 & , \quad \text{διαφορετικά} \end{cases}$$

όπου $\phi_S(W) = \min \left\{ 1, \frac{k c_2 - c_1}{1 - k} (1 + W) \right\}$, $W = X/S$ και $k = \left(\frac{c_1}{c_2} \right)^{n/(n+\gamma)}$.

$$I_{BZ} = \begin{cases} \left(\frac{\phi_{BZ}(W)S}{c_2}, \frac{\phi_{BZ}(W)S}{c_1} \right) & , \quad W > 0 \\ I_0 & , \quad \text{διαφορετικά} \end{cases}$$

όπου $\phi = \phi_{BZ}(W)$, $W = X/S$ είναι η μοναδική λύση της εξίσωσης

$$\int_0^\infty \tau c_2 g(\tau c_2/\phi(w))H(\tau c_2 w/\phi(w))dG(\tau) = \int_0^\infty \tau c_1 g(\tau c_1/\phi(w))H(\tau c_1 w/\phi(w))dG(\tau). \quad (4)$$

Στην Ενότητα 2, κατασκευάζουμε μία διπαραμετρική κλάση ομαλών Δ.Ε. $I_{a,b}$, με $a \geq 1$ και b καθορισμένο, έτσι ώστε να έχει τον ίδιο λόγο άκρων με το ελαχίστου λόγου άκρων διάστημα εμπιστοσύνης I_0 , αλλά μεγαλύτερη πιθανότητα κάλυψης,

χρησιμοποιώντας μεθοδολογία αντίστοιχη με αυτή του Maruyama (1998). Μια υποκλάση της κατασκευής μας, η οποία αντιστοιχεί στο $a = 1$, είναι, ουσιαστικά, μια κλάση βελτιωμένων γενικευμένων Bayes Δ.Ε. για την παράμετρο σ , η οποία περιέχει το Δ.Ε. τύπου Brewster and Zidek, I_{BZ} . Επίσης, για $a > 1$, $I_{a,b}$ έχει αυστηρά μεγαλύτερη πιθανότητα κάλυψης από το I_0 , για κάθε $\varrho = \frac{\mu}{\sigma} \in \mathbb{R}$. Αυτό το αποτέλεσμα είναι ενδιαφέρον γιατί έτσι κατορθώνουμε να κατασκευάσουμε ομαλά Δ.Ε. τα οποία βελτιώνουν την πιθανότητα κάλυψης όταν $\varrho = 0$.

Στην Ενότητα 2, δίνουμε μια εναλλακτική κλάση βελτιωμένων Δ.Ε. $I_{a,b}^S$ με $a \geq 1$ και $b > -1$. Τα Δ.Ε. $I_{a,b}^S$, αν και μη ομαλά, έχουν μια πολύ απλή μορφή και θα μπορούσαν να χρησιμοποιηθούν για πρακτικούς λόγους. Μάλιστα αποδεικνύεται ότι, $I_{1,0}^S = I_S$.

2. ΒΕΛΤΙΩΜΕΝΑ ΔΙΑΣΤΗΜΑΤΑ ΕΜΠΙΣΤΟΣΥΝΗΣ

Σε αυτήν την ενότητα κατασκευάζουμε μία κλάση Δ.Ε. για το σ στο μοντέλο της πολυδιάστατης κατανομής Lomax με τον ίδιο λόγο άκρων όπως το I_0 αλλά μεγαλύτερη πιθανότητα κάλυψης. Χρησιμοποιώντας μια ιδέα του Maruyama (1998), παράγουμε γενικευμένα Bayes Δ.Ε. για το $a\sigma$, $a > 0$, επιλέγοντας μια κατάλληλη εκ των προτέρων κατανομή και στη συνέχεια προτείνουμε αυτά τα Δ.Ε. ως κλάση Δ.Ε. για το σ . Έστω $\eta = \frac{1}{a\sigma} > 0$ και θεωρούμε την εκ των προτέρων κατανομή, $\mu|\eta, \lambda, \tau \sim \mathcal{E}(1/(n\eta\lambda\tau))$, $0 < \lambda < 1$, $\lambda \propto \lambda^b(1-\lambda) I_{(0,1)}(\lambda)$ και $\eta|\tau \propto (\tau\eta)^{b+(a-1)(n-1)} \tau^{(a-1)\gamma+2} e^{-\tau(a-1)} I_{(0,+\infty)}(\eta)$. Τότε η εκ των υστέρων π.π. του η , δοθέντος $X = x$ και $S = s$, $\pi(\eta|x, s)$ είναι,

$$\pi(\eta|x, s) \propto \eta^{b+a(n-1)} \int_0^\infty e^{-\tau(a-1)} \tau^{(a-1)\gamma+2+b+a(n-1)} \int_0^1 \lambda^b e^{-n\eta\tau(x+\lambda s)} d\lambda dG(\tau).$$

Προτείνουμε Δ.Ε. για το $a\sigma$ της μορφής

$$I = \left(\frac{\psi(X, S)}{c_2}, \frac{\psi(X, S)}{c_1} \right) \quad (5)$$

που έχει τον ίδιο λόγο άκρων όπως το I_0 και για κάθε (x, s) , ορίζουμε $\psi = \psi(x, s)$ τ.ω. να μεγιστοποιεί, ως προς ψ , την εκ των υστέρων πιθανότητα κάλυψης

$$\begin{aligned} P\left(\frac{\psi}{c_2} < a\sigma < \frac{\psi}{c_1} | X = x, S = s\right) &= P\left(\frac{c_1}{\psi} < \eta < \frac{c_2}{\psi} | X = x, S = s\right) \\ &= \int_{c_1/\psi}^{c_2/\psi} \pi(\eta|x, s) d\eta \end{aligned}$$

όπου $\pi(\eta|x, s)$ είναι η εκ των υστέρων π.π. της η .

Επομένως,

$$P(a\sigma \in I) \propto \int_{c_1/\psi}^{c_2/\psi} \eta^{b+a(n-1)} \int_0^\infty e^{-\tau(a-1)} \tau^{(a-1)\gamma+2+b+a(n-1)} \int_0^1 \lambda^b e^{-n\eta\tau(x+\lambda s)} d\lambda dG(\tau) d\eta \quad (6)$$

Η ολοκληρωτέα ποσότητα στη Σχέση (6) είναι μια μονοκόρυφη συνάρτηση του η . Έτσι, $\psi = \psi_0(x, s)$ που μεγιστοποιεί την εκ των υστέρων πιθανότητα κάλυψης είναι η μοναδική λύση της εξίσωσης

$$\begin{aligned} c_1^{b+a(n-1)+1} \int_0^\infty e^{-\tau(a-1)} \tau^{(a-1)\gamma+2+b+a(n-1)} \int_0^1 \lambda^b e^{-na\tau(c_1/\psi)(x+\lambda s)} d\lambda dG(\tau) = \\ = c_2^{b+a(n-1)+1} \int_0^\infty e^{-\tau(a-1)} \tau^{(a-1)\gamma+2+b+a(n-1)} \int_0^1 \lambda^b e^{-na\tau(c_2/\psi)(x+\lambda s)} d\lambda dG(\tau) \end{aligned} \quad (7)$$

όπου $b > -1$ και $a \geq 1$. Θέτοντας $w = \frac{x}{s}$, $\frac{s}{\psi} = \frac{1}{\phi}$ και $u = \frac{acw\lambda}{\phi}$, $c = c_1, c_2$, αναπαριστούμε την (7) ως

$$\begin{aligned} c_1^{a(n-1)} \int_0^{awc_1/\phi} u^b \int_0^\infty e^{-\tau(a-1)} \tau^{(a-1)\gamma+2+b+a(n-1)} e^{-n\tau u} e^{-na\tau(c_1/\phi)} dG(\tau) du = \\ c_2^{a(n-1)} \int_0^{awc_2/\phi} u^b \int_0^\infty e^{-\tau(a-1)} \tau^{(a-1)\gamma+2+b+a(n-1)-1} e^{-n\tau u} e^{-na\tau(c_2/\phi)} dG(\tau) du. \end{aligned}$$

Επειδή ϕ εξαρτάται από τα a, b και w , συμβολίζουμε ως $\phi = \phi_{a,b}(w)$.

Άρα $\phi = \phi_{a,b}(w)$ ικανοποιεί την εξίσωση

$$\int_0^\infty c_1 \xi(\tau) g_a \left(\frac{a\tau c_1}{\phi} \right) H_b \left(\frac{aw\tau c_1}{\phi} \right) dG(\tau) = \int_0^\infty c_2 \xi(\tau) g_a \left(\frac{a\tau c_2}{\phi} \right) H_b \left(\frac{aw\tau c_2}{\phi} \right) dG(\tau) \quad (8)$$

όπου $\xi(\cdot)$ είναι η π.π. της κατανομής $Gamma((a-1)\gamma+2, 1/(a-1))$, $g_a(\cdot)$ είναι η π.π. της $Gamma(a(n-1), 1/n)$, $H_b(\cdot)$ είναι η συνάρτηση κατανομής της $Gamma(b+1, 1/n)$, $a \geq 1$ και $b > -1$.

Τελικά προτείνουμε την κλάση των Δ.Ε. για το σ

$$\left\{ I_{a,b} = \begin{cases} \left(\frac{\phi_{a,b}(W)S}{c_2}, \frac{\phi_{a,b}(W)S}{c_1} \right) & , W > 0 \\ I_0 & , \text{διαφορετικά} \end{cases} ; a \geq 1, -1 < b \leq 2(a-1) \right\}$$

όπου $\phi_{a,b}(w)$ ορίζεται στην (8).

Στο επόμενο θεώρημα, αποδεικνύεται το κύριο αποτέλεσμα αυτής της ενότητας.

Θεώρημα 1. Έστω $a \geq 1$ και $-1 < b \leq 2(a-1)$. Τότε ισχύουν τα παρακάτω,

(i) $I_{a,b}$ βελτιώνει, ως προς την πιθανότητα κάλυψης το Δ.Ε. I_0 , δηλ. $P(\sigma \in I_{a,b}) \geq P(\sigma \in I_0)$ και ισχύει αυστηρή ανισότητα για κάθε $\rho = \mu/\sigma \in \mathbb{R}$, αν $a > 1$.

(ii) $\{I_{1,b}; -1 < b \leq 0\}$ είναι μία κλάση βελτιωμένων γενικευμένων Bayes Δ.Ε. για το σ . Επιπλέον $I_{1,b}$, $-1 < b \leq 0$, έχει πιθανότητα κάλυψης αυστηρά μεγαλύτερη από το $1 - \beta$, για κάθε $\rho \in \mathbb{R}$.

Παρατήρηση 2..1. Το Δ.Ε. τύπου *Brewster and Zidek* μπορεί να παραχθεί από την κλάση που κατασκευάσαμε για $a = 1$ και $b = 0$, δηλ. $I_{BZ} = I_{1,0}$. Επισημαίνεται ότι το Δ.Ε. I_{BZ} δε βελτιώνει την πιθανότητα κάλυψης στο ($\varrho = 0$) αλλά τα Δ.Ε. $I_{1,b}$, $-1 < b \leq 0$ βελτιώνουν, όπως αποδεικνύεται στο Θεώρημα 1 (ii).

Παρατήρηση 2..2. Το Δ.Ε. τύπου *Stein* μπορεί να παραχθεί από την κλάση που κατασκευάσαμε, αν $b = a - 1$ και $a \rightarrow +\infty$.

3. ΜΙΑ ΕΝΑΛΛΑΚΤΙΚΗ ΚΛΑΣΗ ΔΙΑΣΤΗΜΑΤΩΝ ΕΜΠΙΣΤΟΣΥΝΗΣ

Στη συνέχεια κατασκευάζουμε μία εναλλακτική κλάση βελτιωμένων Δ.Ε. για το σ στο μοντέλο της Lomax με τον ίδιο λόγο άκρων όπως το I_0 αλλά μεγαλύτερη πιθανότητα κάλυψης. Έστω $w > 0$ και $\phi_{a,b}^1$ τ.ω.,

$$\int_0^\infty c_1 \tau c_1 \xi(\tau) g_a \left(\frac{a\tau c_1}{\phi_{a,b}^1} \right) h_b \left(\frac{a\tau c_1}{\phi_{a,b}^1} \right) dG(\tau) = \int_0^\infty c_2 \tau c_2 \xi(\tau) g_a \left(\frac{a\tau c_2}{\phi_{a,b}^1} \right) h_b \left(\frac{a\tau c_2}{\phi_{a,b}^1} \right) dG(\tau) \quad (9)$$

όπου $g_a(\cdot)$ είναι η π.π. της *Gamma*($a(n-1), 1/n$), $h_b(\cdot)$ είναι η π.π. της *Gamma*($b+1, 1/n$), $a \geq 1$ και $b > -1$.

Από τη Σχέση (9), υπολογίζουμε ότι

$$\phi_{a,b}^1(w) = \frac{kc_2 - c_1}{1 - k}(a + w), \quad k = \left(\frac{c_1}{c_2} \right)^{\frac{a(n-1)+b+1}{a(n+\gamma-1)+b+1}}$$

Για $\phi_{a,b}^S(w) = \min\{1, \phi_{a,b}^1(w)\}$ προτείνουμε την παρακάτω κλάση Δ.Ε. για το σ ,

$$I_{a,b}^S = \begin{cases} \left(\frac{\phi_{a,b}^S(W)S}{c_2}, \frac{\phi_{a,b}^S(W)S}{c_1} \right) & , \quad W > 0 \\ I_0 & , \quad \text{διαφορετικά} \end{cases} \quad ; \quad a \geq 1, \quad -1 < b \leq 2(a-1)$$

και καταλήγουμε στο παρακάτω Θεώρημα,

Θεώρημα 2. Έστω $a \geq 1$ και $-1 < b \leq 2(a-1)$. Τότε $I_{a,b}^S$ βελτιώνει, ως προς την πιθανότητα κάλυψης το I_0 , δηλ. $P(\sigma \in I_{a,b}^S) > P(\sigma \in I_0)$ για κάθε $\varrho = \mu/\sigma \in \mathbb{R}$.

Παρατήρηση 3..1. Το Δ.Ε. $I_{1,0}^S$ συμπίπτει με το Δ.Ε. τύπου *Stein* I_S .

ABSTRACT

Two new classes of improved confidence intervals for the scale parameter σ of a multivariate Lomax distribution with unknown location parameter μ are constructed. The first one is a two-parameter class of smooth intervals $I(a, b)$, for $a \geq 1$ and b in a specified range,

which have the same ratio of endpoints as the minimum ratio of endpoints interval I_0 but greater coverage probability. Within this class, a subclass of generalized Bayes intervals is found which contains, in particular, the Brewster and Zidek-type interval I_{BZ} as a member. Another subclass of smooth intervals that improve the coverage probability for all parameter values is identified. The intervals of the second class, though non-smooth, have a very simple and explicit functional form. The Stein-type interval I_S is a member of this class.

ΒΙΒΛΙΟΓΡΑΦΙΑ

- Brewster, J. F.; Zidek, J. V. (1974). Improving on equivariant estimators. *Ann. Statist.*, **2**, 21–38.
- Bryson, M.C. (1974). Heavy-tailed distributions: properties and tests. *Technometrics*, **16**, 61–68.
- Kubokawa, T. (1994). A unified approach to improving equivariant estimators. *Ann. Statist.*, **22**, 290–299.
- Lomax, K.S. (1954). Business Failures: Another Example of the Analysis of Failure Data. *J. Amer. Statist. Assoc.*, **49**, 847 – 852.
- Maruyama, Y. (1998). Minimax estimators of a normal variance. *Metrika*, **48**, 209–214.
- Petropoulos, C. (2010). A class of improved estimators for the scale parameter of a mixture model of exponential distribution with unknown location. *Comm. in Stat. - Theory and Methods*, **39**, 3153-3162.
- Petropoulos, C. and Kourouklis, S. (2005). Estimation of the scale parameter in mixture models with unknown location. *J. of Stat. Plan. Inf.*, **128**, 191-218.
- Stein, C. (1964). Inadmissibility of the usual estimator for the variance of a normal distribution with unknown mean. *Ann. Inst. Statist. Math.*, **16**, 155–160.



ΣΥΓΚΡΙΣΗ ΥΠΟΚΕΙΜΕΝΙΚΩΝ ΑΠΑΝΤΗΣΕΩΝ ΣΕ ΕΡΕΥΝΕΣ ΚΟΙΝΗΣ ΓΝΩΜΗΣ ΚΑΙ ΤΡΟΠΟΙ ΑΝΤΙΜΕΤΩΠΗΣΗΣ

Σταύρος Πουπάκης
Πανεπιστήμιο Πειραιώς
spoupakis@hotmail.com

ΠΕΡΙΛΗΨΗ

Οι έρευνες γνώμης έχουν αποτελέσει αναμφισβήτητα δυναμικό μοχλό της έρευνας τόσο στον χώρο των κοινωνικών επιστημών, αλλά και σε πολλούς άλλους ακόμα. Σημαντικό μέρος αυτών αποτελούν οι ερωτήσεις αυτοαξιολόγησης για διάφορες καταστάσεις του ερωτώμενου. Η διαφορετικότητα όμως του κάθε ατόμου στο πως αντιλαμβάνεται την πραγματικότητα είναι κάτι που δημιουργεί πρόβλημα στον ερευνητή όταν μελετά έρευνες που προέρχονται από διαφορετικές κοινωνικοοικονομικές ομάδες και ιδιαίτερα από χώρες με διαφορετικά συστήματα διακυβέρνησης και οργάνωσης. Για να ξεπεραστεί αυτό το πρόβλημα έχουν αναπτυχθεί μέθοδοι διερεύνησης των προτύπων που έχει ο κάθε ερωτώμενος και έτσι γίνεται δυνατό να προσαρμοστεί η απάντησή του. Αυτό επιτυγχάνεται με την προσθήκη ερωτήσεων σχετικά με καταστάσεις υποθετικών ατόμων όπου λειτουργούν ως σημείο αναφοράς και έτσι καθίσταται δυνατή η σύγκριση των απαντήσεων των ερωτώμενων. Στόχος της παρούσας εργασίας είναι η αναφορά των μεθόδων που έχουν αναπτυχθεί για την επεξεργασία και μελέτη αυτών των περιπτώσεων. Ειδικότερα, θα παρουσιαστούν οι προσεγγίσεις που έχουν προταθεί, η μεθοδολογία που εφαρμόζεται, οι τρόποι με τους οποίους ξεπερνιούνται τα προβλήματα που δημιουργούνται, καθώς και η εφαρμογή τους.

Λέξεις Κλειδιά: βινιέτες, έρευνες γνώμης, υποκειμενικές ερωτήσεις, chopit, SHARE.

1. ΕΙΣΑΓΩΓΗ

Υποκειμενικές ερωτήσεις αυτοαξιολόγησης της προσωπικής κατάστασης των ερωτωμένων συναντάμε σχεδόν σε κάθε έρευνα γνώμης. Αυτό συμβαίνει διότι είναι ο πιο εύκολος, οικονομικός και πρακτικός τρόπος για να μελετήσουμε πτυχές της ζωής και της κοινωνίας που διαφορετικά θα ήταν πολύ δύσκολο, χωρίς αυτό να γίνεται εις βάρος της εγκυρότητας. Τέτοιες μπορεί να είναι θέματα που αφορούν την υγεία, την ικανοποίηση από την ζωή ή την δουλειά, την οικονομική ευημερία, κ.ά. Χαρακτηριστικά, ακόμα και για την περίπτωση της υγείας, οι αυτοαξιολογήσεις των ερωτώμενων της προσωπικής τους υγείας αποδεικνύονται πράγματι καλοί εκτιμητές της θνησιμότητας (Idler and Benyamini, 1997).

Το κύριο πρόβλημα όμως, όταν ζητάμε από τους ερωτώμενους να αξιολογήσουν τέτοιες καταστάσεις, είναι ότι «τα άτομα καταλαβαίνουν την ίδια ερώτηση με απέραντα διαφορετικούς τρόπους» (Brady, 1985). Έτσι ο ερευνητής μπορεί να οδηγηθεί εύκολα σε λανθασμένα συμπεράσματα, τα οποία να μην ανταποκρίνονται στην πραγματικότητα. Το κάθε άτομο έχει διαφορετικά πρότυπα και σημεία σύγκρισης όταν αξιολογεί διάφορες καταστάσεις από την προσωπική του ζωή. Τα πρότυπα ατόμων διαφορετικού φύλου, εκπαίδευσης, πολιτικού συστήματος, χώρας, ή κλίματος ενδέχεται να διαφέρουν σημαντικά και έτσι να είναι αδύνατη η σύγκριση μεταξύ αυτών. Ο τρόπος για να αντιμετωπιστεί αυτό το πρόβλημα είναι να εντοπιστούν τα πρότυπα αυτά και έτσι να προσαρμοστούν οι απαντήσεις του κάθε ερωτώμενου στα δικά του πρότυπα. Αυτό επιτυγχάνεται με την προσθήκη βινιετών (vignettes) στο ερωτηματολόγιο, δηλαδή κάποιων σύντομων περιγραφών υποθετικών ανθρώπων ή καταστάσεων (Alexander and Becker, 1978).

2. ΜΕΘΟΔΟΛΟΓΙΑ

2.1 Η προσέγγιση

Η προσθήκη προτυποποιημένων βινιετών στο ερωτηματολόγιο και η αξιολόγηση αυτών από τους ερωτώμενους χρησιμεύει στο να χρησιμοποιηθούν αυτές οι απαντήσεις ως άγκυρες (anchoring vignettes), οι οποίες θα προσαρμόσουν τις αυτοαξιολογήσεις και έτσι θα μπορούν να αποκωδικοποιηθούν οι απαντήσεις των ερωτώμενων. Για να είναι δυνατή η χρήση των προτυποποιημένων βινιετών υπάρχουν δύο πολύ σημαντικές υποθέσεις, οι οποίες πρέπει να ικανοποιούνται. Πρώτον, η συνέπεια απαντήσεων, δηλαδή το ότι οι ερωτώμενοι αξιολογούν την ερώτηση αυτοαξιολόγησης με τον ίδιο τρόπο που αξιολογούν και τις βινιέτες. Και δεύτερον, η ένταση της κατάστασης στην κάθε βινιέτα να είναι αντιληπτή από όλους τους ερωτώμενους με τον ίδιο τρόπο και στην ίδια κλίμακα, με ένα τυχαίο σφάλμα μέτρησης (King et al. 2004). Χωρίς όμως αυτό να σημαίνει ότι απαιτείται ο ερωτώμενος να δώσει μοναδικές απαντήσεις σε κάθε βινιέτα ή ότι όλοι οι ερωτώμενοι θα κατατάξουν τις βινιέτες με την ίδια σειρά (King and Wand, 2007).

Απαιτείται λοιπόν να γίνονται κατάλληλοι έλεγχοι αυτών των δύο υποθέσεων για τις προτυποποιημένες βινιέτες που χρησιμοποιούνται στην μέθοδο, είτε αυτοτελώς, είτε και με την χρήση αντικειμενικών μετρήσεων στο ίδιο δείγμα (Bago d' Uva et al. 2009, Van Soest et al. 2011). Επίσης, έχει αναπτυχθεί και μία μέθοδος, η οποία όμως απαιτεί η έρευνα να γίνει σε δύο στάδια. Κατά το πρώτο στάδιο ζητείται από τους ερωτώμενους να επιλέξουν ανάμεσα σε κάποιες περιγραφές, αυτές που χαρακτηρίζουν τους ίδιους. Έπειτα αυτές οι επιλογές συνδέονται ώστε να δημιουργηθεί μια βινιέτα, ξεχωριστή για κάθε ερωτώμενο, η οποία στην πραγματικότητα περιγράφει τον ίδιο τον ερωτώμενο. Έτσι, όταν στο δεύτερο στάδιο χρησιμοποιηθεί η μέθοδος των προτυποποιημένων βινιετών μπορεί να συμπεριληφθεί και η βινιέτα που κατασκευάστηκε, ώστε να γίνει σύγκριση μεταξύ της βινιέτας που περιγράφει τον ερωτώμενο με την αυτοαξιολόγηση που έχει δηλώσει (Kapteyn et al. 2011). Όσον αφορά την χρήση των προτυποποιημένων βινιετών στο ερωτηματολόγιο

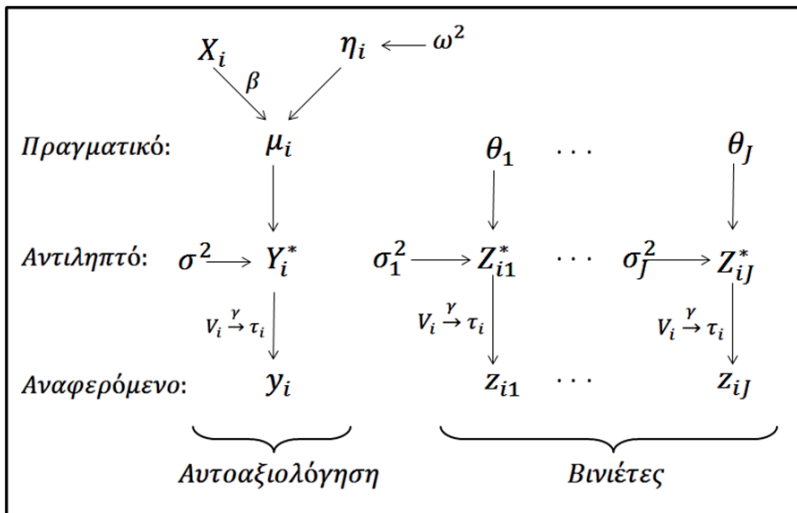
προτείνεται να χρησιμοποιούνται πρώτα οι βινιέτες και μετά οι αυτοαξιολογήσεις (Hopkins and King, 2010).

Η χρήση προτυποποιημένων βινιετών αποδεικνύεται – πέρα από σημαντική – απαραίτητη στην μεθοδολογία ερευνών που μελετούν και συγκρίνουν αποτελέσματα από διαφορετικές χώρες. Όμως, πέρα από την βελτίωση της συγκρισιμότητας μεταξύ χωρών, η μεθοδολογία των βινιετών μπορεί να χρησιμοποιηθεί ακόμα και σε μια συγκεκριμένη χώρα για να διορθωθούν κοινωνικοοικονομικές διαφορές στην δήλωση των ερωτώμενων (O'Donnell et al. 2009).

2.2 Το μοντέλο

Η αξιολόγηση των βινιετών στην επεξεργασία των αυτοαξιολογήσεων των ερωτώμενων γίνεται με την χρήση ενός μοντέλου, το οποίο αποτελεί γενίκευση του διατεταγμένου probit μοντέλου (King et al. 2004). Παρακάτω ακολουθεί μια γραφική αναπαράσταση της μεθοδολογίας, και στην συνέχεια γίνεται εκτενέστερη επεξήγηση.

Εικόνα 1. Γραφική σύνοψη του μοντέλου choirit



Ας θεωρήσουμε ότι ο ερωτώμενος i ($i=1, \dots, n$) αντιλαμβάνεται το πραγματικό επίπεδο της κατάστασης του Y_i^* σε μια συνεχή, μη παρατηρημένη κλίμακα, η οποία ακολουθεί την κανονική κατανομή με μέση τιμή μ_i και διακύμανση σ^2 . Όπου το μ_i είναι ένας γραμμικός συνδυασμός των ερμηνευτικών μεταβλητών X_i και των συντελεστών β , με μια τυχαία επίδραση η_i . Όπου η τυχαία επίδραση κατανέμεται κανονικά με μέση τιμή 0 και διακύμανση ω^2 . Έτσι λοιπόν, η αναφερόμενη κατάσταση του i ερωτώμενου y_i παίρνει τιμές ανάλογα με το ανάμεσα σε ποια σημεία τομής του διαστήματος $(-\infty, \infty)$ βρίσκεται η Y_i^* μέσω του παρακάτω μηχανισμού:

$$y_i = k \text{ αν } \tau_i^{k-1} \leq Y_i^* < \tau_i^k$$

Όπου,

$$\tau_i^0 = -\infty, \tau_i^K = \infty, \tau_i^{k-1} < \tau_i^k \text{ (για } k=1, \dots, K)$$

Αυτά τα σημεία τομής τ_i^k είναι που καθορίζουν σε ποια κατηγορία της κλίμακας 1-5 βρίσκεται ο κάθε ερωτώμενος, τα οποία μοντελοποιούνται ως γραμμικός συνδυασμός των ερμηνευτικών μεταβλητών V_i και των άγνωστων συντελεστών γ . Αυτές οι ερμηνευτικές μεταβλητές ορίζονται V_i , διότι δεν είναι απαραίτητο να είναι οι ίδιες μεταβλητές με τις μεταβλητές X_i που χρησιμοποιήθηκαν προηγουμένως.

$$\tau_i^1 = \gamma^1 V_i \quad (1)$$

$$\tau_i^k = \tau_i^{k-1} + \gamma^k V_i \text{ (για } k=2, \dots, K-1 \text{ και με } \gamma^k \geq 0) \quad (2)$$

Εδώ έρχονται να δώσουν πληροφορία οι προτυποποιημένες βινιέτες, τις οποίες επεξεργαζόμαστε με ανάλογο τρόπο, όπου θεωρούμε ότι η μη παρατηρημένη αντιληπτή αξιολόγηση της j βινιέτας ($j=1, \dots, J$) από τον i ερωτώμενο Z_{ij}^* ακολουθεί κανονική κατανομή με μέση τιμή θ_j και διακύμανση σ_j^2 . Μόνο που εδώ την μέση τιμή την χρησιμοποιούμε ως μια μοναδική τιμή, κοινή για όλους, θεωρώντας ότι η κατάσταση που περιγράφει μια βινιέτα αποτελεί ένα συγκεκριμένο επίπεδο κατάστασης ενός υποθετικού ατόμου. Παρόμοια λοιπόν η αξιολόγηση της κατάστασης που δηλώνει ο i ερωτώμενος z_{ij} παίρνει τιμές ανάλογα με το ανάμεσα σε ποια σημεία τομής του διαστήματος $(-\infty, \infty)$ βρίσκεται η αντιληπτή αξιολόγηση Z_{ij}^* , σύμφωνα με τον παρακάτω μηχανισμό:

$$z_{ij} = k \text{ αν } \tau_i^{k-1} \leq Z_{ij}^* < \tau_i^k$$

Όπου,

$$\tau_i^0 = -\infty, \tau_i^K = \infty, \tau_i^{k-1} < \tau_i^k \text{ (για } k=1, \dots, K)$$

Με τα σημεία τομής να μοντελοποιούνται ως γραμμικός συνδυασμός του διανύσματος μεταβλητών V_i και των συντελεστών γ .

$$\tau_i^1 = \gamma^1 V_i \quad (3)$$

$$\tau_i^k = \tau_i^{k-1} + \gamma^k V_i \text{ (για } k=2, \dots, K-1 \text{ και με } \gamma^k \geq 0) \quad (4)$$

Αυτά τα γ λοιπόν που βρίσκουμε από τις βινιέτες στις (3) και (4) τα χρησιμοποιούμε στην αυτοαξιολόγηση στις (1) και (2). Έτσι λοιπόν, τα κατώφλια στην αυτοαξιολόγηση, δηλαδή τα σημεία που θα καθορίσουν το σε ποια κατάσταση βρίσκεται ο i ερωτώμενος εξαρτώνται από τις αξιολογήσεις που έχει δώσει στις βινιέτες. Διορθώνεται δηλαδή, η αυτοπροσδιοριζόμενη κατάσταση του ερωτώμενου σύμφωνα με τα πρότυπα που θέτει ο ίδιος.

Ένα χαρακτηριστικό παράδειγμα της χρησιμότητας της μεθόδου αναφέρεται από τους King et al. (2004) οι οποίοι συγκρίνοντας την πολιτική αποτελεσματικότητα Μεξικανών και Κινέζων, βρίσκουν ότι οι Κινέζοι αυτό-τοποθετούνται σε υψηλότερο σημείο της κλίμακας πολιτικής αποτελεσματικότητας από τους Μεξικανούς. Για παράδειγμα πάνω από το 50% των Μεξικανών, αλλά λιγότερο από το 30% των

Κινέζων αναφέρουν ότι δεν έχουν κανένα λόγο στην κυβέρνηση. Όταν όμως χρησιμοποιείται ως βινιέτα η κατάσταση ενός υποθετικού ατόμου, που "... στερείται το καθαρό πόσιμο νερό. Θα ήθελε να το αλλάξει αυτό, αλλά δεν μπορεί να ψηφίσει και αισθάνεται ότι κανείς στην κυβέρνηση δεν νοιάζεται για αυτό το θέμα. Έτσι υποφέρει σιωπηλά, ελπίζοντας ότι κάτι θα γίνει στο μέλλον" διαπιστώνουν ότι το 40% των Κινέζων αλλά μόνο το 12% των Μεξικανών θεωρούν ότι έχουν λιγότερη πολιτική αποτελεσματικότητα από αυτό το υποθετικό άτομο.

3. ΕΦΑΡΜΟΓΗ ΣΕ ΔΕΔΟΜΕΝΑ

Παρακάτω παρουσιάζεται μια εφαρμογή του μοντέλου chopit των King et al. (2004) σε δεδομένα από την Έρευνα για την Υγεία, την Γήρανση και την Συνταξιοδότηση στην Ευρώπη (Survey of Health, Ageing and Retirement in Europe, SHARE). Τα δεδομένα προέρχονται από το δεύτερο κύμα (wave 2) του προγράμματος, τα οποία συλλέχθηκαν κατά τα έτη 2006-2007 και αποτελούν αντιπροσωπευτικό δείγμα του πληθυσμού άνω των 50 ετών σε 11 χώρες της Ευρώπης. Οι βινιέτες εντάχθηκαν στην μεθοδολογία του SHARE από το ξεκίνημα του προγράμματος και αποτελούν από τότε αναπόσπαστο μέρος της έρευνας (O'Donnell et al. 2009).

Η συγκεκριμένη εργασία επικεντρώνεται σε μία μόνο πτυχή από τις πολλές που προσφέρει η ύπαρξη προτυποποιημένων βινιέτων στο ερωτηματολόγιο του SHARE, και η οποία αφορά την πολιτική επιρροή στην τοπική αυτοδιοίκηση. Η συγκεκριμένη επιλογή έγινε επειδή αποτελέσματα αυτής δεν υπάρχουν στην βιβλιογραφία. Ενδεχομένως κάποιες άλλες πτυχές να παρουσίαζαν περισσότερο ενδιαφέρον ως προς τις διαφοροποιήσεις που θα δημιουργηθούν, αλλά όσον αφορά τον σκοπό της παρούσας εργασίας, ο οποίος είναι η εφαρμογή της μεθοδολογίας και η ανάδειξη της χρηστικότητας των βινιέτων, το οποιοδήποτε πεδίο μελέτης κρίνεται ικανό. Παρακάτω παρουσιάζονται αποτελέσματα από την κατασκευή δύο μοντέλων. Το πρώτο είναι ένα διατεταγμένο probit μοντέλο και το δεύτερο είναι το σύνθετο ιεραρχικά διατεταγμένο probit μοντέλο (chopit) παραθεμένα το ένα δίπλα στο άλλο για χάριν ευκολίας στην σύγκριση (King et al. 2004). Η ανάλυση των στοιχείων έγινε με το κατάλληλο πακέτο στη γλώσσα προγραμματισμού R (Wand et al. 2011). Για την κατασκευή και των δύο μοντέλων χρησιμοποιήθηκε το δείγμα του SHARE που περιλαμβάνει ερωτήσεις αυτοαξιολόγησης και βινιέτες (vignette sample), το οποίο είναι ένα υπό-δείγμα των 22717 νοικοκυριών που περιλαμβάνονται στο δεύτερο κύμα. Το δείγμα των βινιέτων λοιπόν, αποτελείται από 5336 νοικοκυριά, από τα οποία προέρχονται 7731 άτομα (για περισσότερα βλ. www.share-project.org).

Η εξαρτημένη μεταβλητή y_i είναι η απάντηση του ερωτώμενου στο πόση επιρροή πιστεύει ο ίδιος ότι έχει στην δημοτική αρχή του τόπου του σε μία κλίμακα από 1 έως 5, όπου αντιστοιχούν σε Καθόλου-Μικρή-Κάποια-Μεγάλη-Απεριόριστη. Για την ανάλυση χρησιμοποιήθηκαν ως ερμηνευτικές μεταβλητές οι ίδιες για X_i και V_i , ψευδομεταβλητές 0-1 για κάθε μία χώρα από τις 11 του δείγματος, το φύλο του ερωτώμενου, η ηλικία του σε έτη, και η εκπαίδευση του κωδικοποιημένη σύμφωνα με την ταξινόμηση International Standard Classification of Education (ISCED-97).

Προκειμένου να κατασκευαστεί το μοντέλο *chopit*, χρησιμοποιήθηκαν οι απαντήσεις των ερωτώμενων z_i σε δύο βινιέτες που αφορούν την πολιτική επιρροή, δηλαδή το πως βαθμολογούν οι ερωτώμενοι το επίπεδο πολιτικής επιρροής των δύο υποθετικών ατόμων που περιγράφονται στις βινιέτες στην ίδια κλίμακα από 1 έως 5, όπως ακριβώς στην ερώτηση αυτοαξιολόγησης. Οι περιγραφές, οι οποίες είναι κοινές για όλες τις χώρες, είναι οι εξής:

«Ο Θωμάς ενοχλείται από τους αέριους ρύπους, οι οποίοι προκαλούνται από μια τοπική επιχείρηση. Δεν είναι επιβλαβείς αλλά μερικές φορές δημιουργούν άσχημες οσμές. Ο δήμαρχος μελέτησε το πρόβλημα, αλλά κατέληξε στο ότι η βιομηχανική ανάπτυξη είναι η πιο σημαντική τακτική αυτή τη στιγμή, αντί του καθαρού αέρα.»

«Η Αλίκη ενοχλείται από τους αέριους ρύπους που προκαλεί μια τοπική επιχείρηση. Δεν είναι επιβλαβείς αλλά μερικές φορές δημιουργούν άσχημες οσμές. Αυτή και οι γείτονες της υποστηρίζουν έναν αντιπολιτευόμενο υποψήφιο στις προσεχείς τοπικές εκλογές ο οποίος υποσχέθηκε να αντιμετωπίσει το ζήτημα. Έτσι αυτή και πολλοί ακόμα άνθρωποι στην περιοχή της πιστεύουν ότι ο αντιπολιτευόμενος υποψήφιος θα νικήσει τον τωρινό δήμαρχο.»

Τα αποτελέσματα που προκύπτουν από την επεξεργασία των δεδομένων παρουσιάζονται στον παρακάτω πίνακα:

Πίνακας 1. Σύγκριση των μοντέλων: διατεταγμένο *probit* και *chopit*

Μεταβλητές	Διατεταγμένο <i>Probit</i>		<i>chopit</i>	
	Συντελεστής	(Τ.Σ.)	Συντελεστής	(Τ.Σ.)
Ισπανία	0,412	(0,061) ***	0,433	(0,071) ***
Ιταλία	0,092	(0,056)	0,253	(0,065) ***
Ελλάδα	0,366	(0,059) ***	0,169	(0,069) ***
Γαλλία	-0,010	(0,070)	-0,174	(0,081) **
Κάτω Χώρες	0,262	(0,059) ***	0,203	(0,070) ***
Βέλγιο	-0,107	(0,051) **	-0,107	(0,059) **
Δανία	0,071	(0,048)	0,121	(0,056) **
Σουηδία	0,028	(0,061)	0,042	(0,071)
Τσεχία	-0,148	(0,051) ***	0,083	(0,059) *
Πολωνία	-0,015	(0,060)	-0,006	(0,071)
Γυναίκα	-0,108	(0,026) ***	-0,088	(0,030) ***
Ηλικία	-0,008	(0,001) ***	-0,003	(0,001) ***
Εκπαίδευση	0,098	(0,013) ***	0,052	(0,015) ***
Κατώφλι 1	-0,607	(0,113)		
Κατώφλι 2	0,487	(0,113)		
Κατώφλι 3	1,331	(0,115)		
Κατώφλι 4	1,881	(0,118)		
N	7310		7310	
Λογ. Πιθανοφ.	-8680,52		-8554,90	

Σημείωση: Επίπεδο σημαντικότητας για * 10%, για ** 5%, και για ***1%.

Για την κατασκευή και των δύο μοντέλων ως βάση χρησιμοποιήθηκε το δείγμα που προέρχεται από την Γερμανία, γι' αυτό και δεν αναφέρεται στον πίνακα. Στα αριστερά του πίνακα φαίνονται οι συντελεστές και τα τυπικά σφάλματα της κάθε μεταβλητής που περιλαμβάνεται στο διατεταγμένο probit και τα τέσσερα κατώφλια, οι τιμές δηλαδή που χωρίζουν την επιρροή του ερωτώμενου στις πέντε κατηγορίες της απάντησης του. Στα δεξιά του πίνακα παρατίθενται οι συντελεστές των ίδιων μεταβλητών που χρησιμοποιήθηκαν για την κατασκευή του chorbit μοντέλου, χωρίς όμως να υπάρχουν κατώφλια. Εδώ τα κατώφλια μοντελοποιούνται και έτσι παράγουν διαφορετικά κατώφλια για κάθε ερωτώμενο, ανάλογα με τις απαντήσεις που έχει δώσει ο ερωτώμενος στις βινιέτες (ως ερμηνευτικές μεταβλητές των καταφλιών χρησιμοποιήθηκαν οι ίδιες με του μοντέλου). Έτσι οι συντελεστές των ερμηνευτικών μεταβλητών είναι διαφορετικοί από ότι στο διατεταγμένο probit. Αυτό επιτρέπει να δούμε τις επιρροές των μεταβλητών απαλλαγμένες από τα στερεότυπα που δεσμεύουν την ανάδειξη των πραγματικών συνεισφορών. Έτσι οι συντελεστές των χωρών μπορούν να αλλάξουν σημαντικά. Πράγμα που δείχνει ότι εάν είχε χρησιμοποιηθεί το πρώτο μοντέλο θα μπορούσαν να εξαχθούν λάθος συμπεράσματα. Αντιθέτως, το δεύτερο μοντέλο δίνει τους πραγματικούς συντελεστές κάθε χώρας, αφού συμπεριλαμβάνει και τις προτυποποιημένες βινιέτες στην κατασκευή του σύμφωνα με την μεθοδολογία που προαναφέρθηκε.

Οι συντελεστές με θετικό πρόσημο δείχνουν ποια χαρακτηριστικά αυξάνουν τον βαθμό πολιτικής επιρροής, ενώ αντίστοιχα αυτοί με αρνητικό πρόσημο τον μειώνουν. Από τα αστέρια φαίνεται για κάθε συντελεστή το αν είναι στατιστικά σημαντικός για τρία διαφορετικά επίπεδα σημαντικότητας. Τα διαφορετικά αποτελέσματα των δύο μοντέλων υποδεικνύουν την σημασία της χρήσης των βινιετών. Χαρακτηριστικά, κοιτάζοντας στο διατεταγμένο probit μοντέλο τον αρνητικό συντελεστή της Τσεχίας μαζί με το τυπικό σφάλμα κάποιος ερευνητής θα μπορούσε να καταλήξει για παράδειγμα στο συμπέρασμα ότι το να είναι κάποιος – άνω των 50 ετών – κάτοικος Τσεχίας μειώνει την πολιτική επιρροή που έχει σε σχέση με το να είναι από την Γερμανία. Εξετάζοντας όμως το chorbit μοντέλο φαίνεται ότι η κατάσταση είναι αντίστροφη, και έτσι θα προέκυπταν λάθος συμπεράσματα, τα οποία εν πρώτοις φαίνονταν σωστά.

Αν και η μεθοδολογία των προτυποποιημένων βινιετών έχει συμπεριληφθεί σε πολλές έρευνες και οργανισμούς παγκοσμίως (Παγκόσμια Τράπεζα, Παγκόσμιος Οργανισμός Υγείας, SHARE στην Ευρώπη, ELSA στην Αγγλία, HRS στις ΗΠΑ), ακόμη δεν είναι ευρέως γνωστή. Σε κάθε περίπτωση πάντως, οι βινιέτες είναι καλοδεχούμενες με όποιον τρόπο και αν χρησιμοποιούνται και σε όποιον βαθμό είναι δυνατόν να χρησιμοποιηθούν. Και αυτό γιατί έρχονται να δώσουν πληροφορία που διαφορετικά δεν θα μπορούσαμε να έχουμε, καθώς και να διορθώσουν αποτελέσματα τα οποία ενδεχομένως να οδηγούσαν σε λανθασμένα συμπεράσματα.

ABSTRACT

Undoubtedly, the surveys have been a powerful keystone for research in social science, as in many other fields. A significant part of them is self-assessment questions about several situations of respondent's life. Individuals' differentiation on understanding reality is

something that causes problems to the researcher who studies surveys from different socioeconomic groups, and particularly countries with different systems of governance and organization. To overcome this problem, methods have been developed which identify the standards of each respondent, and so it is possible to adjust their answers. This is achieved by the imputation of questions about situations of hypothetical persons, which are used as benchmarks and so it is possible to compare respondents' answers. The aim of this paper is to report the methods which have been developed for the elaboration and study of these cases. Especially, there will be a presentation of the proposed approaches, the implementation of methodology, ways to overcome any occurred problems, and their application.

Ευχαριστίες: Θα ήθελα να ευχαριστήσω τον καθηγητή μου Πλάτωνα Τήνιο για την αμέριστη υποστήριξη και καθοδήγηση που μου έδειξε καθ' όλη την διάρκεια της ενασχόλησης μου με την επεξεργασία των στοιχείων του SHARE. Καθώς επίσης, και τον Βενέδικτο Καπετανάκη για τις πολύτιμες και καιρίες συμβουλές του, δίχως των οποίων θα ήταν αδύνατη η συγγραφή της παρούσας εργασίας.

ΑΝΑΦΟΡΕΣ

- Alexander, C., and Becker, H. (1978). The Use of Vignettes in Survey Research. *Public Opinion Quarterly*, **42**, 93-104.
- Bago d'Uva, T., Lindeboom, M., O'Donnell, O., and Van Doorslaer, E. (2009). Slipping Anchor: Testing the vignettes approach to identification and correction of reporting heterogeneity. *Netspar Discussion Paper*, No. 11/2009-038.
- Brady, H. (1985). The perils of survey research: Interpersonally incomparable responses. *Political Methodology*, **11**, 269-291.
- Hopkins, D., and King, G. (2010). Improving Anchoring Vignettes: Designing Surveys to Correct Interpersonal Incomparability. *Public Opinion Quarterly*, **74**, 201-222.
- Idler, E., and Benyamini, Y. (1997). Self-rated Health and Mortality: A Review of Twenty-seven Community Studies. *Journal of Health and Social Behavior*, **38**, 21-37.
- Kapteyn, A., Smith, J., Van Soest, A., and Vonkova, H. (2011). Anchoring vignettes and response consistency. *RAND Working Paper Series*, WR-840.
- King, G., Murray, C., Salomon, J., and Tandon, A. (2004). Enhancing the validity and cross-cultural comparability of measurement in survey research. *American Political Science Review*, **98**, 191-207.
- King, G., and Wand, J. (2007). Comparing incomparable survey responses: New tools for anchoring vignettes. *Political Analysis*, **15**, 46-66.
- O'Donnell, O., Bago d'Uva, T., και Δεληπάλλα, Σ. (2009). Βελτίωση συγκρισιμότητας στοιχείων βασισμένων σε υποκειμενικές απαντήσεις χρησιμοποιώντας προτυποποιημένες vignettes. Στο Λυμπεράκη Α., Πλάτων Τ., Φιλαλήθης Τ. (επιμ.), *Ζωή 50+*. Αθήνα: Εκδόσεις Κριτική, Κεφ. 3, 83-103.
- Van Soest, A., Delaney, L., Harmon, C., Kapteyn, A., and Smith, J. (2011). Validating the use of vignettes for subjective threshold scales. *Journal of the Royal Statistical Society A*, forthcoming.
- Wand, J., King, G., and Lau, O. (2011). anchors: Software for anchoring vignette data. *Journal of Statistical Software*, forthcoming.



ΤΕΣΤ ΕΚΚΙΝΗΣΗΣ ΑΠΟΔΟΧΗΣ ΕΞΑΡΤΗΜΑΤΩΝ ΜΕ ΔΥΟ ΕΠΙΠΕΔΑ ΑΠΟΤΥΧΙΑΣ

A. X. Ρακιτζής¹, Δ. Α. Αντζουλάκος²

¹Τμήμα Στατιστικής και Αναλ.-Χρημ. Μαθηματικών, Πανεπιστήμιο Αιγαίου
arakitz@aegean.gr

²Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης, Πανεπιστήμιο Πειραιά
dantz@unipi.gr

ΠΕΡΙΛΗΨΗ

Τα τεστ εκκίνησης αποτελούν διαδικασίες του ποιοτικού ελέγχου που μας βοηθούν να αποφασίσουμε για την ικανότητα μιας συσκευής (ή ενός μηχανήματος) να τεθεί σε λειτουργία. Τα τεστ εκκίνησης είναι δίτιμα (*binary*), με την έννοια ότι είναι δυνατά μόνο δύο αποτελέσματα: επιτυχημένη ή αποτυχημένη εκκίνηση. Στην παρούσα εργασία προτείνονται και μελετώνται τεστ εκκίνησης, όπου σε κάθε δοκιμή εκκίνησης της συσκευής είναι δυνατή η εμφάνιση τριών αποτελεσμάτων: επιτυχημένη εκκίνηση (S), αποτυχημένη εκκίνηση πρώτου είδους (F_1) και αποτυχημένη εκκίνηση δεύτερου είδους (F_2). Οι κανόνες απόφασης για την αποδοχή μιας συσκευής βασίζονται στην εμφάνιση συνεχόμενων επιτυχιών (*consecutive S*) και συνεχόμενων / συνολικών αποτυχιών πρώτου και δεύτερου είδους (*consecutive / total F_1, F_2*). Η μελέτη των νέων τεστ εκκίνησης γίνεται χρησιμοποιώντας θεωρία Μαρκοβιανών αλυσίδων.

Λέξεις Κλειδιά: Τεστ εκκίνησης, Κατανομή χρόνου αναμονής, Μαρκοβιανές αλυσίδες.

1. ΕΙΣΑΓΩΓΗ

Ένα τεστ εκκίνησης είναι μια διαδικασία η οποία εξετάζει την αξιοπιστία μιας συσκευής σε σχέση με την ικανότητά της να τεθεί σε λειτουργία. Οι Hahn and Gage (1983) εισήγαγαν και μελέτησαν αυτή τη διαδικασία χρησιμοποιώντας ως κανόνα αποδοχής της συσκευής την εμφάνιση k συνεχόμενων επιτυχημένων εκκινήσεων της. Θα αναφερόμαστε σε αυτό το τεστ εκκίνησης ως CS (*consecutive successes*) τεστ. Για τη μελέτη του μοντέλου αυτού υπέθεσαν ότι οι διαδοχικές εκκινήσεις λειτουργίας της συσκευής αποτελούν μια ακολουθία ανεξάρτητων και ισόνομων δοκιμών Bernoulli με πιθανότητα επιτυχίας p . Η κατανομή του αριθμού των εκκινήσεων μέχρι τον τερματισμό του CS τεστ είναι γνωστή στη βιβλιογραφία ως γεωμετρική κατανομή τάξης k (Philippou et al. (1983)). Οι Viveros and Balakrishnan (1993) μελέτησαν περαιτέρω CS τεστ και έδωσαν εκτιμήσεις για την άγνωστη πιθανότητα επιτυχούς εκκίνησης p . Πλήθος επεκτάσεων για το CS τεστ έχουν καταγραφεί στη βιβλιογραφία και αφορούν (i) τεστ με σύνθετους κανόνες αποδοχής / απόρριψης

(Balakrishnan and Chan (2000), Smith and Griffith (2005, 2008), Antzoulakos et al. (2009)), (ii) τεστ με Μαρκοβιανή εξάρτηση μεταξύ των διαδοχικών εκκινήσεων του εξοπλισμού (Balakrishnan et al. (1995, 1997), Martin (2004)), (iii) μεθόδους εκτίμησης της άγνωστης πιθανότητας p για την επιτυχημένη εκκίνηση του εξοπλισμού (Chan et al. (2008), Scollnik (2010, 2011)).

Όλα τα παραπάνω τεστ εκκίνησης έχουν το κοινό χαρακτηριστικό ότι σε κάθε εκκίνηση υπάρχουν δύο δυνατά αποτελέσματα, επιτυχημένη ή αποτυχημένη εκκίνηση. Όμως σε πρακτικές εφαρμογές (δείτε π.χ. Koutras and Alexandrou (1997), Balakrishnan et al. (2009)), οι διαδοχικές εκκινήσεις ενός εξοπλισμού μπορούν να περιγραφούν από ένα μοντέλο στο οποίο υπάρχουν περισσότερες από δύο πιθανές καταστάσεις. Για παράδειγμα, έστω ότι μια εξωλέμβια μηχανή θαλάσσης, υπόκειται σε διαδοχικές εκκινήσεις. Σε κάθε εκκίνηση, υπάρχουν τρία δυνατά ενδεχόμενα (i) να μην ξεκινήσει η λειτουργία της, (ii) να έχει περιορισμένη χρονική περίοδο λειτουργίας (π.χ. κάποια δευτερόλεπτα) πριν σβήσει και (iii) να ξεκινήσει χωρίς πρόβλημα τη λειτουργία της. Στην περίπτωση αυτή, ένα πιο ακριβές μοντέλο περιγραφής της αξιοπιστίας της μηχανής απαιτεί τη χρησιμοποίηση τριών καταστάσεων αντί για δύο.

Στην παρούσα εργασία εισάγουμε και μελετάμε τρίτιμα (*trinary*) τεστ εκκίνησης υποθέτοντας ότι κάθε εκκίνηση του εξοπλισμού μπορεί να είναι επιτυχημένη (S), αποτυχημένη τύπου 1 (F_1) ή αποτυχημένη τύπου 2 (F_2). Τα τεστ που προτείνονται είναι τα εξής:

- $CSCF_1CF_2$: Ο εξοπλισμός απορρίπτεται εάν εμφανιστούν r_1 συνεχόμενες αποτυχημένες εκκινήσεις τύπου 1 ή εάν εμφανιστούν r_2 συνεχόμενες αποτυχημένες εκκινήσεις τύπου 2, πριν από την εμφάνιση k συνεχόμενων επιτυχημένων εκκινήσεων. Διαφορετικά, ο εξοπλισμός γίνεται αποδεκτός.
- $CSCF_1TF_2$: Ο εξοπλισμός απορρίπτεται εάν εμφανιστούν r_1 συνεχόμενες αποτυχημένες εκκινήσεις τύπου 1 ή εάν εμφανιστούν συνολικά r_2 αποτυχημένες εκκινήσεις τύπου 2, πριν από την εμφάνιση k συνεχόμενων επιτυχημένων εκκινήσεων. Διαφορετικά, ο εξοπλισμός γίνεται αποδεκτός.
- $CSTF_1TF_2$: Ο εξοπλισμός απορρίπτεται εάν εμφανιστούν συνολικά r_1 αποτυχημένες εκκινήσεις τύπου 1 ή εάν εμφανιστούν συνολικά r_2 αποτυχημένες εκκινήσεις τύπου 2, πριν από την εμφάνιση k συνεχόμενων επιτυχημένων εκκινήσεων. Διαφορετικά, ο εξοπλισμός γίνεται αποδεκτός.

Στην Ενότητα 2 εισάγονται συμβολισμοί και οι αναγκαίες ποσότητες που περιγράφουν την απόδοση ενός τρίτιμου τεστ εκκίνησης ενώ στην Ενότητα 3 δίνονται τα βασικά θεωρητικά αποτελέσματα για τη μελέτη των προτεινόμενων τεστ. Στην Ενότητα 4 δίνονται οι Μαρκοβιανές αλυσίδες και οι πιθανότητες μετάβασης, οι οποίες περιγράφουν τη λειτουργία των τεστ ενώ στην Ενότητα 5 δίδονται ενδεικτικά αριθμητικά αποτελέσματα.

2. ΟΡΙΣΜΟΙ ΚΑΙ ΣΥΜΒΟΛΙΣΜΟΙ

Έστω $\{Z_n, n \geq 1\}$ ακολουθία ανεξάρτητων και ισόνομων τρίτιμων δοκιμών με πιθανότητα επιτυχίας (επιτυχημένης εκκίνησης) ίση με $p = \Pr(Z_n = 0)$ και πιθανότητα αποτυχίας (αποτυχημένης εκκίνησης) τύπου j ίση με $p_j = \Pr(Z_n = j), j = 1, 2$. Έστω E_0 το ενδεχόμενο που αντιστοιχεί στην εμφάνιση μιας ροής επιτυχιών μήκους k , E_{jC} ($j = 1, 2$) το ενδεχόμενο που αντιστοιχεί στην εμφάνιση μιας ροής αποτυχίας τύπου j μήκους r_1 και E_{jT} το ενδεχόμενο που αντιστοιχεί στην εμφάνιση r_2 συνολικά αποτυχιών τύπου j . Έστω Y η τυχαία μεταβλητή που συμβολίζει το χρόνο αναμονής μέχρι την εμφάνιση του E_0 , του E_{1U} ή του E_{2V} , οποιοδήποτε από τα δύο εμφανιστεί νωρίτερα ($U, V \in \{C, T\}$). Για το χρόνο αναμονής Y εισάγουμε τους ακόλουθους συμβολισμούς

$$f_Y(n) = P(Y = n), \quad S_Y(n) = P(Y > n), \quad F_Y(n) = P(Y \leq n),$$

$$G_Y(s) = \sum_{n=0}^{\infty} f_Y(n)s^n, \quad H_Y(s) = \sum_{n=0}^{\infty} S_Y(n)s^n.$$

Η πιθανότητα λήψης απόφασης για την τύχη της συσκευής στη n -οστή δοκιμή είναι ίση με $f_Y(n)$.

Έστω Y_0, Y_{1U} και Y_{2V} οι χρόνοι αναμονής μέχρι την εμφάνιση για πρώτη φορά του ενδεχομένου $E = E_0 \cup E_{1U} \cup E_{2V}$ λόγω της εμφάνισης του E_0 , του E_{1U} και του E_{2V} , αντίστοιχα. Για το χρόνο αναμονής $Y_D, D = 0, 1U, 2V$, θα χρησιμοποιήσουμε ανάλογους συμβολισμούς με αυτούς για το χρόνο αναμονής Y , δηλαδή

$$f_{Y_D}(n) = P(Y_D = n), \quad S_{Y_D}(n) = P(Y_D > n), \quad F_{Y_D}(n) = P(Y_D \leq n),$$

$$G_{Y_D}(s) = \sum_{n=0}^{\infty} f_{Y_D}(n)s^n, \quad H_{Y_D}(s) = \sum_{n=0}^{\infty} S_{Y_D}(n)s^n.$$

Μπορούμε να διαπιστώσουμε ότι

$$f_Y(n) = f_{Y_0}(n) + f_{Y_{1U}}(n) + f_{Y_{2V}}(n) \text{ και } G_Y(s) = G_{Y_0}(s) + G_{Y_{1U}}(s) + G_{Y_{2V}}(s).$$

Η πιθανότητα αποδοχής (απόρριψης) της συσκευής στη n -οστή δοκιμή είναι ίση με $f_{Y_0}(n)$ ενώ η πιθανότητα αποδοχής P_A της συσκευής είναι ίση με $P_A = G_{Y_0}(1)$. Όμοια, διαπιστώνουμε ότι η πιθανότητα απόρριψης P_R της συσκευής είναι ίση με

$$P_R = G_{Y_{1U}}(1) + G_{Y_{2V}}(1) = P_{1U} + P_{2V} = 1 - P_A$$

όπου P_{1U}, P_{2V} είναι οι πιθανότητες απόρριψης του εξοπλισμού λόγω εμφάνισης των ενδεχομένων E_{1U} και του E_{2V} , αντίστοιχα.

3. ΘΕΩΡΗΤΙΚΑ ΑΠΟΤΕΛΕΣΜΑΤΑ

Στην παρούσα παράγραφο διεξάγεται μια συστηματική μελέτη του χρόνου αναμονής Y χρησιμοποιώντας τη μεθοδολογία εμφύτευσης σε πεπερασμένη Μαρκοβιανή αλυσίδα παρόμοια με αυτή των Smith and Griffith (2008) και Antzoulakos et al. (2009).

Ορίζουμε μια Μαρκοβιανή αλυσίδα $\{X_t, t \geq 1\}$ με πεπερασμένο χώρο καταστάσεων $\Omega = T \cup A$, όπου $T = \{1, 2, \dots, t\}$ και $A = \{t+1, t+2, \dots, t+a\}$ είναι τα σύνολα των μεταβατικών και απορροφητικών καταστάσεων, αντίστοιχα. Ο πίνακας πιθανοτήτων μετάβασης \mathbf{P} γράφεται στη μορφή

$$\mathbf{P} = \begin{bmatrix} \mathbf{Q} & \mathbf{R} \\ \mathbf{0} & \mathbf{I}_a \end{bmatrix}_{(t+a) \times (t+a)},$$

όπου \mathbf{Q} είναι ο $t \times t$ πίνακας πιθανοτήτων μετάβασης μεταξύ των μεταβατικών καταστάσεων της αλυσίδας, \mathbf{R} είναι ο $t \times a$ πίνακας πιθανοτήτων μετάβασης μεταξύ των μεταβατικών καταστάσεων της αλυσίδας, $\mathbf{0}$ είναι ένας $a \times t$ μηδενικός πίνακας και ως \mathbf{I}_n συμβολίζουμε τον $n \times n$ μοναδιαίο πίνακα. Το διάνυσμα των αρχικών πιθανοτήτων είναι

$$\boldsymbol{\pi} = (P(X_0 = 1), P(X_0 = 2), \dots, P(X_0 = t+a)) = (1, 0, 0, \dots, 0).$$

Για το χρόνο αναμονής Y έχουμε τα ακόλουθα αποτελέσματα (δείτε π.χ. Antzoulakos et al. (2009), Smith and Griffith (2008))

$$F_Y(n) = \mathbf{e}_1(\mathbf{I}_t - \mathbf{Q}^n)(\mathbf{I}_t - \mathbf{Q})^{-1}\mathbf{R}\mathbf{1}'_a, \quad n \geq 1, \quad S_Y(n) = \mathbf{e}_1\mathbf{Q}^n(\mathbf{I}_t - \mathbf{Q})^{-1}\mathbf{R}\mathbf{1}'_a, \quad n \geq 0, \\ f_Y(n) = \mathbf{e}_1\mathbf{Q}^{n-1}\mathbf{R}\mathbf{1}'_a, \quad n \geq 1$$

όπου \mathbf{e}_i ($i = 1, 2, \dots, t$) είναι το μοναδιαίο διάνυσμα του \mathbb{R}^t με 1 στην i -οστή συνιστώσα και 0 διαφορετικά, και $\mathbf{1}_n$ είναι το $1 \times n$ διάνυσμα με 1 σε όλες τις συνιστώσες του.

Οι γεννήτριες συναρτήσεις $G_Y(s)$, $H_Y(s)$ και οι καθοδικές παραγοντικές ροπές m -τάξης $\mu'_{[m]} = E[T(T-1)\dots(T-m+1)]$ δίνονται από τις σχέσεις

$$H_Y(s) = \mathbf{e}_1(\mathbf{I}_t - s\mathbf{Q})^{-1}(\mathbf{I}_t - \mathbf{Q})^{-1}\mathbf{R}\mathbf{1}'_a, \quad G_Y(s) = s\mathbf{e}_1(\mathbf{I}_t - s\mathbf{Q})^{-1}\mathbf{R}\mathbf{1}'_a \\ \mu'_{[m]} = \begin{cases} \mathbf{e}_1(\mathbf{I}_t - \mathbf{Q})^{-2}\mathbf{R}\mathbf{1}'_a, & m = 1 \\ m!\mathbf{e}_1(\mathbf{I}_t - \mathbf{Q})^{-m}\mathbf{Q}^{m-1}(\mathbf{I}_t - \mathbf{Q})^{-1}\mathbf{R}\mathbf{1}'_a, & m \geq 2. \end{cases}$$

Ειδικότερα,

$$\mu'_{[1]} = E(Y) = \mathbf{e}_1(\mathbf{I}_t - \mathbf{Q})^{-2}\mathbf{R}\mathbf{1}'_a, \quad V(Y) = \mu'_{[2]} + \mu'_{[1]} - (\mu'_{[1]})^2, \\ \mu'_{[2]} = E(Y(Y-1)) = 2\mathbf{e}_1(\mathbf{I}_t - \mathbf{Q})^{-2}\mathbf{Q}(\mathbf{I}_t - \mathbf{Q})^{-1}\mathbf{R}\mathbf{1}'_a.$$

Η μελέτη του χρόνου αναμονής Y_D γίνεται αντίστοιχα με αυτή του Y με κατάλληλες τροποποιήσεις στα στοιχεία του πίνακα πιθανοτήτων μετάβασης \mathbf{P} . Γενικά, μπορεί να διαπιστωθεί ότι ο πίνακας πιθανοτήτων μετάβασης για τη μελέτη του χρόνου αναμονής Y_B , με $B \subseteq A$, ο οποίος εκφράζει το χρόνο αναμονής μέχρι να εισέλθει για πρώτη φορά η αλυσίδα στο υποσύνολο B των απορροφητικών καταστάσεων είναι ο

$$\mathbf{P}_B = \begin{bmatrix} \mathbf{Q} & \mathbf{R}_B \\ \mathbf{0} & \mathbf{I}_a \end{bmatrix}_{(t+a) \times (t+a)}$$

όπου ο \mathbf{R}_B προκύπτει από τον \mathbf{R} , αντικαθιστώντας με 0 τα μη-μηδενικά στοιχεία των στηλών οι οποίες αντιστοιχούν στις απορροφητικές καταστάσεις του συνόλου $A-B$. Αντικαθιστώντας τον \mathbf{R} με τον \mathbf{R}_B στους παραπάνω τύπους είναι δυνατή η μελέτη του χρόνου αναμονής Y_D .

Τέλος, για τον αριθμό W_B των εκκινήσεων μέχρι το τέλος του τεστ δεδομένου ότι η αλυσίδα εισήλθε στο σύνολο των απορροφητικών καταστάσεων B έχουμε

$$P(W_B = n) = (P_B)^{-1} P(Y_B = n), \quad n \geq 1$$

όπου $P_B = G_{Y_B}(1) = \mathbf{e}_1(\mathbf{I}_t - \mathbf{Q})^{-1} \mathbf{R}_B \mathbf{1}'_a$.

4. ΜΑΡΚΟΒΙΑΝΗ ΑΝΑΛΥΣΗ ΤΩΝ ΠΡΟΤΕΙΝΟΜΕΝΩΝ ΕΛΕΓΧΩΝ

Στην παρούσα παράγραφο δίνονται οι Μαρκοβιανές αλυσίδες για τη μελέτη των τεστ εκκίνησης $CSCF_1CF_2$, $CSCF_1TF_2$ και $CSTF_1TF_2$.

4.1 Το Τεστ Εκκίνησης $CSCF_1CF_2$

Για το τεστ εκκίνησης $CSCF_1CF_2$ ορίζουμε τη Μαρκοβιανή αλυσίδα $\{X_t, t \geq 1\}$ με χώρο καταστάσεων

$$\Omega = \{(0, 0, 0)\} \cup \{(x, 0, 0) : 1 \leq x \leq k-1\} \\ \cup \{(0, y, 0) : 1 \leq y \leq r_1-1\} \cup \{(0, 0, z) : 1 \leq z \leq r_2-1\} \cup \{a_0, a_1, a_2\},$$

όπου $X_n = (x, y, z)$ σημαίνει ότι μετά τη n -οστή εκκίνηση υπάρχουν x συνεχόμενες επιτυχημένες εκκινήσεις S , y συνεχόμενες αποτυχημένες εκκινήσεις F_1 και z συνεχόμενες αποτυχημένες εκκινήσεις F_2 . Η κατάσταση $(0, 0, 0)$ είναι η αρχική κατάσταση της αλυσίδας ενώ οι καταστάσεις a_0, a_1 και a_2 είναι οι απορροφητικές καταστάσεις και αντιστοιχούν στην εμφάνιση των ενδεχομένων E_0, E_{1C}, E_{2C} . Οι πιθανότητες μετάβασης είναι

$$P(X_n = (x+1, 0, 0) | X_{n-1} = (x, y, z)) = p, \quad P(X_n = (0, y+1, 0) | X_{n-1} = (x, y, z)) = q_1, \\ P(X_n = (0, 0, z+1) | X_{n-1} = (x, y, z)) = q_2.$$

4.2 Το Τεστ Εκκίνησης $CSCF_1TF_2$

Για το τεστ εκκίνησης $CSCF_1TF_2$ ορίζουμε τη Μαρκοβιανή αλυσίδα $\{X_t, t \geq 1\}$ με χώρο καταστάσεων

$$\Omega = \{(0, 0, 0)\} \cup \{(x, 0, z) : 1 \leq x \leq k-1, 0 \leq z \leq r_2-1\} \\ \cup \{(0, y, z) : 1 \leq y \leq r_1-1, 0 \leq z \leq r_2-1\} \cup \{(0, 0, z) : 1 \leq z \leq r_2-1\} \cup \{a_0, a_1, a_2\},$$

όπου $X_n = (x, y, z)$ σημαίνει ότι μετά τη n -οστή εκκίνηση υπάρχουν x συνεχόμενες επιτυχημένες εκκινήσεις S , y συνεχόμενες αποτυχημένες εκκινήσεις F_1 και συνολικά z αποτυχημένες εκκινήσεις F_2 . Η κατάσταση $(0, 0, 0)$ είναι η αρχική κατάσταση της αλυσίδας ενώ οι καταστάσεις a_0, a_1 και a_2 είναι οι απορροφητικές καταστάσεις και

αντιστοιχούν στην εμφάνιση των ενδεχομένων E_0, E_{1C}, E_{2T} . Οι πιθανότητες μετάβασης είναι

$$P(X_n = (x+1, 0, z) | X_{n-1} = (x, y, z)) = p, \quad P(X_n = (0, y+1, z) | X_{n-1} = (x, y, z)) = q_1, \\ P(X_n = (0, 0, z+1) | X_{n-1} = (x, y, z)) = q_2.$$

4.3 Το Τεστ Εκκίνησης $CSTF_1TF_2$

Για το τεστ εκκίνησης $CSTF_1TF_2$ ορίζουμε τη Μαρκοβιανή αλυσίδα $\{X_t, t \geq 1\}$ με χόρο καταστάσεων

$$\Omega = \{(x, y, z) : 0 \leq x \leq k-1, 0 \leq y \leq r_1-1, 0 \leq z \leq r_2-1\} \cup \{a_0, a_1, a_2\},$$

όπου $X_n = (x, y, z)$ σημαίνει ότι μετά τη n -οστή εκκίνηση υπάρχουν x συνεχόμενες επιτυχημένες εκκινήσεις S , συνολικά y αποτυχημένες εκκινήσεις F_1 και συνολικά z αποτυχημένες εκκινήσεις F_2 . Η κατάσταση $(0, 0, 0)$ είναι η αρχική κατάσταση της αλυσίδας ενώ οι καταστάσεις a_0, a_1 και a_2 είναι οι απορροφητικές καταστάσεις και αντιστοιχούν στην εμφάνιση των ενδεχομένων E_0, E_{1T}, E_{2T} . Οι πιθανότητες μετάβασης είναι

$$P(X_n = (x+1, y, z) | X_{n-1} = (x, y, z)) = p, \quad P(X_n = (0, y+1, z) | X_{n-1} = (x, y, z)) = q_1, \\ P(X_n = (0, y, z+1) | X_{n-1} = (x, y, z)) = q_2.$$

Από τον τρόπο με τον οποίο ορίστηκαν οι παραπάνω ομογενείς Μαρκοβιανές αλυσίδες, μπορούμε να βρούμε τον πίνακα πιθανοτήτων μετάβασης \mathbf{P} και χρησιμοποιώντας τα αποτελέσματα της Ενότητας 2 είναι εφικτή η μελέτη των προτεινόμενων τεστ εκκίνησης.

5. ΑΡΙΘΜΗΤΙΚΑ ΠΑΡΑΔΕΙΓΜΑΤΑ

Στην παρούσα παράγραφο δίνονται αριθμητικά παραδείγματα σχετικά με τα μέτρα απόδοσης (πιθανότητα αποδοχής P_A , μέση τιμή $E(Y)$ και διασπορά $V(Y)$ του αριθμού των εκκινήσεων) των προτεινόμενων τεστ εκκίνησης.

Πίνακας 1: $P_A, E(Y)$ και $V(Y)$ για το τεστ $CSCF_1CF_2$

Test	$k = 8, r_1 = 5, r_2 = 2$			$k = 8, r_1 = 4, r_2 = 2$			$k = 8, r_1 = 3, r_2 = 2$		
$CSCF_1CF_2$	$p=0.80$ $q_1=0.15$ $q_2=0.05$	$p=0.80$ $q_1=0.10$ $q_2=0.10$	$p=0.80$ $q_1=0.05$ $q_2=0.15$	$p=0.80$ $q_1=0.15$ $q_2=0.05$	$p=0.80$ $q_1=0.10$ $q_2=0.10$	$p=0.80$ $q_1=0.05$ $q_2=0.15$	$p=0.80$ $q_1=0.15$ $q_2=0.05$	$p=0.80$ $q_1=0.10$ $q_2=0.10$	$p=0.80$ $q_1=0.05$ $q_2=0.15$
$E(Y)$	23.38	20.24	16.70	23.19	20.20	16.70	21.34	19.88	16.67
$V(Y)$	17.81	15.17	12.20	17.63	15.14	12.20	16.53	14.84	12.17
P_A	0.943	0.816	0.673	0.935	0.815	0.673	0.885	0.801	0.672

Πίνακας 2: $P_A, E(Y)$ και $V(Y)$ για το τεστ $CSCF_1TF_2$

Test	$k = 8, r_1 = 5, r_2 = 2$			$k = 8, r_1 = 4, r_2 = 2$			$k = 8, r_1 = 3, r_2 = 2$		
$CSCF_1TF_2$	$p=0.80$ $q_1=0.15$ $q_2=0.05$	$p=0.80$ $q_1=0.10$ $q_2=0.10$	$p=0.80$ $q_1=0.05$ $q_2=0.15$	$p=0.80$ $q_1=0.15$ $q_2=0.05$	$p=0.80$ $q_1=0.10$ $q_2=0.10$	$p=0.80$ $q_1=0.05$ $q_2=0.15$	$p=0.80$ $q_1=0.15$ $q_2=0.05$	$p=0.80$ $q_1=0.10$ $q_2=0.10$	$p=0.80$ $q_1=0.05$ $q_2=0.15$
$E(Y)$	17.18	12.20	9.40	17.09	12.19	9.40	16.49	12.10	9.39
$V(Y)$	10.35	6.27	4.26	10.28	6.26	4.26	9.89	6.20	4.25
P_A	0.693	0.492	0.379	0.689	0.492	0.379	0.665	0.488	0.378

Πίνακας 3: P_A , $E(Y)$ και $V(Y)$ για το τεστ $CSTF_1TF_2$

Test	$k = 8, r_1 = 5, r_2 = 2$			$k = 8, r_1 = 4, r_2 = 2$			$k = 8, r_1 = 3, r_2 = 2$		
$CSTF_1TF_2$	$p = 0.80$	$p = 0.80$	$p = 0.80$	$p = 0.80$	$p = 0.80$	$p = 0.80$	$p = 0.80$	$p = 0.80$	$p = 0.80$
	$q_1 = 0.15$	$q_1 = 0.10$	$q_1 = 0.05$	$q_1 = 0.15$	$q_1 = 0.10$	$q_1 = 0.05$	$q_1 = 0.15$	$q_1 = 0.10$	$q_1 = 0.05$
	$q_2 = 0.05$	$q_2 = 0.10$	$q_2 = 0.15$	$q_2 = 0.05$	$q_2 = 0.10$	$q_2 = 0.15$	$q_2 = 0.05$	$q_2 = 0.10$	$q_2 = 0.15$
$E(Y)$	14.48	11.87	9.39	13.19	11.48	9.35	11.34	10.69	9.22
$V(Y)$	6.11	5.52	4.22	5.12	5.01	4.15	4.19	4.31	3.95
P_A	0.584	0.479	0.378	0.532	0.463	0.377	0.457	0.431	0.372

Από τους Πίνακες 1-3 παρατηρούμε ότι για δεδομένες τιμές των k , r_1 και r_2 το τεστ $CSCF_1CF_2$ έχει το μεγαλύτερο αναμενόμενο αριθμό εκκινήσεων ενώ το τεστ $CSTF_1TF_2$ έχει το μικρότερο. Το αποτέλεσμα αυτό είναι σύμφωνο με το γεγονός ότι για την εμφάνιση r_1 (r_2) συνεχόμενων αποτελεσμάτων F_1 (F_2) απαιτείται η εμφάνιση τουλάχιστον r_1 (r_2) συνολικά αποτελεσμάτων F_1 (F_2). Ίδια συμπεριφορά παρατηρούμε και για την πιθανότητα αποδοχής του εξοπλισμού P_A . Επίσης, αξίζει να αναφέρουμε ότι η $E(Y)$ είναι αύξουσα συνάρτηση του r_1 για δεδομένες τιμές των k και r_2 . Ανάλογα συμπεράσματα (σε σχέση με τη $E(Y)$ ή τη P_A) μπορούμε να διατυπώσουμε και για τις παραμέτρους k και r_2 .

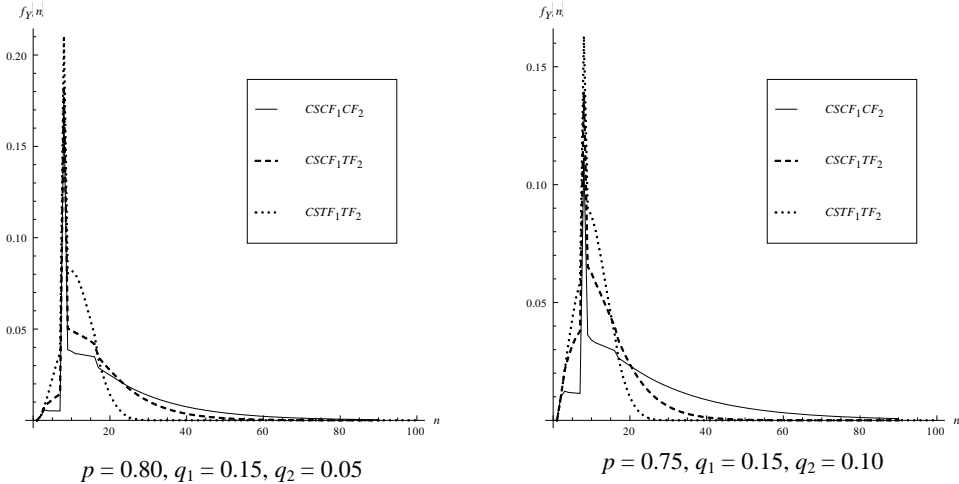
Σε ένα τεστ εκκίνησης σημαντικό ενδιαφέρον παρουσιάζει ο προσδιορισμός των k , r_1 και r_2 έτσι ώστε να γίνεται αποδεκτή με μεγάλη (μικρή) πιθανότητα μια συσκευή αξιοπιστίας $p \geq p_U$ ($p \leq p_L$). Τα p_L, p_U καθορίζουν το επίπεδο αξιοπιστίας του εξοπλισμού και ικανοποιούν την προφανή σχέση $p_L < p_U$. Έστω P_A^U (P_A^L) η πιθανότητα αποδοχής του εξοπλισμού για $p = p_U$ ($p = p_L$). Χρησιμοποιώντας τη μεθοδολογία των Smith and Griffith (2008) (δείτε επίσης Gera (2010)), επιλέγουμε τα k , r_1 και r_2 έτσι ώστε η $E(Y)$ για $p = p_U$ να πάρει τη μικρότερη δυνατή τιμή και ταυτόχρονα, για μικρές τιμές των α και β , να επαληθεύονται οι ανισώσεις $P_A^U > 1 - \beta$ και $P_A^L < \alpha$. Για $\alpha = 0.05$, $\beta = 0.10$, $p_U = 0.90$, $p_L = 0.60$ δίνουμε τον Πίνακα 4, έχοντας υποθέσει επιπλέον (για λόγους παρουσίασης του μοντέλου βελτιστοποίησης) ότι για τις πιθανότητες αποτυχημένης εκκίνησης q_1, q_2 ισχύει ότι $q_1 = cq_2$. Αξίζει να σημειώσουμε ότι στη γενική περίπτωση οι βέλτιστες τιμές των k, r_1 και r_2 δύναται να προκύψουν για οποιοδήποτε ζεύγος (q_1, q_2) με $0 < q_1, q_2 < 1$ και $q_1 + q_2 = 1 - p$.

Πίνακας 4: Βέλτιστες τιμές των παραμέτρων k, r_1 και r_2

Τεστ	c	k	r_1	r_2	P_A^U	P_A^L	$E(Y)$	$V(Y)^{1/2}$
$CSCF_1CF_2$	9/1	11	3	2	0.984	0.042	21.51	13.59
	8/2	9	2	2	0.909	0.047	14.38	14.38
	6/4	10	2	2	0.916	0.034	17.10	10.25
$CSCF_1TF_2$	9/1	11	3	2	0.955	0.031	20.88	12.62
	8/2	9	2	3	0.904	0.044	14.29	7.88
	6/4	10	2	4	0.913	0.033	17.06	9.37
$CSTF_1TF_2$	9/1	10	5	3	0.902	0.033	16.84	7.57
	8/2	9	4	4	0.902	0.049	14.25	6.20
	6/4	9	4	3	0.903	0.049	14.27	6.28

Τέλος, στο Σχήμα 1 δίνεται η γραφική παράσταση της συνάρτησης πιθανότητας του αριθμού εκκινήσεων μέχρι το τέλος των τριών τεστ για $k = 8$, $r_1 = 3$, $r_2 = 2$ και δύο ενδεικτικά σύνολα τιμών για τις παραμέτρους p , q_1 και q_2 .

Σχήμα 1. Συνάρτηση πιθανότητας της Y για διάφορες τιμές των p , q_1 , q_2 , k , r_1 και r_2



ABSTRACT

In this paper, we introduce and study three new start-up demonstration tests with two types of unsuccessful start-ups. The proposed tests use run and frequency quotas acceptance/rejection criteria, providing extensions of binary start-up demonstration tests that have been previously discussed in the literature. Using a Markov chain approach we establish general formulas for the distribution of the number of start-ups and its mean, as well as for the acceptance probability of the equipment. Numerical results are presented in order to illustrate our method.

ΑΝΑΦΟΡΕΣ

- Antzoulakos, D.L., Koutras, M.V. and Rakitzis, A.C. (2009). Start-up demonstration tests based on run and scan statistics. *Journal of Quality Technology*, **41**, 48-59.
- Balakrishnan, N., Balasubramanian, K. and Viveros, R. (1995). Start-up demonstration tests under correlation and corrective action, *Naval Research Logistics*, **42**, 1271-1276.
- Balakrishnan, N., Bersimis, S. and Koutras, M.V. (2009). Run and frequency quota rules in process monitoring and acceptance sampling. *Journal of Quality Technology*, **41**, 66-81.
- Balakrishnan, N. and Chan, P.S. (2000). Start-up demonstration tests with rejection of units upon observing d failures. *Annals of the Institute of Statistical Mathematics*, **52**, 184-196.
- Balakrishnan, N., Mohanty, S.G. and Aki, S. (1997). Start-up demonstration tests under Markov dependence model with corrective actions. *Annals of the Institute of Statistical Mathematics*, **49**, 155-169.

- Chan, P.S., NG, T.H.K. and Balakrishnan, N. (2008). Statistical inference for Start-up demonstration tests with rejection of units upon observing d failures. *Journal of Applied Statistics*, **35**, 867-878.
- Gera, A.E. (2010). A new start-up demonstration test. *IEEE Transactions on Reliability*, **59**, 128-131.
- Hahn, G.J. and Gage, J. B. (1983). Evaluation of start-up demonstration test. *Journal of Quality Technology*, **15**, 103-106.
- Koutras, M.V. and Alexandrou, V.A. (1997). Sooner waiting time problems in a sequence of trinary trials. *Journal of Applied Probability*, **34**, 593-609.
- Martin, D. E. K. (2004). Markovian start-up demonstration tests with rejection of units upon observing d failures. *European Journal of Operational Research*, **155**, 474-486.
- Philippou, A.N., Georghiou, G. and Philippou, G.N. (1983). A generalized geometric distribution and some properties. *Statistics and Probability Letters*, **1**, 171-175.
- Scollnik, D.P.M. (2010). Bayesian statistical inference for start-up demonstration tests with rejection of units upon observing d failures. *Journal of Applied Statistics*, **37**, 1113-1121.
- Scollnik, D.P.M. (2011). Bayesian inference for a class of start-up demonstration tests with rejection of units upon the observation of d failures. *Communications in Statistics-Theory and Methods*, **40**, 2528-2538.
- Smith, M.L.P. and Griffith S.W. (2005). Start-up demonstration tests based on consecutive successes and total failures. *Journal of Quality Technology*, **37**, 186-198.
- Smith, M.L.D. and Griffith, W.S. (2008). The analysis and comparison of start-up demonstration tests. *European Journal of Operational Research*, **186**, 1029-1045.
- Viveros, R. and Balakrishnan, N. (1993). Statistical inference from start-up demonstration test data. *Journal of Quality Technology*, **25**, 119-130.



ΠΕΡΙ ΤΗΣ ΚΑΤΑΝΟΜΗΣ ΤΟΥ ΑΘΡΟΙΣΜΑΤΟΣ ΔΙΑΚΟΠΗΣ: ΕΦΑΡΜΟΓΕΣ ΣΤΟΝ ΠΟΙΟΤΙΚΟ ΕΛΕΓΧΟ

A. X. Ρακιτζής¹, M. B. Μπούτσικας², Δ. Α. Αντζουλάκος²

¹Τμήμα Στατιστικής και Αναλ.-Χρημ. Μαθηματικών, Πανεπιστήμιο Αιγαίου
arakitz@aegean.gr

²Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης, Πανεπιστήμιο Πειραιά
{mbouts, dantz}@unipi.gr

ΠΕΡΙΛΗΨΗ

Έστω T ένας χρόνος διακοπής (*stopping time*) που βασίζεται σε μια ακολουθία ανεξάρτητων και ισόνομων τ.μ. Z_1, Z_2, \dots , και έστω $S_T = Z_1 + Z_2 + \dots + Z_T$ το αντίστοιχο άθροισμα διακοπής (*stopped sum*). Χρησιμοποιώντας τη θεμελιώδη ταυτότητα του Wald είναι δυνατόν να εξαχθεί η κατανομή του S_T όταν είναι γνωστή η κατανομή του T , και αντίστροφα. Ως παράδειγμα, παρουσιάζονται δύο εφαρμογές από την περιοχή του ποιοτικού ελέγχου. Η πρώτη αφορά την εύρεση της κατανομής του αριθμού των ελαττωματικών προϊόντων S_T τα οποία προκύπτουν κατά τη διαδικασία επιθεώρησης σωρών αντικειμένων, ενώ η δεύτερη αφορά την εύρεση της από κοινού κατανομής του αριθμού των εκκινήσεων T και του αριθμού των επιτυχημένων εκκινήσεων S_T σε ένα τεστ εκκίνησης αποδοχής εξαρτημάτων.

Λέξεις Κλειδιά: Άθροισμα διακοπής, Χρόνος διακοπής, Δειγματοληψία αποδοχής, Τεστ εκκίνησης αποδοχής εξαρτημάτων

1. ΕΙΣΑΓΩΓΗ

Σε πολλές περιοχές της εφαρμοσμένης έρευνας, αντικείμενο μελέτης αποτελεί ο χρόνος T (τ.μ.) μέχρι τη λήψη μιας δεδομένης απόφασης, ο οποίος σχετίζεται με την ακολουθία των τ.μ. Z_1, Z_2, \dots , καθώς επίσης και τα αντίστοιχα μερικά αθροίσματα $S_n = Z_1 + Z_2 + \dots + Z_n, n = 1, 2, \dots$. Ο χρόνος αναμονής T και το αντίστοιχο τυχαίο άθροισμα S_T είναι γνωστά ως χρόνος διακοπής (*stopping time*) και άθροισμα διακοπής (*stopped sum*). Πρακτικά προβλήματα που σχετίζονται με χρόνους διακοπής συναντώνται σε διάφορες επιστημονικές περιοχές όπως στην ακολουθιακή ανάλυση, στον ποιοτικό έλεγχο, στα αναλογιστικά και χρηματοοικονομικά μαθηματικά κλπ.

Η μελέτη των ιδιοτήτων των κατανομών των T και S_T ξεκίνησε, ουσιαστικά, από τον Wald (1945) ο οποίος έδωσε δύο πολύ γνωστές εξισώσεις οι οποίες συνδέουν τα T και S_T . Οι εξισώσεις αυτές είναι γνωστές ως η *Θεμελιώδης Ταυτότητα* του Wald

(Wald's fundamental identity) και η *Πρώτη Ταυτότητα* του Wald (Wald's first equation) και δίνονται, αντίστοιχα, από τις σχέσεις

$$E\left(\left(E(e^{wZ})\right)^{-T} e^{wS_T}\right) = 1 \text{ και } E(S_T) = E(Z)E(T),$$

όπου $E(e^{wZ})$ είναι η ροπογεννήτρια των $Z_i, i=1, 2, \dots$ και $E(Z)$ η μέση τιμή αυτών. Στη βιβλιογραφία υπάρχει πλήθος αναφορών και επεκτάσεων σχετικά με τις παραπάνω εξισώσεις. Ενδεικτικά αναφέρουμε τις εργασίες των Blom (1949), Franken and Lisek (1982) και Kallenberg (1989).

Πρόσφατα, οι Antzoulakos and Boutsikas (2007) κατέληξαν σε μια ειδική σχέση για τις κατανομές των T και S_T . Πιο συγκεκριμένα, αν T_r είναι ο χρόνος αναμονής μέχρι την r -οστή εμφάνιση ενός σχηματισμού \mathcal{E} σε μια ακολουθία ανεξάρτητων και ισόνομων δίτιμων δοκιμών Z_1, Z_2, \dots και S_{T_r} είναι ο αριθμός των επιτυχιών μέχρι εκείνη τη χρονική στιγμή, τότε χρησιμοποιώντας την πιθανογεννήτρια συνάρτηση της τ.μ. T_r μπορούμε άμεσα να βρούμε την από κοινού πιθανογεννήτρια συνάρτηση των (T_r, S_{T_r}) . Στην παρούσα εργασία δίδεται μια επέκταση αυτών των αποτελεσμάτων για οποιαδήποτε κατανομή των Z_i και για οποιοδήποτε χρόνο διακοπής T .

Η δομή της εργασίας είναι η εξής: Στην 2^η Ενότητα δίδονται τα βασικά αποτελέσματα τα οποία συνδέουν τις κατανομές των T και S_T και στη συνέχεια παρουσιάζονται δύο εφαρμογές από την περιοχή του στατιστικού ελέγχου ποιότητας. Συγκεκριμένα, στην 3^η Ενότητα εξάγεται η κατανομή του αριθμού των ελαττωματικών προϊόντων όπως αυτά προκύπτουν κατά τη διαδικασία επιθεώρησης μέσω ενός συστήματος δειγματοληψίας αποδοχής, και στην 4^η Ενότητα εξάγεται η από κοινού κατανομή του αριθμού των εκκινήσεων και του αριθμού των επιτυχημένων εκκινήσεων σε ένα τεστ εκκίνησης αποδοχής εξαρτημάτων. Δίνονται επίσης ενδεικτικά αριθμητικά παραδείγματα.

2. ΒΑΣΙΚΑ ΑΠΟΤΕΛΕΣΜΑΤΑ

Έστω T ένας χρόνος διακοπής ορισμένος σε μια ακολουθία ανεξάρτητων και ισόνομων τ.μ. Z_1, \dots, Z_n (δηλ., $[T = n] = \{\omega \in \Omega : T(\omega) = n\} \in \mathcal{F}_n = \sigma(Z_1, \dots, Z_n) \forall n = 1, 2, \dots$, και έστω το αντίστοιχο άθροισμα διακοπής $S_T = Z_1 + Z_2 + \dots + Z_T$. Το επόμενο Θεώρημα συνδέει την κατανομή του T με την κατανομή του S_T (για την απόδειξη, δείτε Boutsikas et al. (2011)):

Θεώρημα: Έστω T ένας χρόνος διακοπής ορισμένος σε μια ακολουθία τ.μ. Z_1, Z_2, \dots και έστω τ.μ. Y τέτοια ώστε η $Y \cdot I_{[T=n]}$ να είναι \mathcal{F}_n -μετρήσιμη. Τότε

$$E\left(Y e^{wS_T} I_{[T<\infty]}\right) = \tilde{E}_w\left(Y \prod_{i=1}^T E(e^{wZ_i}) I_{[T<\infty]}\right), \quad (1)$$

για όλα τα w για τα οποία υπάρχουν οι παραπάνω μέσες τιμές.

Η μέση τιμή $\tilde{E}_w(\cdot)$ υπολογίζεται ως προς το μέτρο πιθανότητας \tilde{P}_w υπό το οποίο ισχύει ότι οι τ.μ. Z_1, Z_2, \dots είναι ανεξάρτητες και ισόνομες με συνάρτηση (πυκνότητας) πιθανότητας

$$f_w(z) = e^{wz}f(z)/E(e^{wZ}),$$

όπου $f(z)$ είναι η συνάρτηση (πυκνότητας) πιθανότητας των τ.μ. Z_1, Z_2, \dots (υπό το αρχικό μέτρο πιθανότητας P) και $E(e^{wZ})$ είναι η αντίστοιχη ροπογεννήτρια συνάρτηση της κατανομής των $Z_i, i = 1, 2, \dots$. Αξίζει επίσης να αναφέρουμε ότι μπορούμε πάντοτε να αλλάξουμε το αρχικό μέτρο πιθανότητας P σε ένα νέο μέτρο πιθανότητας \tilde{P}_w στον (Ω, \mathcal{F}) υπό το οποίο οι τ.μ. Z_1, Z_2, \dots είναι ανεξάρτητες αλλά με $Z_i \sim f_w(z), i = 1, 2, \dots$. Για μια αυστηρή κατασκευή του χώρου πιθανότητας $(\Omega, \mathcal{F}, \tilde{P}_w)$ παραπέμπουμε στην εργασία των Boutsikas *et al.* (2011).

Σημείωση: Η σχέση (1) αποτελεί ειδική μορφή της *Ταυτότητας Λόγου Πιθανοφαινοίων* του Wald (*Wald's Likelihood Ratio Identity*, δείτε π.χ. Siegmund (1985) ή Lai (2004)).

Οι Antzoulakos and Boutsikas (2007) διατύπωσαν το παρακάτω αποτέλεσμα: Αν Z_1, Z_2, \dots είναι μια ακολουθία ανεξάρτητων και ισόνομων δίτιμων τ.μ. με $P(Z_i = 1) = p, P(Z_i = 0) = 1 - p$ και T είναι ο χρόνος αναμονής (αριθμός δοκιμών) μέχρι την εμφάνιση ενός σχηματισμού \mathcal{E} , τότε η από κοινού πιθανογεννήτρια συνάρτηση των (T, S_T) προκύπτει από την πιθανογεννήτρια της T μέσω της σχέσης

$$E(u^T t^{S_T}) = \tilde{E}_t \left((u(1-p + pt))^T \right),$$

για όλα τα t, u σε μια περιοχή του 0, όπου η αναμενόμενη τιμή \tilde{E}_t υπολογίζεται υπό το μέτρο πιθανότητας \tilde{P}_t , για το οποίο ισχύει ότι

$$\tilde{P}_t(Z_i = 1) = 1 - \tilde{P}_t(Z_i = 0) = (pt) / (1 - p + pt).$$

Μια γενίκευση του παραπάνω αποτελέσματος για οποιαδήποτε κατανομή των Z_i και για οποιοδήποτε χρόνο διακοπής T δίδεται στην επόμενη πρόταση, η οποία προκύπτει με άμεση εφαρμογή του παραπάνω Θεωρήματος:

Πρόταση: Αν $P(T < \infty) = \tilde{P}_w(T < \infty) = 1$ τότε

$$E(u^T e^{wS_T}) = \tilde{E}_w \left((uE(e^{wZ}))^T \right), \quad (2)$$

για όλα τα u, w για τα οποία υπάρχουν οι παραπάνω μέσες τιμές.

Απόδειξη: Έστω ακολουθία ανεξάρτητων και ισόνομων τ.μ. Z_1, Z_2, \dots . Τότε η (2) προκύπτει άμεσα από την (1) για $Y = u^T$ (αφού $u^T I_{[T=n]} = u^n I_{[T=n]}$ είναι \mathcal{F}_n -μετρήσιμη). ■

Θέτοντας $u = 1$ στη σχέση (2), προκύπτει η ροπογεννήτρια της κατανομής του S_T

$$E(e^{wS_T}) = \tilde{E}_w \left((E(e^{wZ}))^T \right).$$

Στην περίπτωση που η τ.μ. S_T είναι διακριτή θέτουμε στη (2) το e^w ίσο με t και προκύπτει η από κοινού πιθανογεννήτρια συνάρτηση των T και S_T ,

$$E(u^T t^{S_T}) = \tilde{E}_t \left((uE(t^Z))^T \right), \quad (3)$$

καθώς και η πιθανογεννήτρια συνάρτηση του S_T (θέτοντας $u = 1$ στη σχέση (3))

$$E(t^{S_r}) = \tilde{E}_t \left((E(t^Z))^T \right). \quad (4)$$

Στη συνέχεια, δίνουμε δύο πρακτικές εφαρμογές των παραπάνω αποτελεσμάτων από την επιστημονική περιοχή του Στατιστικού Ελέγχου Ποιότητας.

3. Η ΚΑΤΑΝΟΜΗ ΤΟΥ ΑΡΙΘΜΟΥ ΤΩΝ ΕΛΑΤΤΩΜΑΤΙΚΩΝ ΑΝΤΙΚΕΙΜΕΝΩΝ ΣΕ ΕΝΑ ΣΥΣΤΗΜΑ ΔΕΙΓΜΑΤΟΛΗΨΙΑΣ

Έστω ότι σωροί n προϊόντων περνούν από ολική (100%) επιθεώρηση, p είναι το ποσοστό των ελαττωματικών προϊόντων που αποδίδει η παραγωγική διεργασία και $Z_i, i = 1, 2, \dots$, ο αριθμός των ελαττωματικών προϊόντων στον i -οστό επιθεωρούμενο σωρό. Προφανώς τα $Z_i, i = 1, 2, \dots$, ακολουθούν διωνυμική κατανομή με παραμέτρους n, p (συμβ. $Z_i \sim B(n, p)$). Θεωρούμε ότι το σύστημα δειγματοληψίας ξεκινάει με κανονικό έλεγχο (*normal sampling level*) και μπορεί να μεταπηδήσει μόνο σε αυστηρό έλεγχο (*tightened level*). Έστω c ο αριθμός αποδοχής του κανονικού ελέγχου, δηλ. ο σωρός απορρίπτεται αν κατά την επιθεώρηση βρεθούν περισσότερα από c ελαττωματικά αντικείμενα. Ορίζουμε τον επόμενο κανόνα εναλλαγής (*switching rule*) στο επίπεδο επιθεώρησης:

Κανόνας: Το επίπεδο επιθεώρησης αλλάζει από κανονικό σε αυστηρό αν βρεθούν k συνεχόμενοι σωροί με περισσότερα από c ελαττωματικά αντικείμενα ο καθένας.

Έστω T ο αριθμός των επιθεωρημένων σωρών μέχρι την αλλαγή του επιπέδου επιθεώρησης (χρόνος διακοπής), ο οποίος εκφράζεται ως

$$T = \inf \{ l \geq k : Z_{l-k+1} > c, \dots, Z_l > c \}$$

και $S_T = Z_1 + Z_2 + \dots + Z_T$ ο συνολικός αριθμός των ελαττωματικών αντικειμένων που βρέθηκαν μέχρι την εναλλαγή στο επίπεδο επιθεώρησης (άθροισμα διακοπής).

Για την εύρεση της κατανομής του S_T μπορούμε να χρησιμοποιήσουμε την (4). Αφού η πιθανογεννήτρια των τ.μ. Z_i είναι $E(t^Z) = (1 - p + pt)^n$, η κατανομή των Z_i υπό το μέτρο πιθανότητας \tilde{P}_t είναι

$$\tilde{P}_t(Z_i = x) = \frac{t^x P(Z_i = x)}{E(t^Z)} = \binom{n}{x} \left(\frac{pt}{1 - p + pt} \right)^x \left(\frac{1 - p}{1 - p + pt} \right)^{n-x}, \quad x = 0, 1, \dots, n,$$

δηλ. $Z_i \sim B(n, p_t)$ με $p_t = pt / (1 - p + pt)$.

Η κατανομή του T είναι η γεωμετρική κατανομή τάξης k (Philippou et al. (1983)) με πιθανότητα επιτυχίας $q = P(Z_i > c)$ και πιθανογεννήτρια συνάρτηση,

$$E(u^T) = \frac{(qu)^k (1 - qu)}{1 - u + (1 - q)q^k u^{k+1}}, \quad u \in [0, 1]. \quad (5)$$

Παρατηρώντας ότι υπό το μέτρο πιθανότητας \tilde{P}_t έχουμε ότι

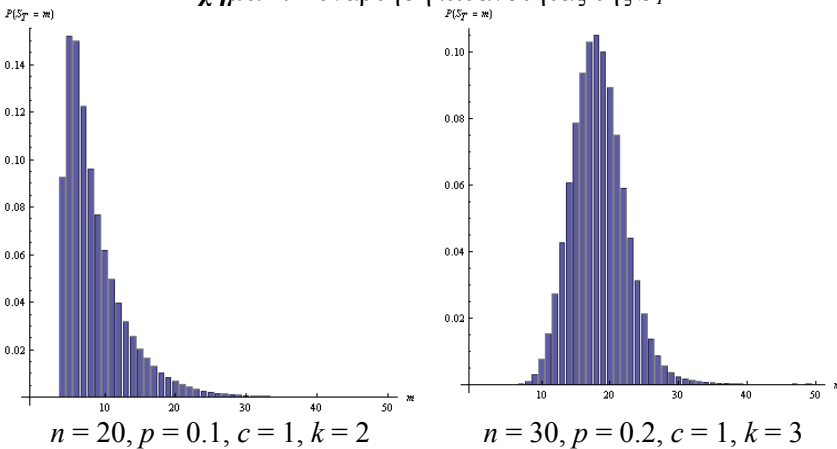
$$q_t = \tilde{P}_t(Z_i > c) = 1 - \sum_{x=0}^c \binom{n}{x} p_t^x (1 - p_t)^{n-x},$$

προκύπτει άμεσα από τις (3) και (5) ότι η από κοινού πιθανογεννήτρια συνάρτηση της (T, S_T) είναι

$$E(u^T t^{S_T}) = \tilde{E}_t \left[(u(1-p+pt)^n)^T \right] = \frac{(q_t u(1-p+pt)^n)^k (1-q_t u(1-p+pt)^n)}{1-u(1-p+pt)^n + (1-q_t)q_t^k (u(1-p+pt)^n)^{k+1}}.$$

Η πιθανογεννήτρια συνάρτηση $E(t^{S_T})$ του S_T προκύπτει από τον παραπάνω τύπο για $u=1$. Η κατανομή του S_T , της οποίας το στήριγμα είναι το $\{k(c+1), k(c+1)+1, \dots\}$ μπορεί να υπολογιστεί για διάφορες τιμές των παραμέτρων n, p, c και k , όπως φαίνεται στα δύο γραφήματα του Σχήματος 1. Για τον υπολογισμό των ζητούμενων πιθανοτήτων χρησιμοποιήθηκε η συνάρτηση `SeriesCoefficient` του *Mathematica*.

Σχήμα 1. Συνάρτηση πιθανότητας της S_T



Η μεθοδολογία που παρουσιάσαμε μας δίνει τη δυνατότητα υπολογισμού των ποσοστιαίων σημείων της κατανομής του S_T , τα οποία περιγράφουν πληρέστερα την απόδοση του συστήματος δειγματοληψίας αποδοχής. Ενδεικτικά παραθέτουμε τον επόμενο πίνακα με τη μέση τιμή και τα ποσοστιαία σημεία της κατανομής του S_T για $n=40, c=1, k=3$ και $p=0.020(0.005)0.100$.

Πίνακας 1. Μέση τιμή και ποσοστιαία σημεία της κατανομής του S_T

p	$E(S_T)$	5%	25%	50%	75%	95%
0.020	142.04	13	45	100	195	414
0.025	72.32	9	25	52	98	205
0.030	44.97	8	17	33	60	123
0.035	31.93	7	13	24	42	84
0.040	24.87	7	11	19	32	63
0.045	20.69	8	10	16	26	50
0.050	18.08	8	10	14	23	42
0.055	16.39	7	9	13	20	36
0.060	15.28	7	9	13	19	32
0.065	14.55	7	9	12	17	29

0.070	14.09	7	10	12	17	27
0.075	13.84	7	10	12	16	26
0.080	13.72	8	10	12	16	25
0.085	13.73	8	10	12	16	24
0.090	13.82	8	10	13	16	24
0.095	13.99	8	11	13	16	23
0.100	14.23	9	11	13	16	23

Η γνώση της κατανομής του συνολικού αριθμού των ελαττωματικών αντικειμένων μέχρι την εναλλαγή στο επίπεδο (ή το τέλος) της επιθεώρησης είναι σημαντική, κυρίως σε προγράμματα επιδιορθωτικού ελέγχου (*rectifying inspection programs*). Για παράδειγμα, για $p = 0.02$ προκύπτει από τον Πίνακα 1 ότι ο αναμενόμενος αριθμός ελαττωματικών αντικειμένων μέχρι την αλλαγή στο επίπεδο επιθεώρησης είναι περίπου 142. Επίσης, η διάμεσος είναι ίση με 100 το οποίο συνεπάγεται ότι η πιθανότητα να υπερβεί ο αριθμός των ελαττωματικών αντικειμένων, μέχρι την αλλαγή στο επίπεδο επιθεώρησης, την τιμή 100 είναι μικρότερη από 50%. Τέλος, ο υπολογισμός της $E(S_T)$ μπορεί να γίνει και με εφαρμογή της πρώτης εξίσωσης του Wald για $E(Z) = np$ και $E(T) = (1 - q^k)/(q^k(1 - q))$.

Αξίζει να αναφέρουμε ότι η παραπάνω μεθοδολογία μπορεί να εφαρμοστεί με τον ίδιο τρόπο για τη μελέτη και πιο σύνθετων κανόνων εναλλαγής, όπου οι τ.μ. Z_i προέρχονται από μια κατανομή F (συνεχή ή διακριτή), και η εναλλαγή στο επίπεδο επιθεώρησης συμβαίνει σύμφωνα με έναν χρόνο διακοπής τύπου k -από- m . Αυτό που χρειάζεται να γνωρίζουμε είναι η κατανομή του χρόνου διακοπής T , η οποία στην περίπτωση αυτή είναι η γεωμετρική κατανομή τάξης k/m που έχει μελετηθεί εκτενώς στη βιβλιογραφία (δείτε, π.χ. Balakrishnan and Koutras (2002), σελ. 291).

Τέλος, αντίστροφα, είναι δυνατή η εύρεση της κατανομής του T όταν είναι γνωστή η κατανομή του S_T αντιστρέφοντας κατάλληλα τη σχέση (2). Παράδειγμα εφαρμογής αυτού αποτελεί η εύρεση της κατανομής του χρόνου πρώτης εξόδου από το διάστημα $(-a, b)$ ($a, b > 0$) ενός τυχαίου περιπάτου με εκθετικά άλματα. Για περισσότερες λεπτομέρειες δείτε Boutsikas *et al.* (2011).

4. Η ΚΑΤΑΝΟΜΗ ΤΟΥ ΑΡΙΘΜΟΥ ΤΩΝ ΕΠΙΤΥΧΗΜΕΝΩΝ ΕΚΚΙΝΗΣΕΩΝ ΣΕ ΕΝΑ ΤΕΣΤ ΕΚΚΙΝΗΣΗΣ

Ένα τεστ εκκίνησης αποδοχής εξαρτημάτων (*start-up demonstration test*) είναι μια διαδικασία η οποία εξετάζει την ικανότητα μιας συσκευής να τεθεί σε λειτουργία (Hahn and Gage (1983), Viveros and Balakrishnan (1993)). Εκτός από την κατανομή του αριθμού T των εκκινήσεων μέχρι τον τερματισμό του τεστ και τη λήψη απόφασης (αποδοχή ή απόρριψη του εξοπλισμού), σημαντικό ρόλο παίζει και η κατανομή του αριθμού των επιτυχημένων εκκινήσεων μέχρι τον τερματισμό του τεστ, η οποία π.χ. μπορεί να χρησιμοποιηθεί για την εκτίμηση της άγνωστης πιθανότητας επιτυχημένης εκκίνησης p .

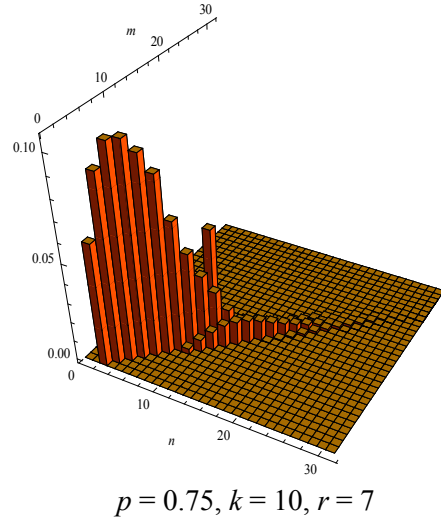
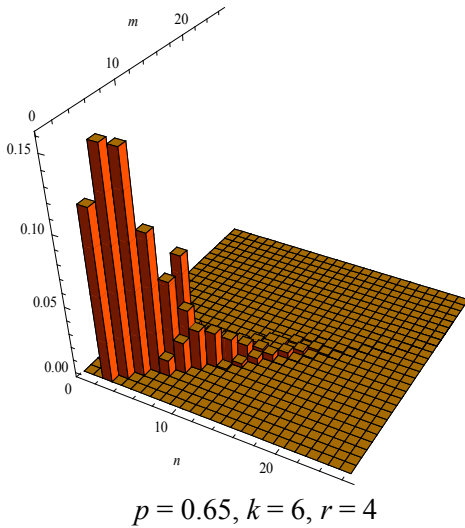
Έστω ότι ένας εξοπλισμός υπόκειται στο παρακάτω τεστ εκκίνησης αποδοχής εξαρτημάτων με δύο δυνατά αποτελέσματα (επιτυχημένη, αποτυχημένη εκκίνηση): 0

εξοπλισμός γίνεται αποδεκτός εάν συμβούν k -συνεχόμενες επιτυχημένες εκκινήσεις πριν από την εμφάνιση 2 αποτυχημένων οι οποίες απέχουν το πολύ r δοκιμές, διαφορετικά απορρίπτεται ($2 \leq r \leq k$). Έστω T ο αριθμός των εκκινήσεων του εξοπλισμού μέχρι τη λήψη απόφασης (χρόνος διακοπής) και S_T ο αριθμός των επιτυχημένων εκκινήσεων του εξοπλισμού μέχρι εκείνη τη στιγμή (άθροισμα διακοπής). Μπορεί να διαπιστωθεί (δείτε Antzoulakos et al. (2009)) ότι η πιθανογεννήτρια συνάρτηση της τ.μ. T δίνεται από τη σχέση

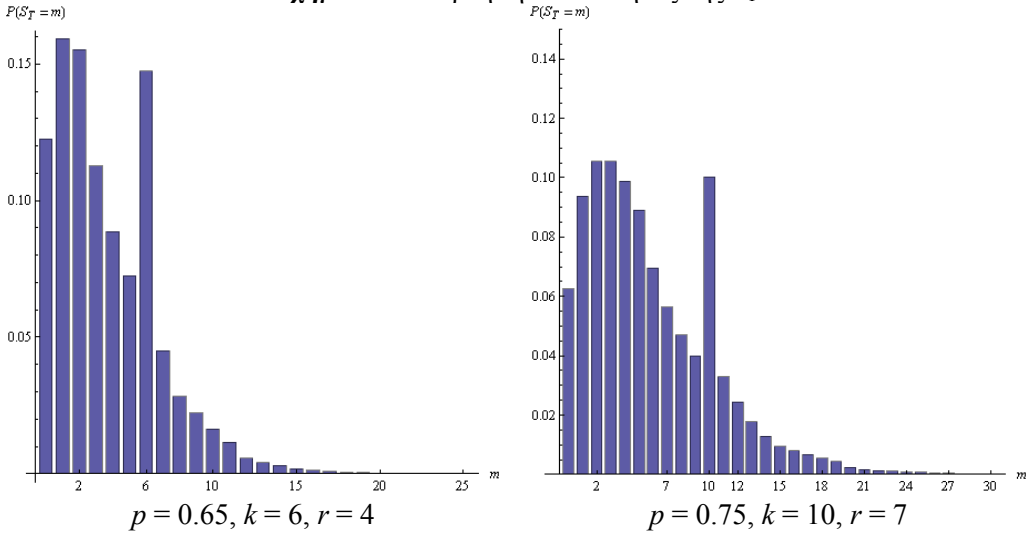
$$E(z^T) = \frac{(1-pz)(pz)^k(1-pz+qz-(qz)(pz)^{r-1})+(qz)^2(1-(pz)^k)(1-(pz)^{r-1})}{(1-pz)(1-pz-(qz)(pz)^{r-1}+(qz)(pz)^k)}$$

Η από κοινού πιθανογεννήτρια συνάρτηση της διδιάστατης τ.μ. (T, S_T) μπορεί να βρεθεί από την (3) θέτοντας $E(t^Z) = (q+pt)$ και παρατηρώντας ότι υπό το \tilde{P}_t έχουμε ότι $Z_i \sim B(1, (pt)/(q+pt))$. Χρησιμοποιώντας την από κοινού πιθανογεννήτρια της (T, S_T) μπορούμε να βρούμε την από κοινού συνάρτηση πιθανότητας της (T, S_T) και τη συνάρτηση πιθανότητας της S_T . Ενδεικτικά παραθέτουμε τις ακόλουθες γραφικές παραστάσεις.

Σχήμα 2. Από κοινού συνάρτηση πιθανότητας της (T, S_T)



Σχήμα 3. Συνάρτηση πιθανότητας της S_T



ABSTRACT

Let T be a stopping time associated with a sequence of i.i.d. r.v.'s Z_1, Z_2, \dots , and let $S_T = Z_1 + Z_2 + \dots + Z_T$ be the corresponding stopped sum. By applying *Wald's fundamental identity* we present relations between the distributions of T and S_T . These relations imply that, when the distribution of S_T is known, then the distribution of T is also known and vice versa. Two interesting applications from the area of statistical quality control are offered. In the first one we derive the distribution of the number of non-conforming items S_T which occurs during an inspection procedure, according to an acceptance sampling system based on a k -run switching rule. In the second one we derive the joint distribution of the random vector (T, S_T) where T is the number of start-ups until the end of a specific binary start-up demonstration test and S_T is the corresponding number of successful start-up attempts until the end of the test.

Ευχαριστίες: Ευχαριστούμε τους κριτές για τις εύστοχες παρατηρήσεις τους. Η έρευνα του πρώτου συγγραφέα υποστηρίχθηκε από το Ίδρυμα Κρατικών Υποτροφιών.

ΑΝΑΦΟΡΕΣ

- Antzoulakos, D. L. and Boutsikas, M. (2007). A direct method to obtain the joint distribution of successes, failures and patterns in enumeration problems. *Statistics and Probability Letters*, **77**, 32-39.
- Antzoulakos, D. L., Koutras, M. V. and Rakitzis, A.C. (2009). Start-up demonstration tests based on run and scan statistics. *Journal of Quality Technology*, **41**, 48-59.
- Balakrishnan, N. and Koutras, M. V. (2002). *Runs and Scans with Applications*, New York: John Wiley.
- Blom, G. (1949). A generalization of Wald's fundamental identity. *The Annals of Mathematical Statistics*, **20**, 439-444.

- Boutsikas, M. V., Rakitzis, A. C. and Antzoulakos, D. L. (2011). On the relation between the distributions of stopping time and stopped sum via Wald's identity with applications, arXiv: 1008.0116v1.
- Franken, P. and Lisek, B. (1982). On Wald's identity for dependent variables. *Probability Theory and Related Fields*, **60**, 143-150.
- Hahn, G. J. and Gage, J. B. (1983). Evaluation of start-up demonstration test. *Journal of Quality Technology*, **15**, 103-106.
- Kallenberg, O. (1989). General Wald-type identities for exchangeable sequences and processes. *Probability Theory and Related Fields*, **83**, 447-487.
- Lai, T. L. (2004). Likelihood ratio identities and their applications to sequential analysis. *Sequential Analysis*, **23**, 467-497.
- Philippou, A. N., Georghiou, G. and Philippou, G. N. (1983). A generalized geometric distribution and some properties, *Statistics and Probability Letters*, **1**, 171-175.
- Siegmund, D. (1985). *Sequential Analysis*, New York: Springer.
- Viveros, R. and Balakrishnan, N. (1993). Statistical inference from start-up demonstration test data. *Journal of Quality Technology*, **25**, 119-130.
- Wald, A. (1945). Sequential tests of statistical hypothesis. *Annals of Mathematical Statistics*, **16**, 117-186.



ΕΦΑΡΜΟΓΗ ΜΕΘΟΔΩΝ ΠΟΛΥΜΕΤΑΒΛΗΤΗΣ ΣΤΑΤΙΣΤΙΚΗΣ ΑΝΑΛΥΣΗΣ ΓΙΑ ΤΗΝ ΑΝΙΧΝΕΥΣΗ & ΠΕΡΙΓΡΑΦΗ ΤΩΝ ΠΡΟΤΙΜΗΣΕΩΝ ΤΟΥ ΕΛΛΗΝΙΚΟΥ ΠΛΗΘΥΣΜΟΥ ΩΣ ΠΡΟΣ ΤΗΝ ΚΑΤΑΝΑΛΩΣΗ ΑΛΚΟΟΛΟΥΧΩΝ ΠΟΤΩΝ

Ρεκούτη Αγγελική, Αλεβίζος Φίλιππος
Τμήμα Μαθηματικών, Πανεπιστήμιο Πατρών,
Πανεπιστημιούπολη Ρίου, 26500 Πάτρα
rekouti.ageliki@gmail.com, philipos@math.upatras.gr

ΠΕΡΙΛΗΨΗ

Η αξιολόγηση των πληροφοριών που συλλέγονται στην έρευνα αγοράς για τη λήψη αποφάσεων με όσο το δυνατόν μικρότερο ρίσκο και μεγαλύτερο κέρδος δημιουργεί προβλήματα πολυδιάστατου χαρακτήρα τα οποία επιλύονται με την εφαρμογή πολυμεταβλητών στατιστικών μεθόδων. Στην εργασία αυτή ανιχνεύουμε και περιγράφουμε σε δείγμα πραγματικών δεδομένων τις καταναλωτικές συνήθειες του Ελληνικού πληθυσμού ως προς την εβδομαδιαία κατανάλωση 9 κατηγοριών αλκοολούχων ποτών εφαρμόζοντας μεθόδους στατιστικής ανάλυσης. Τα δεδομένα που χρησιμοποιήθηκαν είναι πνευματική ιδιοκτησία της Focus Bari AE και προέρχονται από τις βάσεις δεδομένων της, οι οποίες προκύπτουν από την πανελλήνια έρευνα που διεξάγει ετησίως και καταγράφει την εξέλιξη στις καταναλωτικές συνήθειες και συνήθειες lifestyle των Ελλήνων. Αρχικά αναπτύσσουμε τη μέθοδο επίλυσης ως προς τα διάφορα στάδια σχεδιασμού και τις προϋποθέσεις της. Στη συνέχεια μελετάμε και περιγράφουμε το δείγμα μας με πολυμεταβλητές μεθόδους στατιστικής ανάλυσης. Τέλος, επικυρώνουμε τα αποτελέσματά μας ως προς τη σταθερότητα που εμφανίζουν στο δείγμα και τη γενίκευση τους στον πληθυσμό.

Λέξεις Κλειδιά: Πολυμεταβλητές Στατιστικές Μέθοδοι, Έρευνα Αγοράς

1. ΕΙΣΑΓΩΓΗ

Η Πολυμεταβλητή Ανάλυση είναι η εφαρμογή της Πολυδιάστατης Στατιστικής, η οποία περιλαμβάνει την ταυτόχρονη παρατήρηση και ανάλυση περισσότερων από μία μεταβλητών. Σκοπός είναι να μετρήσει, να εξηγήσει και να προβλέψει τον βαθμό της αλληλεπίδρασης μεταξύ των μεταβλητών [1].

Η ταξινόμηση των πολυδιάστατων μεθόδων και η επιλογή της κατάλληλης μεθόδου στηρίζονται σε τρεις αποφάσεις για τη μορφή των δεδομένων [1]:

- Μπορούν οι μεταβλητές να χωριστούν σε εξαρτημένες και ανεξάρτητες;

- Αν ναι, πόσες είναι οι εξαρτημένες;
- Ποιο είναι το είδος των ανεξάρτητων και των εξαρτημένων μεταβλητών;

Η μέθοδος εξαρτημένων μεταβλητών μπορεί να οριστεί σαν μία μέθοδος κατά την οποία μία μεταβλητή (ή ένα σύνολο μεταβλητών) που χαρακτηρίζεται ως εξαρτημένη μεταβλητή προβλέπεται ή ερμηνεύεται από άλλες μεταβλητές, τις ανεξάρτητες. Αντιθέτως, στις μεθόδους αλληλεξάρτησης όλες οι μεταβλητές αναλύονται ταυτόχρονα για να προσδιοριστεί η δομή του συνόλου των δεδομένων [1].

Οι μέθοδοι της Πολυμεταβλητής Ανάλυσης βρίσκουν εφαρμογή σε προβλήματα που παρουσιάζονται στον τομέα της Έρευνας Αγοράς λόγω του πολυδιάστατου χαρακτήρα τους. Η έρευνα αγοράς, μέσω της συλλογής και αξιολόγησης πληροφοριών από ομάδες του πληθυσμού, στοχεύει στο να βοηθήσει την εκάστοτε επιχείρηση στη λήψη αποφάσεων με όσο το δυνατό μικρότερο ρίσκο [2].

Στην εργασία μας το πρόβλημα με το οποίο ήρθαμε αντιμέτωποι είναι η ανίχνευση των καταναλωτικών συνηθειών ως προς την κατανάλωση ορισμένων αλκοολούχων ποτών. Επιλύθηκε με τη μέθοδο της Παραγοντικής Ανάλυσης, συμπυκνώνοντας σε ένα μικρότερο σύνολο συνθετικών μεταβλητών την πληροφορία που προέρχεται από τις αρχικές, με όσο το δυνατό μικρότερη απώλεια αυτής. Οι ομάδες των καταναλωτών που προέκυψαν περιγράφησαν ως προς τις ομοιότητες και τις διαφορές τους με τη χρήση της Ανάλυσης κατά Συστάδων και των γραφικών αναπαραστάσεων των παραγόντων.

2. ΜΕΘΟΔΟΙ ΕΠΙΛΥΣΗΣ ΚΑΙ ΠΕΙΡΑΜΑΤΙΚΑ ΑΠΟΤΕΛΕΣΜΑΤΑ

2.1 Παραγοντική Ανάλυση (ΠΑ)

Η ΠΑ είναι μία τεχνική αλληλεξάρτησης με πρωταρχικό στόχο τον προσδιορισμό της βασικής δομής των συσχετίσεων ενός μεγάλου συνόλου μεταβλητών και τη δημιουργία υποσυνόλων αλληλένδετων μεταβλητών, που ονομάζονται παράγοντες και συγκεντρώνουν το μεγαλύτερο δυνατό κομμάτι πληροφορίας αντιπροσωπεύοντας τις διαστάσεις των δεδομένων [1, 3, 4]. Το αντικείμενο της ΠΑ πρέπει να αναζητηθεί σε τέσσερα θέματα [1, 2, 3]:

- στον προσδιορισμό της μονάδας της ανάλυσης, απαντώντας στο ερώτημα «Τι ή ποιον μελετάμε;».
- στη συμπύκνωση των δεδομένων με στόχο πάντα να διατηρείται η φύση και ο χαρακτήρας των αρχικών μεταβλητών.
- στην ύπαρξη ισχυρών σχέσεων μεταξύ των μεταβλητών που συμμετέχουν στους παράγοντες.
- στη χρήση των αποτελεσμάτων της ΠΑ μαζί με άλλες πολυμεταβλητές τεχνικές.

Ο σχεδιασμός της ΠΑ αποτελείται από τη μελέτη των μεταβλητών ως προς το πλήθος και το είδος τους, από τους υπολογισμούς των δεδομένων για να προσδιοριστούν οι στόχοι της ομαδοποίησης και από το μέγεθος του δείγματος σαν ανεξάρτητο πλήθος, αλλά και σαν συνάρτηση του αριθμού των μεταβλητών [1, 2, 3].

Οι στατιστικές προϋποθέσεις της μεθόδου, οι οποίες απαιτούμε να ικανοποιούνται είναι

- το ποσοστό της διακύμανσης της μεταβλητής που ερμηνεύεται από όλους τους παράγοντες συνολικά να υπερβαίνει το 0,3. Στο κείμενο χρησιμοποιούμε τον όρο «communalities» [1, 2, 3].
- το μέτρο επάρκειας της δειγματοληψίας, MSA, των Kaiser-Meyer-Olkin να είναι μεγαλύτερο από 0,5 για κάθε μεταβλητή [1, 3].
- υψηλές τιμές στον έλεγχο σφαιρικότητας του Bartlett [1, 3, 5].
- ο βαθμός συσχέτισης της μεταβλητής με τον παράγοντα να είναι μεγαλύτερος από 0,3. Στο κείμενο χρησιμοποιούμε τον όρο «factor loadings» [1, 2].
- η μη ύπαρξη ακραίων τιμών [1, 5].

Τα δύο κριτήρια που χρησιμοποιούνται συχνότερα για την επιλογή του πλήθους των παραγόντων είναι το κριτήριο του Kaiser και το scree plot. Εφαρμόζονται συνδυαστικά λόγω των φαινομένων της υπερεκτίμησης ή υποεκτίμησης του πλήθους των αξόνων στο κριτήριο του Kaiser και της μεροληπτικής επιλογής του σημείου αλλαγής της κλίσης στο scree plot [1, 3].

Στο τελευταίο στάδιο πρέπει να αξιολογηθεί ο βαθμός γενίκευσης των αποτελεσμάτων στον πληθυσμό. Η πιο άμεση μέθοδος επικύρωσης των αποτελεσμάτων είναι να επιβεβαιωθούν από κάποιο άλλο δείγμα δεδομένων, είτε από ένα τυχαίο υποσύνολο του ήδη υπάρχοντος, αν το μέγεθος του το επιτρέπει [1].

2.2 Ανάλυση Κύριων Συνιστωσών

Η ανάλυση κύριων συνιστωσών μετατρέπει, μέσα από έναν ορθογώνιο μετασχηματισμό που εφαρμόζεται πάνω στο μητρώο συσχετίσεων, ένα σύνολο μεταβλητών που πιθανώς να είναι συσχετισμένες σε ένα σύνολο ασυσχέτιστων μεταβλητών [1].

2.3 Ανάλυση κατά Συστάδες

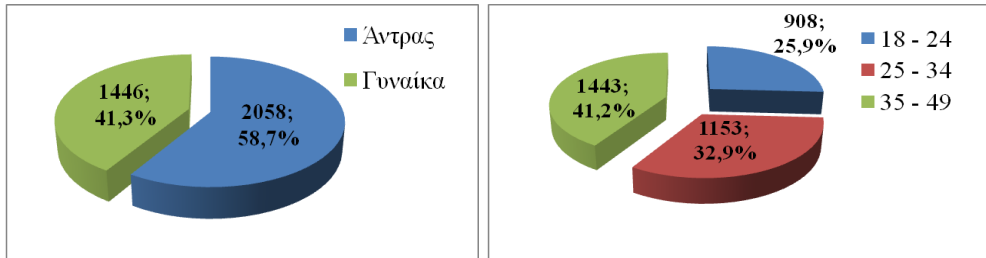
Η ανάλυση κατά συστάδες είναι μία αναλυτική μέθοδος ταξινόμησης ενός συνόλου ατόμων ή αντικειμένων σε έναν μικρότερο αριθμό αμοιβαία αποκλειόμενων ομάδων βασισμένων στις ομοιότητες μεταξύ των οντοτήτων τους. Το δείγμα πρέπει να είναι αντιπροσωπευτικό του πληθυσμού και είναι επιθυμητό να αποτελείται από ασυσχέτιστους παράγοντες [1, 3].

2.4 Παρουσίαση των πειραματικών αποτελεσμάτων

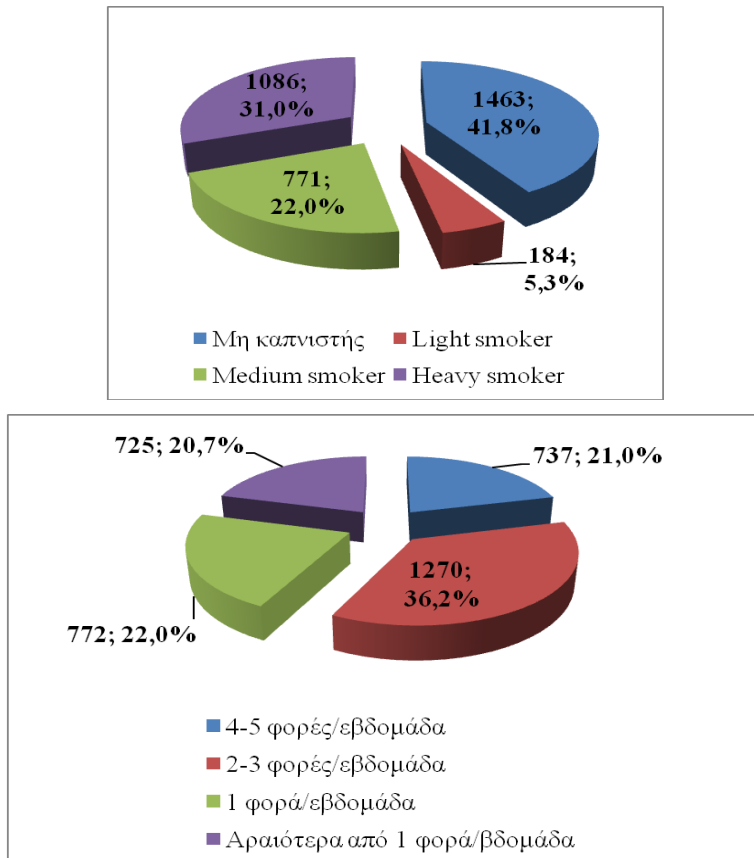
Τα δεδομένα που χρησιμοποιήθηκαν στην έρευνα αφορούν τη χρονική περίοδο Μάρτιος 2007 έως Μάρτιος 2008. Αρχικά χρησιμοποιήθηκαν 3731 πανελλαδικές συνεντεύξεις ατόμων που έχουν ηλικία από 18 έως 49 ετών, καπνίζουν τσιγάρα έτοιμα σε πακέτο ή στριφτά και κατανάλωσαν τουλάχιστον μία από τις 9 κατηγορίες αλκοολούχων ποτών (Πίνακας 1) την τελευταία εβδομάδα. Οι συνεντεύξεις αυτές μελετήθηκαν ως προς τις ακραίες και χαμένες τιμές καταλήγοντας στο τελικό σύνολο δεδομένων των 3504 καταναλωτών. Η κατανομή των δύο φύλων στο δείγμα διαφέρει ελάχιστα, με τους άνδρες να υπερέχουν των γυναικών. Η ηλικία είναι χωρισμένη σε τρεις ομάδες, 18-24 ετών, 25-34 ετών και 35-49 ετών. Οι καταναλωτές έχουν επιπλέον ομαδοποιηθεί ως προς τον αριθμό τσιγάρων που καπνίζουν την ημέρα σε

«Heavy smoker» (πάνω από 40 τσιγάρα/ημέρα), «Medium smoker» (από 20 έως 39 τσιγάρα/ημέρα), «Light smoker» (λιγότερα από 20 τσιγάρα/ημέρα) και σε «No smoker» (μη καπνιστές). Τέλος, τους ταξινομήσαμε και ως προς τον αριθμό εξόδων για διασκέδαση τη βδομάδα. Τα ποσοστά όλων αυτών των χαρακτηριστικών παρουσιάζονται στα γραφήματα που ακολουθούν στις εικόνες 1 και 2.

Εικόνα 1. Δημογραφικά χαρακτηριστικά καταναλωτών



Εικόνα 2. Συνήθειες καθημερινότητας καταναλωτών



Οι κατηγορίες των αλκοολούχων ποτών αποτελούν τις προς μελέτη μεταβλητές, δέχονται αριθμητικές τιμές και μετρώνε, σε κάθε κατηγορία ξεχωριστά, την ποσότητα

αλκοόλ (σε ποτήρια) που κατανάλωσε ο ερωτώμενος συνολικά τις τελευταίες 7 μέρες εντός και εκτός σπιτιού. Ονομαστικά οι 9 αυτές κατηγορίες και τα ποσοστά της εβδομαδιαίας κατανάλωσης κάθε μίας στο δείγμα παρουσιάζονται στον πίνακα 1.

Πίνακας 1. Εβδομαδιαία κατανάλωση ανά κατηγορία

Κατηγορία	Ποσοστό κατανάλωσης
εμφιαλωμένα αναψυκτικά με αλκοόλ	5%
ελληνική μύρα	65,8%
εισαγόμενη μύρα	9%
ούζο	14,5%
ουίσκι	29,2%
vodka	12,8%
gin	6,1%
βερμούτ / απεριτίφ	2,8%
ρούμι	4,1%

Τα ποσοστά κατανάλωσης σε κάποιες κατηγορίες είναι μικρά με αποτέλεσμα να προκύπτει αραιό μητρώο δεδομένων. Για το λόγο αυτό συνθέσαμε τις δύο μεταβλητές «Βερμούτ/Απεριτίφ» και «Εμφιαλωμένα αναψυκτικά με αλκοόλ» σε μία νέα μεταβλητή που ονομάσαμε «Ελαφριά ποτά», με συνολικό ποσοστό κατανάλωσης 7,6%. Επιπλέον, διασταυρώνοντας τις κατηγορίες των χαρακτηριστικών των καταναλωτών (φύλο, ηλικία, κάπνισμα, εβδομαδιαία διασκέδαση), δημιουργήσαμε 72 ομάδες καταναλωτών. Οι ομάδες αυτές αποτελούν τις γραμμές του τελικού μητρώου δεδομένων, το οποίο έχει ως στήλες τις 8 πλέον κατηγορίες εβδομαδιαίας κατανάλωσης. Τα στοιχεία του μητρώου είναι οι μέσες τιμές της κατανάλωσης κάθε ομάδας καταναλωτών στην αντίστοιχη κατηγορία αλκοολούχου ποτού.

Η μέθοδος που εφαρμόστηκε είναι ο R-τύπος της ΠΑ με μέθοδο εξαγωγής των παραγόντων την ανάλυση σε κύριες συνιστώσες και κριτήρια επιλογής τους το κριτήριο του Kaiser και το scree plot. Η ανάλυση των δεδομένων έγινε με τη χρήση του SPSS 17.0.

Το μέγεθος του δείγματος είναι ικανοποιητικό και 9 φορές μεγαλύτερο από το πλήθος των μεταβλητών. Η ερμηνεία της συνδιακύμανσης είναι ισχυρή αφού οι τιμές όλων των μεταβλητών ξεπερνάνε κατά πολύ το 0,3 που είναι το ελάχιστο αποδεκτό όριο. Το μέτρο επάρκειας της δειγματοληψίας συνολικά είναι $0,601 > 0,5$. Από το μητρώο μερικών συσχετίσεων δύο από τα στοιχεία της κυρίας διαγωνίου δεν είναι αποδεκτά ($MSA < 0,5$). Η τιμή του ελέγχου σφαιρικότητας των δεδομένων του Bartlett είναι 181,591. Ελέγξαμε το δείγμα ως προς τις ακραίες τιμές μέσω της ΠΑ και βρέθηκαν έξι τιμές μεγαλύτερες κατά απόλυτη τιμή από 3.

Στη συνέχεια αποκλείστηκαν από την ανάλυση οι συγκεκριμένες τιμές και εφαρμόστηκε η ΠΑ ξανά στο μητρώο διαστάσεων 66×8 . Η ερμηνεία της συνδιακύμανσης είναι και σε αυτή τη περίπτωση ισχυρή, ο γενικός δείκτης επάρκειας είναι μεγαλύτερος από πριν ($KMO = 0,694$) και όλα τα στοιχεία της κυρίας διαγωνίου του μητρώου μερικών συσχετίσεων είναι αποδεκτά ($MSA > 0,5$). Η τιμή

του ελέγχου σφαιρικότητας είναι 140,994. Τελικά, επιλέχτηκε προς ανάλυση το μητρώο των 66 γραμμών και 8 στηλών.

Σύμφωνα με το κριτήριο του Kaiser τρεις παράγοντες έχουν ιδιοτιμή μεγαλύτερη της μονάδος και ερμηνεύουν συνολικά το 68,53% της συνδιασποράς των μεταβλητών. Το scree plot μας δίνει δύο πιθανές ομάδες παραγόντων που μπορούν να εξαχθούν. Η μία αποτελείται από τρεις παράγοντες και η άλλη από τέσσερις. Καταλήξαμε στη λύση των τριών παραγόντων, όπου εφαρμόσαμε τη μέθοδο ορθογωνίας περιστροφής Varimax, που μεγιστοποιεί τη διακύμανση των τετραγώνων των factor loadings των παραγόντων (στήλες μητρώου) για όλες τις μεταβλητές (γραμμές μητρώου), έτσι ώστε να αποδίδεται, όσο αυτό είναι εφικτό, κάθε μεταβλητή σε ένα μόνο παράγοντα [1, 3]. Τα factor loadings των παραγόντων δίνονται στον πίνακα 2.

Πίνακας 2. Factor loadings παραγόντων σε μητρώο με περιστροφή

Μεταβλητή	Factor 1	Factor 2	Factor 3
Ουίски	0,885		
Ούζο	0,870		
Ελληνική μύρα	0,675	0,329	-0,315
Ρούμι		0,791	
Εισαγόμενη μύρα		0,713	
Gin		0,625	0,542
Εμφ/να & Βερμούτ			0,760
Vodka	0,389		0,716

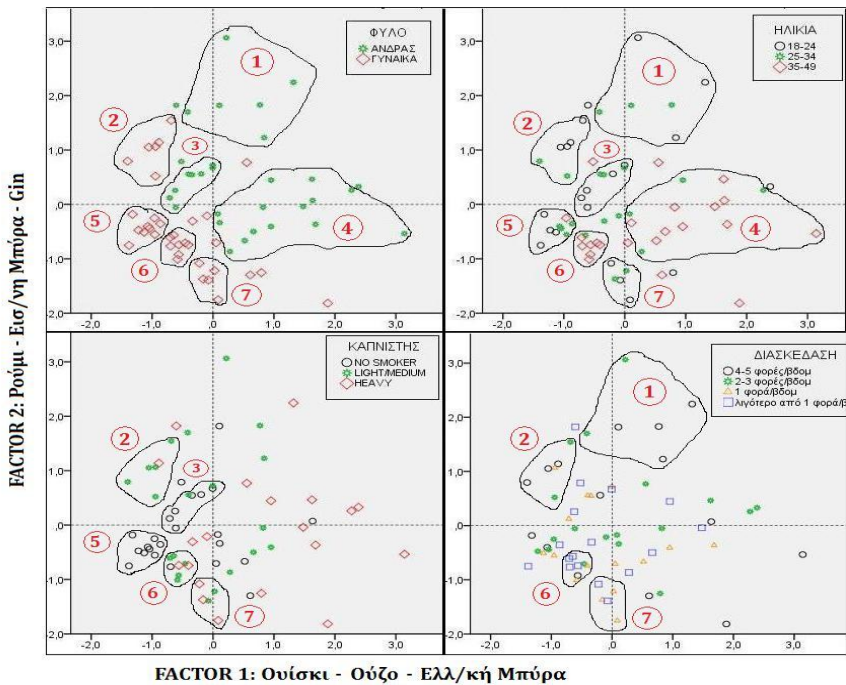
Η επικύρωση των παραπάνω αποτελεσμάτων έγινε σε ένα υποσύνολο του αρχικού δείγματος μεγέθους 1952 ατόμων. Ακολουθώντας την ίδια μεθοδολογία καταλήξαμε και σε αυτή τη περίπτωση σε μητρώο δεδομένων διαστάσεων 66 x 8. Ελέγχθηκε το δείγμα ως προς τις στατιστικές προϋποθέσεις ($KMO = 0,532$ αποδεκτό, Bartlett's test = 88,672, τρεις μεταβλητές με $MSA < 0,5$, communalities $> 0,5$) και επιλέχτηκε σύμφωνα με τα δύο κριτήρια λύση τριών αξόνων με ποσοστό ερμηνείας 59,32%. Εφαρμόστηκε η Varimax περιστροφή και τα factor loadings δίνονται στον πίνακα 3.

Πίνακας 3. Factor loadings παραγόντων για επικύρωση αποτελεσμάτων

Μεταβλητή	Factor 1	Factor 2	Factor 3
Ούζο	0,876		
Ουίски	0,747		
Ελληνική μύρα	0,724		
Εισαγόμενη μύρα		0,719	
Gin		0,717	0,327
Ρούμι		0,558	
Εμφ/να & Βερμούτ			0,737
Vodka	0,333		0,689

Για την εξαγωγή συμπερασμάτων από την παραγοντική λύση που προέκυψε αναλύσαμε τους παράγοντες ανά ζεύγη αναπαριστώντας τους γραφικά. Εντοπίσαμε 7 ομάδες καταναλωτών, εκ των οποίων οι γυναικείες ομάδες εμφανίζουν κατά κύριο λόγο υψηλή κατανάλωση στα ελαφριά ποτά και στη vodka (3^{ος} παράγοντας) και χαμηλή στις υπόλοιπες κατηγορίες (1^ο και 2^ο παράγοντα). Αντιθέτως, οι άντρες καταναλώνουν μεγαλύτερες ποσότητες από τις κατηγορίες των δύο πρώτων παραγόντων, με εξαίρεση μία ομάδα νεαρών αντρών μη καπνιστών που καταναλώνουν ελαφριά ποτά και vodka. Στην εικόνα 3 παρουσιάζονται οι 7 ομάδες καταναλωτών ως προς τους δύο πρώτους παράγοντες.

Εικόνα 3. Ομάδες καταναλωτών ως προς τους παράγοντες 1 και 2



Εφαρμόζοντας την ανάλυση κατά συστάδες, με τις μεθόδους between groups linkage και k-means, στους τρεις παράγοντες καταλήξαμε σε δύο υπέρ-ομάδες καταναλωτών, 44 και 22 παρατηρήσεων. Μέσα στην ομάδα των 44 παρατηρήσεων εντοπίσαμε τις γυναίκες που καταναλώνουν ελαφριά ποτά και vodka και τους νεαρούς άντρες μη καπνιστές που καταναλώνουν και αυτοί τις ίδιες κατηγορίες. Η ομάδα των 22 παρατηρήσεων που προέκυψε είναι οι άντρες καπνιστές που βγαίνουν πάνω από 1 φορά τη βδομάδα και καταναλώνουν ούζο, μπίρα, ούισκι, ρούμι και gin.

2.5 Συμπεράσματα

Σε αυτή την εργασία εφαρμόσαμε μεθόδους Πολυμεταβλητής Στατιστικής Ανάλυσης για να προσεγγίσουμε τις διαφορετικές ομάδες καταναλωτών 9 κατηγοριών αλκοολούχων ποτών ως προς τις συνήθειες τους και τα δημογραφικά χαρακτηριστικά

τους. Από την παραγοντική ανάλυση προέκυψε μία λύση τριών αξόνων που ερμηνεύει το 68,53% της συνολικής πληροφορίας. Τα αποτελέσματα αυτά είναι σταθερά στο συγκεκριμένο δείγμα και μπορούν να γενικευθούν στον πληθυσμό.

Με την παρατήρηση της γραφικής αναπαράστασης των αξόνων και την εφαρμογή της ανάλυσης κατά συστάδες σε αυτούς εντοπίσαμε δύο βασικές ομάδες καταναλωτών, οι οποίες χωρίζονται σε 7 υποομάδες. Τα γενικά συμπεράσματα αυτών των ομάδων είναι ότι η πλειοψηφία των γυναικών δεν καταναλώνει μεγάλες ποσότητες ούζου, ούισκι και ελληνικής μπύρας, αλλά καταναλώνει περισσότερη βότκα και ελαφριά ποτά. Ξεχώρισε μία ομάδα γυναικών κάτω των 35 ετών που καπνίζουν, βγαίνουν πάνω από 2 φορές την εβδομάδα και καταναλώνουν ρούμι, εισαγόμενη μπύρα και gin. Οι μη καπνιστές άντρες κάτω των 35 ετών, που βγαίνουν μέχρι 3 φορές τη βδομάδα, καταναλώνουν βότκα και ελαφριά ποτά. Τέλος, οι καπνιστές άντρες που βγαίνουν τουλάχιστον μία φορά τη βδομάδα είναι καταναλωτές ούζου, μπύρας, gin, ρούμι και ούισκι.

ABSTRACT

The evaluation of the information collected in market research for decision making with as little risk and as much profit as possible creates problems of many kinds which are resolved by applying multivariate statistical methods. In this work we detect and describe in a sample of actual data taken from the Greek population, the weekly consumer of 9 categories of alcoholic beverages, by applying methods of statistical analysis. The data used are intellectual property of Focus Bari SA and are taken from its databases. These databases are created by a nationwide survey conducted annually by Focus Bari SA that records the changes in consumer and lifestyle habits of the Greeks. We initially present the solution method and its various design stages and constraints. We then study and describe our sample via multivariate statistical analysis methods. Finally, we validate our results in terms of the stability that they exhibit in the sample as well as their generalization in the population.

Ευχαριστίες: Οι συγγραφείς ευχαριστούν θερμά τη Διοίκηση της εταιρίας Focus Bari A.E. για την παραχώρηση των δεδομένων που χρησιμοποιήθηκαν στην παρούσα εργασία

ΑΝΑΦΟΡΕΣ

- Anderson R. E., Babin B. J., Black W. C., Hair J. F. Jr (2010). *Multivariate Data Analysis*, 7th edition, Prentice Hall.
- Βασιλικπούλου Α. Ι., Σιώμκος Γ. Ι. (2005). *Εφαρμογή Μεθόδων Ανάλυσης στην Έρευνα Αγοράς*, Αθήνα: εκδόσεις Σταμούλης.
- Garson D. G., *Topics in Multivariate Analysis*, <http://faculty.chass.ncsu.edu/garson>
- Hill T., Lewicki P. (2011). *STATISTICS Methods and Applications*, Tulsa: StatSoft, Inc. Electronic Statistics Textbook
- http://www.utexas.edu/courses/schwab/sw388r7/SolvingProblems/PrincipalComponentAnalysis_Outliers_Validation_Reliability.ppt



ΔΙΕΡΕΥΝΗΣΗ ΤΗΣ ΣΤΑΤΙΣΤΙΚΗΣ ΣΗΜΑΝΤΙΚΟΤΗΤΑΣ ΤΗΣ ΣΥΣΧΕΤΙΣΗΣ ΜΕΤΑΞΥ ΧΡΟΝΟΣΕΙΡΩΝ ΓΙΑ ΤΗ ΔΗΜΙΟΥΡΓΙΑ ΔΙΚΤΥΩΝ

Ε.Σιγγιρίδου¹, Δ.Κουγιουμτζής², Ε.Μπόρα-Σέντα¹

¹ Μαθηματικό Τμήμα, Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης
elsa_siggiridou@hotmail.com, bora@math.auth.gr

² Γενικό Τμήμα, Πολυτεχνική Σχολή, Αριστοτέλειο Πανεπιστήμιο
Θεσσαλονίκης
dkugiu@gen.auth.gr

ΠΕΡΙΛΗΨΗ

Για τη μελέτη πολύπλοκων συστημάτων που παρατηρούνται με πολυδιάστατες χρονοσειρές προτείνεται η δημιουργία δικτύων συσχέτισης με κόμβους τις χρονοσειρές και συνδέσεις τις συσχετίσεις μεταξύ τους. Ως μέτρο συσχέτισης θεωρούμε τη διασυσχέτιση και τη μερική διασυσχέτιση για μηδενική και μη-μηδενική υστέρηση, όπου η μη-μηδενική υστέρηση παραπέμπει σε εκτίμηση αιτιότητας και κατευθυνόμενες συνδέσεις δικτύου. Εξετάζουμε αν η στατιστική σημαντικότητα της συσχέτισης υποδεικνύει σωστά την επιλογή των συνδέσεων και για αυτό χρησιμοποιούμε παραμετρικό έλεγχο και έλεγχο τυχαιοποίησης, συμπεριλαμβανοντας επίσης και τη διόρθωση λόγω πολλαπλών ελέγχων με χρήση του λανθασμένου ποσοστού ανακαλύψεων (false discovery rate). Δημιουργήσαμε δίκτυα συσχέτισης με σταθμισμένες συνδέσεις από τις τιμές των εκτιμητών συσχέτισης, καθώς και με απλές συνδέσεις όπως διαμορφώνονται από τον έλεγχο σημαντικότητας και συγκρίναμε τους δύο τύπους δικτύων με βάση κάποια στατιστικά μέτρα δικτύων (κατανομή βαθμού, συντελεστής σύμπλεξης και μήκος μονοπατιού). Προσομοιώσεις που έγιναν σε γνωστές πολυμεταβλητές στοχαστικές διαδικασίες έδειξαν πως η στατιστική σημαντικότητα δεν αποδίδει σωστά τη δομή του δικτύου, προσθέτοντας συνδέσεις που δεν αντιστοιχούν σε ουσιαστική συσχέτιση ή αιτιότητα. Τέλος, έγινε εφαρμογή των δικτύων συσχέτισης και αιτιότητας στην παγκόσμια χρηματιστηριακή αγορά.

Λέξεις Κλειδιά: διασυσχέτιση, μερική διασυσχέτιση, χρονοσειρές, δίκτυα συσχέτισης.

1. ΕΙΣΑΓΩΓΗ

Το πεδίο της θεωρίας δικτύων και κυρίως των λεγόμενων πολύπλοκων δικτύων γνωρίζει τα τελευταία χρόνια ραγδαία ανάπτυξη με πολλές εφαρμογές, όπως στην οικονομία (Tumminello et al, 2010), κλιματολογία (Donges et al, 2009) και γενετική (Khanin & Wit, 2007). Σε δίκτυα συσχέτισης ή αιτιότητας από πολυμεταβλητές χρονοσειρές, κάθε χρονοσειρά αποτελεί κόμβο δικτύου και η εκτιμώμενη συσχέτιση σε δύο χρονοσειρές αποτελεί τη σύνδεση των

αντίστοιχων κόμβων, ενώ η εκτιμώμενη σχέση αιτιότητας αποτελεί την κατευθυνόμενη σύνδεση. Η επιλογή των συνδέσεων γίνεται συνήθως εμπειρικά με χρήση κατωφλίου ή διατηρώντας ένα προκαθορισμένο ποσοστό των πιο σημαντικών συσχετίσεων (Donges et al, 2009). Σε αυτήν την εργασία χρησιμοποιούμε τη διασυσχέτιση και μερική διασυσχέτιση με μηδενική υστέρηση για αναζήτηση συσχέτισης και με μη-μηδενική υστέρηση για αναζήτηση αιτιότητας. Σκοπός μας είναι να διερευνήσουμε αν η στατιστική σημαντικότητα της συσχέτισης ή αιτιότητας από παραμετρικό έλεγχο ή έλεγχο τυχαιοποίησης σε χρονοσειρές ορίζει τις κατάλληλες συνδέσεις δικτύου, καθώς τέτοια μελέτη δεν έχει γίνει από όσο γνωρίζουμε. Για το σκοπό αυτό χρησιμοποιούμε προσομοιώσεις σε γνωστές πολυδιάστατες στοχαστικές διαδικασίες. Με βάση τα αποτελέσματα των προσομοιώσεων αξιολογούμε τη μορφή του δικτύου συσχέτισης από χρονοσειρές παγκόσμιων χρηματιστηριακών αγορών.

2. ΜΕΘΟΔΟΛΟΓΙΑ

Θεωρούμε τη χρονοσειρά $\{\mathbf{x}_t\}_{t=1}^N = \{x_{1,t}, x_{2,t}, \mathbf{K}, x_{n,t}\}_{t=1}^N$ ως πραγματοποίηση μιας στάσιμης στοχαστικής διαδικασίας διακριτού χρόνου με τυχαία μεταβλητή $\mathbf{X}_t = \{X_{1,t}, X_{2,t}, \mathbf{K}, X_{n,t}\}$ για χρόνο t . Από τα στοιχεία $c_{ji}(\tau)$ του εκτιμητή του πίνακα διασπορών-συνδιασπορών $C(\tau) = 1/(N-\tau) \sum_{t=1}^{N-\tau} (\mathbf{x}_t - \bar{\mathbf{x}})(\mathbf{x}_{t+\tau} - \bar{\mathbf{x}})^T$ για υστέρηση τ , όπου $\bar{\mathbf{x}}$ το διάνυσμα μέσων τιμών, ορίζεται η εκτίμηση της διασυσχέτισης δύο συνιστωσών τυχαίων μεταβλητών $X_{i,t}$, $X_{j,t+\tau}$ ως $r_{ji}(\tau) \equiv r_{X_j X_i}(\tau) = c_{ji}(\tau) / \sqrt{c_{ii}(\tau)c_{jj}(\tau)}$. Η $r_{ji}(\tau)$ για $\tau > 0$ μπορεί να θεωρηθεί και ως δείκτης επίδρασης της X_i στη X_j σε επόμενο χρόνο τ , υποδηλώνοντας αιτιότητα κατά Granger και αντίστροφα για $\tau < 0$ (Wei, 2006, Κεφ 14).

Η διασυσχέτιση δε διακρίνει άμεσες συσχετίσεις μεταξύ των X_i και X_j από έμμεσες συσχετίσεις, π.χ. λόγω της συσχέτισης της κάθε μιας με μια άλλη τυχαία μεταβλητή X_k . Για τον αποκλεισμό των έμμεσων συσχετίσεων των X_i και X_j ως προς όλες τις άλλες συνιστώσες $X_{k_1}, X_{k_2}, \mathbf{K}, X_{k_{n-2}}$ ορίζεται η εκτίμηση της μερικής διασυσχέτισης $\phi_{ji}(\tau) \equiv r_{X_j X_i | S_{n-2}}(\tau)$, όπου $S_{n-2} = \{X_{k_1}, X_{k_2}, \mathbf{K}, X_{k_{n-2}}\}$. Πρώτα υπολογίζονται τα υπόλοιπα της πολλαπλής γραμμικής παλινδρόμησης του $X_{i,t}$ ως προς $X_{k_1,t}, X_{k_2,t}, \mathbf{K}, X_{k_{n-2},t}$ και της $X_{j,t+\tau}$ ως προς $X_{k_1,t}, X_{k_2,t}, \mathbf{K}, X_{k_{n-2},t}$ και στη συνέχεια το $\phi_{ji}(\tau)$ είναι ο συντελεστής συσχέτισης υπολογισμένος στα δείγματα υπολοίπων του $X_{i,t}$ και $X_{j,t+\tau}$ (Kolaczyk, 2009, Κεφ 7, σελ. 207-216).

2.1. ΣΤΑΤΙΣΤΙΚΗ ΣΗΜΑΝΤΙΚΟΤΗΤΑ

Ο έλεγχος σημαντικότητας της διασυσχέτισης r , όπου r είναι $r_{ji}(\tau)$ ή $\phi_{ji}(\tau)$, γίνεται με το Student στατιστικό $t = r\sqrt{(N-2)/(1-r^2)}$ υποθέτοντας ότι το ζεύγος (X_i, X_j) ακολουθεί από κοινού διμεταβλητή κανονική κατανομή. Στην περίπτωση πολλαπλών ελέγχων (για τα διάφορα i και j) μπορεί να γίνει διόρθωση της p -τιμής των ελέγχων με χρήση της διαδικασίας λανθασμένου ποσοστού ανακαλύψεων (false discovery rate, FDR) (Benjamini and Hochberg, 1995). Οι p -τιμές m ελέγχων διατάσσονται σε αύξουσα σειρά $p_{(1)} \leq p_{(2)} \dots \leq p_{(m)}$. Η απόφαση απόρριψης σε επίπεδο σημαντικότητας α ισχύει για όλες τις μηδενικές υποθέσεις που η p -τιμή τους είναι μικρότερη από κάποιο $p_{(i)}$, όπου $p_{(i)}$ είναι η μεγαλύτερη p -τιμή για την οποία ισχύει $p_{(i)} \leq i \cdot \alpha/m$.

Επειδή η υπόθεση κανονικής κατανομής μπορεί πάντα να μην ικανοποιείται, χρησιμοποιούμε επίσης έλεγχο τυχαιοποίησης. Για αυτό δημιουργούνται M χρονοσειρές από επαναδειγματοληψία για κάθε συνιστώσα της $\{x_t\}_{t=1}^N$ μετατοπίζοντας τις αρχικές παρατηρήσεις κάθε χρονοσειράς $\{x_{j,t}\}_{t=1}^N$ με ένα τυχαίο χρονικό βήμα w_j , έτσι ώστε η νέα χρονοσειρά επαναδειγματοληψίας να είναι $\{x_{j,t}^*\} = \{x_{j,w_j+1}, x_{j,w_j+2}, \dots, x_{j,N}, x_{j,1}, \dots, x_{j,w_j}\}$. Με αυτόν τον τρόπο καταστρέφουμε τυχόν διασυσχετίσεις ώστε οι χρονοσειρές να είναι σύμφωνες με τη μηδενική υπόθεση, αλλά διατηρούμε τυχόν αυτοσυσχετίσεις καθώς και την περιθώρια κατανομή της κάθε χρονοσειράς. Υπολογίζουμε το στατιστικό r στο αρχικό ζεύγος χρονοσειρών, έστω r_0 , και σε κάθε ένα από τα M ζεύγη χρονοσειρών επαναδειγματοληψίας, έστω r_1, r_2, \dots, r_M . Αν το r_0 βρίσκεται στις ουρές της εμπειρικής κατανομής που σχηματίζεται από τα r_1, r_2, \dots, r_M απορρίπτεται η μηδενική υπόθεση. Η τάξη του r_0 , έστω i_0 , στο σύνολο των $M+1$ τιμών ορίζει την p -τιμή. Για δίπλευρο έλεγχο αν $i_0 < (M+1)/2$ τότε $p = 2(i_0 - 0.326)/(M+1+0.348)$, αλλιώς $p = 2(1 - (i_0 - 0.326)/(M+1+0.348))$ (Yu and Huang, 2001). Η διόρθωση του FDR γίνεται και στις p -τιμές που παίρνουμε από έλεγχο τυχαιοποίησης. Οι p -τιμές από τον παραμετρικό έλεγχο και τον έλεγχο τυχαιοποίησης, καθώς και οι διορθώσεις τους με το FDR, ορίζουν τη σημαντικότητα των $r_{ji}(\tau)$ ή $\phi_{ji}(\tau)$ για κάποιο α . Έτσι ορίζονται 4 τρόποι σύνδεσης των κόμβων του δικτύου, που συμβολίζονται αντίστοιχα P-VALUE, FDR, P-VALUE-R και FDR-R (με R δηλώνεται η επαναδειγματοληψία).

2.2. ΔΙΚΤΥΑ ΚΑΙ ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ

Η δομή συσχέτισης των χρονοσειρών, που εκφράζει τη δομή αλληλεπίδρασης των μεταβλητών του συστήματος, μπορεί να παρασταθεί με δίκτυο συσχέτισης.

Για τη μελέτη των χαρακτηριστικών του συστήματος που καθορίζονται από τις αλληλεπιδράσεις και αλληλεξαρτήσεις των δομικών μονάδων χρησιμοποιούνται κάποια στατιστικά μέτρα δικτύων. Στον Πίνακα 1 παρουσιάζονται τρία βασικά στατιστικά μέτρα, ορισμένα για κάθε κόμβο δικτύου (Antonioni and Tsompra, 2007). Θα χρησιμοποιήσουμε το μέσο όρο των μέτρων αυτών από όλους τους κόμβους για κάθε δίκτυο που προκύπτει από κάθε πραγματοποίηση.

Πίνακας 1: Στατιστικά μέτρα για κόμβο i μη-σταθμισμένου δικτύου, όπου $a_{ij}=1$ αν ο κόμβος i συνδέεται με τον κόμβο j , αλλιώς $a_{ij}=0$, και σταθμισμένου δικτύου, όπου w_{ij} είναι το βάρος της αντίστοιχης σύνδεσης.

	Βαθμός κόμβου	Μήκος μονοπατιού	Συντελεστής σύμπλεξης κόμβου
Μη σταθμισμένα δίκτυα	$k_i = \sum_j a_{ij}$ πλήθος των κόμβων που συνδέονται με τον κόμβο i	d_{ij} : η συντομότερη διαδρομή από τον κόμβο i στον j	$c_i = \frac{\sum_{j,k} a_{ij} a_{jk} a_{ki}}{\sum_{j,k} a_{ij} a_{ik}}$
Σταθμισμένα δίκτυα	$s_i = \sum_j w_{ij}$ άθροισμα βαρών του κόμβου i	d_{ij} : η συντομότερη διαδρομή από τον κόμβο i στον j με βάση τα βάρη	$c_i = \frac{\sum_j \sum_k (w_{ij} w_{jk} w_{ki})^{1/3}}{k_i (k_i - 1)}$

3. ΠΡΟΣΟΜΟΙΩΣΕΙΣ - ΑΠΟΤΕΛΕΣΜΑΤΑ

3.1. ΑΞΙΟΛΟΓΗΣΗ ΜΕΤΡΩΝ ΔΙΑΣΥΣΧΕΤΙΣΗΣ

Πρώτα ελέγξαμε τη σημαντικότητα και ισχύ των ελέγχων των $r_{ji}(\tau)$ και $\phi_{ji}(\tau)$ με Monte Carlo προσομοιώσεις σε γνωστά συστήματα 5 μεταβλητών. Για $N=1000$ και $N=100$, έγιναν 1000 πραγματοποιήσεις του κάθε συστήματος και υπολογίστηκαν τα $r_{ji}(\tau)$ και $\phi_{ji}(\tau)$ ($\tau = -2, -1, 0, 1, 2$), καθώς και οι 4 p -τιμές.

Για να ελέγξουμε τη σημαντικότητα θεωρήσαμε ένα σύστημα 5 ανεξάρτητων διαδικασιών λευκού θορύβου καθώς και ένα σύστημα 5 ανεξάρτητων αυτοπαλινδρομούμενων διαδικασιών τάξης ένα, AR(1). Και στα δύο συστήματα ο αριθμός των σημαντικών διασυσχετίσεων και μερικών δασυσχετίσεων κυμάνθηκε στο επίπεδο του 5% των 1000 πραγματοποιήσεων. Το μήκος της χρονοσειράς δεν επηρέασε τα αποτελέσματα.

Είναι γνωστό πως σε πολυμεταβλητές χρονοσειρές η ύπαρξη αυτοσυσχέτισης μπορεί να οδηγήσει σε λανθασμένες εκτιμήσεις διασυσχέτισης και για αυτό συνίσταται η απαλοιφή της με τη διαδικασία της λεγόμενης προ-λευκάνσης (Chatfield, 2005, Κεφ. 8). Κάναμε προ-λευκάνση σε όλα τα συστήματα παρακάτω, δηλαδή προσαρμόσαμε σε κάθε χρονοσειρά ξεχωριστά μοντέλο AR(1) και η εκτίμηση διασυσχέτισης έγινε στα υπόλοιπα της προσαρμογής (δεν παρατηρήσαμε αλλαγή στα αποτελέσματα για μεγαλύτερη τάξη του AR).

Για την ισχύ των ελέγχων θεωρήσαμε διανυσματική αυτοπαλινδρομούμενη διαδικασία τάξης ένα σε 5 μεταβλητές, VAR(1), με πίνακα συντελεστών A όπως δίνεται παρακάτω. Επίσης υπολογίσαμε τον κανονικοποιημένο πίνακα συντελεστών B από προσαρμογή μοντέλου VAR(1) σε χρονοσειρές μεγάλου μήκους (μετά από προ-λευκανση) από το σύστημα VAR(1) (που ορίστηκε με τον πίνακα A):

$$A = \begin{pmatrix} 0.8 & -0.5 & 0.3 & 0.9 & 0.1 \\ 0 & -0.7 & -0.2 & 0.8 & 0.7 \\ 0 & 0 & -0.4 & -0.25 & 0.4 \\ 0 & 0 & 0 & 0.5 & -0.4 \\ 0 & 0 & 0 & 0 & 0.6 \end{pmatrix} \quad B = \begin{pmatrix} 0 & -0.1157 & 0.0445 & 0.4018 & 0.0382 \\ 0.0587 & 0 & -0.1399 & 0.4485 & 0.4539 \\ -0.0948 & -0.0172 & 0 & -0.3850 & 0.4647 \\ -0.0401 & -0.1208 & -0.0784 & 0 & -0.3624 \\ 0.0274 & -0.0315 & 0.0516 & 0.0023 & 0 \end{pmatrix}$$

Παρατηρούμε από τον πίνακα B πως η διαδικασία της προ-λευκανσης δημιουργεί μη-μηδενικούς συντελεστές που δεν υπάρχουν στο αρχικό σύστημα, όπως αυτός στη θέση (4,2). Αντίστοιχα δημιουργούνται και μη-μηδενικές διασυσχετίσεις και μερικές διασυσχετίσεις, όπως φαίνεται και από τις μέσες τιμές των $r_{ji}(1)$ και $\phi_{ji}(1)$ από 1000 πραγματοποιήσεις που δίνονται στους πίνακες $R(1)$ και $\Phi(1)$:

$$R(1) = \begin{pmatrix} 0 & -0.3220 & -0.0146 & 0.6061 & 0.0594 \\ 0.1360 & 0 & -0.2434 & 0.4652 & 0.4447 \\ -0.0933 & 0.0125 & 0 & -0.3080 & 0.3322 \\ -0.0071 & -0.1038 & -0.0756 & 0 & -0.3614 \\ 0.0004 & 0.0001 & -0.0023 & 0.0006 & 0 \end{pmatrix} \quad \Phi(1) = \begin{pmatrix} 0 & -0.3162 & -0.0080 & 0.5926 & 0.0702 \\ 0.1332 & 0 & -0.2356 & 0.4597 & 0.4408 \\ -0.0824 & 0.0161 & 0 & -0.2925 & 0.3184 \\ -0.0204 & -0.1133 & -0.0646 & 0 & -0.3460 \\ 0.0146 & 0.0047 & -0.0102 & 0.0109 & 0 \end{pmatrix}$$

Παρόλο που οι τιμές των ψευδών συσχετίσεων είναι μικρές, για μεγάλο μήκος χρονοσειρών μπορεί να είναι σημαντικές, όπως διαπιστώσαμε για $N=1000$ και $\alpha=0.05$. Όπως δίνεται στον Πίνακα 2, για το ζεύγος μεταβλητών $X_{1,t}$ και $X_{3,t+1}$ που δε συσχετίζονται, οι απορρίψεις του ελέγχου σημαντικότητας με το στατιστικό διασυσχέτισης και μερικής διασυσχέτισης στις 1000 επαναλήψεις αυξάνουν δραματικά αλλάζοντας το μήκος χρονοσειρών από 100 σε 1000.

Πίνακας 2: Πλήθος απορρίψεων υπόθεσης σημαντικότητας για $r_{3,1}(1)$ και $\phi_{3,1}(1)$ για τους 4 ελέγχους από 1000 πραγματοποιήσεις του συστήματος VAR(1) για $N=1000$ και σε παρένθεση για $N=100$.

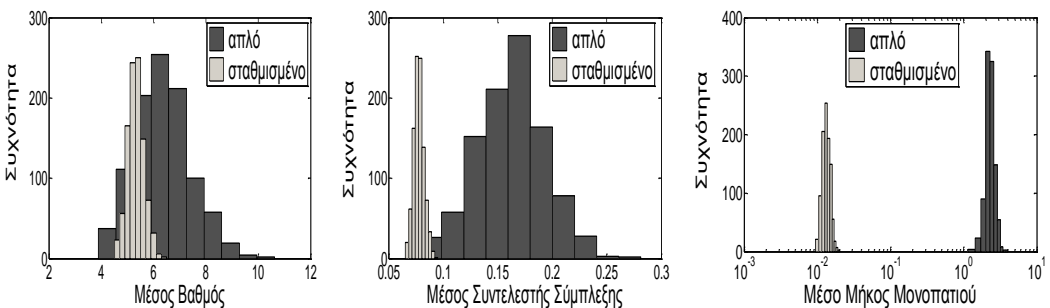
	P-VALUE	FDR	P-VALUE-R	FDR-R
$r_{3,1}(1)$	787 (119)	629 (17)	788 (70)	598 (0)
$\phi_{3,1}(1)$	829 (154)	704 (36)	788 (115)	540 (27)

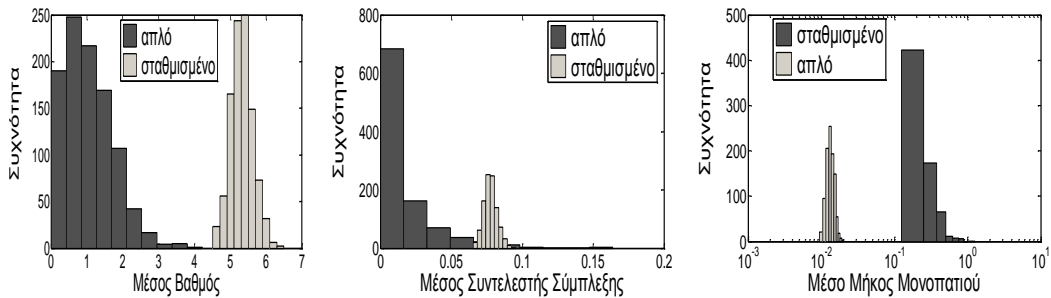
Παρατηρούμε λοιπόν ότι η διαδικασία της προλεύκανσης δεν εξαλείφει το πρόβλημα εμφάνισης ψευδών συσχετίσεων και άρα συνδέσεων στο δίκτυο, ειδικότερα όταν το N είναι μεγάλο.

3.2. ΣΤΑΘΜΙΣΜΕΝΑ ΚΑΙ ΑΠΛΑ ΔΙΚΤΥΑ

Για να προσομοιώσουμε δίκτυα από πολλές μεταβλητές, προσαρμόσαμε μοντέλο VAR(1) σε χρονοσειρές αποδόσεων δεκτών από 55 χρηματιστηριακές αγορές (πηγή: Morgan Stanley Capital International's market, <http://www.msci.com/products/indices>) και μηδενίσαμε όλους τους συντελεστές μικρότερους κατά απόλυτη τιμή του 0.25. Από αυτό το VAR(1) δημιουργήσαμε 1000 πραγματοποιήσεις για $N=1000$ και 100. Για κάθε πραγματοποίηση δημιουργήσαμε σταθμισμένα δίκτυα για $r_{ji}(\tau)$ και $\phi_{ji}(\tau)$ ($\tau = -2,-1,0,1,2$) και τα αντίστοιχα απλά δίκτυα για κάθε έναν από τους 4 στατιστικούς ελέγχους σημαντικότητας, στα οποία υπολογίσαμε τα τρία μέτρα δικτύων. Τα αποτελέσματα έδειξαν πως η κατανομή του κάθε μέτρου δικτύου διαφέρει στο σταθμισμένο και το αντίστοιχο απλό δίκτυο. Όπως φαίνεται στο Σχήμα 1, για $N=100$ και $r_{ji}(1)$, ο έλεγχος P-VALUE έδωσε δίκτυο με πολλές συνδέσεις ενώ ο FDR με λίγες συνδέσεις, με αποτέλεσμα όμως η κατανομή των μέτρων μέσου βαθμού και συντελεστή σύμπλεξης να είναι δεξιά (P-VALUE) ή αριστερά (FDR) από την αντίστοιχη κατανομή για το σταθμισμένο δίκτυο. Οι έλεγχοι τυχαιοποίησης έδωσαν αντίστοιχα μειωμένο αριθμό συνδέσεων χωρίς όμως να προσεγγίζουν καλύτερα τα χαρακτηριστικά του σταθμισμένου δικτύου.

Σχήμα 1: Οι κατανομές του μέσου βαθμού, μέσου συντελεστή σύμπλεξης και μέσου μήκους μονοπατιού σε σταθμισμένα και απλά δίκτυα για $N=100$ με βάση το στατιστικό $r_{ji}(1)$ και έλεγχο P-VALUE (σειρά 1), FDR (σειρά 2).



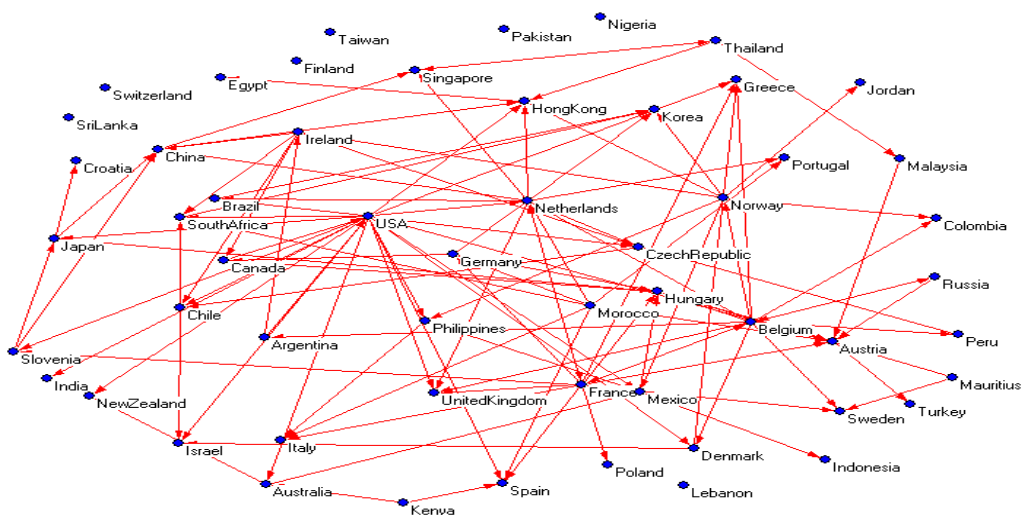


Οι αποκλίσεις των κατανομών επιβεβαιώθηκαν με έλεγχο προσαρμογής κατανομής Kolmogorov-Smirnov για κάθε υστέρηση τ των $r_{ji}(\tau)$ και $\phi_{ji}(\tau)$. Μάλιστα για $N=1000$ οι αποκλίσεις ήταν μεγαλύτερες, αφού τα απλά δίκτυα είχαν περισσότερες συνδέσεις, δηλαδή στατιστικά σημαντικές διασυσχετίσεις.

4. ΕΦΑΡΜΟΓΗ

Εφαρμόσαμε τα $r_{ji}(\tau)$ και $\phi_{ji}(\tau)$ ($\tau=-2,-1,0,1,2$), καθώς και τους 4 ελέγχους σημαντικότητας στα προαναφερθέντα δεδομένα των 55 χρηματιστηριακών αγορών. Για το $r_{ji}(\tau)$ όλα τα δίκτυα ήταν σχεδόν πλήρη για υστέρησεις $-1,0,1$ και με λιγότερες συνδέσεις για $-2,2$ και δεν επέτρεπαν την εξαγωγή χρήσιμων συμπερασμάτων. Για το $\phi_{ji}(\tau)$ ο έλεγχος τυχαιοποίησης P-VALUE-R ήταν πιο αυστηρός και έδινε απλά δίκτυα με λιγότερες συνδέσεις. Μάλιστα με τον έλεγχο του FDR-R όλοι οι κόμβοι ήταν ασύνδετοι μεταξύ τους. Στο Σχήμα 2 δίνεται το δίκτυο αιτιότητας που προέκυψε για υστέρηση 1 από τον P-VALUE-R έλεγχο.

Σχήμα 2: Δίκτυο για την μερική διασυσχέτιση για υστέρηση 1 με την χρήση p -τιμών από επαναδειγματοληψία για $\alpha=0.005$.



Παρατηρούμε ότι οι χρηματιστηριακές αγορές των Η.Π.Α. παίζουν κυρίαρχο ρόλο στο δίκτυο καθώς επηρεάζουν πολλές χώρες αλλά και επηρεάζονται από τις αγορές του Καναδά και της Αργεντινής. Η χρηματιστηριακή αγορά της Ελλάδας φαίνεται να επηρεάζεται από του Βελγίου, των Η.Π.Α., της Γαλλίας και της Δανίας. Πολλές νέες αγορές, κυρίως της Ασίας και της Αφρικής, είναι απομονωμένες και είναι αξιοσημείωτο ότι αυτό παρατηρείται και για την Ελβετία.

5. ΣΥΜΠΕΡΑΣΜΑΤΑ

Το πρώτο βασικό συμπέρασμα της εργασίας είναι ότι η διαδικασία της προ-λεύκανσης μπορεί να διώχνει τις αυτοσυσχετίσεις αλλά δημιουργεί κάποιες ψευδείς διασυσχετίσεις. Επίσης ο έλεγχος τυχαιοποίησης προέκυψε πιο αυστηρός από τον αντίστοιχο παραμετρικό έλεγχο, που τείνει να δίνει πολλές συνδέσεις και σχεδόν πλήρη δίκτυα. Αντίθετα ο έλεγχος με τη χρήση του FDR μειώνει το πλήθος σημαντικών συσχετίσεων καθώς αυξάνει το πλήθος των μεταβλητών και για αυτό δεν είναι χρήσιμος στη δημιουργία απλών δικτύων με πολλούς κόμβους. Άρα, παρά το μεγάλο υπολογιστικό κόστος στον υπολογισμό των p -τιμών, ο έλεγχος τυχαιοποίησης είναι ο πιο κατάλληλος για την κατασκευή δικτύων.

ABSTRACT

A novel way to study complex systems observed by multivariate time-series is by means of correlation networks, where the time-series form the nodes and their correlation form the connections. For the correlation, we consider the cross-correlation and the partial cross-correlation, where for non-zero lag the measures estimate causal relationships and imply directed network connections. We investigate whether parametric and randomization tests for the correlation significance identify correctly the true connections. We consider also the correction for multiple testing using the false discovery rate. We formed correlation networks with weighted connections defined directly by the correlation estimates, as well as zero-one connections determined by the significance tests, and compared the two network types in terms of standard network measures (degree distribution, cluster coefficient and shortest path). Simulations on some known multivariate stochastic processes showed that statistical significance doesn't ascribe correctly the network structure, adding connections that do not correspond to substantial correlation or causality. Finally, the significance tests and the correlation and causality networks were applied to indices of the global stock market.

ΑΝΑΦΟΡΕΣ

- Antoniou I.E. and Tsompa E.T. (2007). Statistical analysis of weighted networks, *Discrete Dynamics in Nature and Society*, **2008**, 375452
- Benjamini Y. and Hochberg Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society. Series B*, **57**, No. 1, 289-300
- Chatfield C. (2005). *The Analysis of Time Series: An Introduction*, Sixth Edition, Chapman & Hall.

- Donges J.F., Zou Y., Marwan N. and Kurths J. (2009). The backbone of the climate network, *Europhysics Letters*, **87**, 48007.
- Khanin R. and Wit E. (2007). Construction of malaria gene expression network using partial correlations, *Methods of Microarray Data Analysis V*, Springer, 75-88
- Kolaczyk D. (2009). *Statistical Analysis of Network Data*, Springer
- Tumminello M, Lillo F and Mantegna R.M. (2010) Correlation, hierarchies, and networks in financial markets, *Journal of Economic Behavior & Organization*, **75**, No 1, 40-58
- Wei W.(2006). *Time Series Analysis: Univariate and Multivariate Methods*, Addison-Wesley
- Yu G.-H. and Huang C.-C. (2001) A distribution free plotting position, *Stochastic Environmental Research and Risk Assessment*, **15**, 462-476



ΜΕΛΕΤΗ ΙΔΙΟΤΗΤΩΝ ΜΙΑΣ ΓΕΝΙΚΗΣ ΚΛΑΣΗΣ ΤΥΧΑΙΩΝ ΜΕΤΑΒΛΗΤΩΝ ΠΟΥ ΣΧΕΤΙΖΟΝΤΑΙ ΜΕ ΚΑΤΑΝΟΜΕΣ ΔΙΑΤΕΤΑΓΜΕΝΩΝ ΠΑΡΑΤΗΡΗΣΕΩΝ

I. Σ. Τριανταφύλλου¹, M. B. Κούτρας¹

¹Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης, Πανεπιστήμιο Πειραιώς
itrantal@unipi.gr, mkoutras@unipi.gr

ΠΕΡΙΛΗΨΗ

Στην εργασία του Jones (2004), ορίζεται μία νέα οικογένεια κατανομών η οποία προκύπτει γενικεύοντας τον τύπο της συνάρτησης πυκνότητας των διατεταγμένων παρατηρήσεων ενός τυχαίου δείγματος που προέρχεται από μία κατανομή F . Πιο συγκεκριμένα, θεμελιώνεται μία νέα κλάση κατανομών, με τη χρησιμοποίηση μίας αρχικής κατανομής F και δύο πρόσθετων παραμέτρων που σχετίζονται με τη λοξότητα και τις ουρές των νέων αυτών κατανομών.

Τα τελευταία χρόνια έχουν δημοσιευθεί ποικίλα αποτελέσματα που αφορούν διάφορες ειδικές περιπτώσεις της οικογένειας κατανομών του Jones (2004), για την οποία χρησιμοποιείται συνήθως η ονομασία beta-generated κλάση κατανομών. Για παράδειγμα, οι Nadarajah & Kotz (2006) ασχολήθηκαν με μία γενίκευση της Εκθετικής κατανομής που ανήκει στην προαναφερθείσα κλάση ενώ ο Zografos (2008) έδωσε αρκετά αποτελέσματα, που σχετίζονται κυρίως με μέτρα πληροφορίας, για διάφορες κατανομές της κλάσης των beta-generated κατανομών και μελέτησε διεξοδικά τις ιδιότητες της κατανομής beta-Weibull.

Στην παρούσα εργασία, αποδεικνύονται γενικά αποτελέσματα που αφορούν τη μονοτονία του ρυθμού αποτυχίας της κλάσης των beta-generated κατανομών, και δίνονται κλειστοί τύποι υπολογισμού ροπών πρώτης τυχαίων μεταβλητών που σχετίζονται με τις κατανομές αυτές.

1. ΕΙΣΑΓΩΓΗ

Η γέννηση νέων κλάσεων κατανομών με στόχο την καλύτερη εφαρμογή σε αληθινά δεδομένα, αποτελεί ένα ιδιαίτερα σημαντικό και ενδιαφέρον αντικείμενο μελέτης στο χώρο των Πιθανοτήτων και της Στατιστικής. Στα πλαίσια αυτά, έχει αναπτυχθεί τα τελευταία χρόνια μία έντονη ερευνητική δραστηριότητα σχετικά με τον ορισμό νέων οικογενειών κατανομών που προκύπτουν με λογαριθμικό μετασχηματισμό τυχαίων μεταβλητών που ακολουθούν την κατανομή Βήτα. Πιο συγκεκριμένα, ο Jones (2004) όρισε μία νέα οικογένεια κατανομών η οποία προκύπτει γενικεύοντας τον τύπο της Η εργασία χρηματοδοτήθηκε από το Ίδρυμα Κρατικών Υποτροφιών (I.K.Y.)

συνάρτησης πυκνότητας των διατεταγμένων παρατηρήσεων ενός τυχαίου δείγματος που προέρχεται από μία κατανομή F . Τα συστατικά της νέας αυτής κλάσης κατανομών περιλαμβάνουν μία συνεχή κατανομή F με συνάρτηση πυκνότητας f και δύο πραγματικές, θετικές παραμέτρους a, b . Λέμε ότι η τυχαία μεταβλητή $T = X_F^{(a,b)}$ ακολουθεί μία beta-generated κατανομή, όταν η συνάρτηση πυκνότητας της T δίνεται από τον τύπο:

$$g_B(t) = g_F(t; a, b) = \frac{1}{B(a, b)} (F(t))^{a-1} (1 - F(t))^{b-1} f(t), \quad t \in R, \quad a, b > 0,$$

όπου F είναι η αρχική κατανομή και $B(\cdot, \cdot)$ η γνωστή συνάρτηση Βήτα. Η αθροιστική συνάρτηση κατανομής της T δίνεται ως ακολούθως

$$G_B(t) = G_F(t; a, b) = \int_{-\infty}^t g_B(s) ds = \frac{B_{F(t)}(a, b)}{B(a, b)} = I_{F(t)}(a, b),$$

όπου οι συναρτήσεις

$$B_x(a, b) = \int_0^x s^{a-1} (1-s)^{b-1} ds, \quad 0 \leq x \leq 1,$$

$$I_x(a, b) = \frac{B_x(a, b)}{B(a, b)}, \quad 0 \leq x \leq 1$$

είναι γνωστές ως μη πλήρη συνάρτηση Βήτα και μη πλήρη αναλογία Βήτα αντίστοιχα.

Είναι σαφές ότι χρησιμοποιώντας μία διαφορετική κάθε φορά αρχική κατανομή F προκύπτει με τη βοήθεια των παραπάνω τύπων, μία νέα κατανομή που ανήκει στην κλάση των beta-generated κατανομών. Για παράδειγμα, οι Nadarajah & Kotz (2006) θεμελίωσαν μία γενίκευση της Εκθετικής κατανομής γνωστή ως beta-exponential κατανομή, ενώ οι Akinsete *et al.* (2008) εισήγαγαν την beta-Pareto κατανομή.

2. ΡΟΠΕΣ ΜΙΑΣ ΒΕΤΑ-GENERATED ΤΥΧΑΙΑΣ ΜΕΤΑΒΛΗΤΗΣ

Στη συγκεκριμένη παράγραφο θα αποδείξουμε αποτελέσματα που σχετίζονται με τις ροπές πρώτης και δεύτερης τάξης μίας τυχαίας μεταβλητής T που ακολουθεί κατανομή που ανήκει στην beta-generated οικογένεια κατανομών.

Στα πλαίσια της παρούσας εργασίας, ας συμβολίσουμε τη δίγαμμα συνάρτηση ως εξής:

$$\psi(x) = \frac{d}{dx} \ln \Gamma(x).$$

Πρόταση 2.1. Αν $T = X_F^{(a,b)}$ είναι μια beta-generated τυχαία μεταβλητή με αθροιστική συνάρτηση κατανομής G_B , τότε

$$(i) E(\ln F(T)) = \psi(a) - \psi(a + b)$$

$$(ii) E(\ln(1 - F(T))) = \psi(b) - \psi(a + b).$$

Απόδειξη. (i) Ας συμβολίσουμε με f_Y the τη συνάρτηση πυκνότητας της τυχαίας μεταβλητής $Y = F(T)$. Συνεπώς ισχύει

$$E(\ln Y) = \int_{-\infty}^{\infty} \ln t f_Y(t) dt.$$

Η αθροιστική συνάρτηση κατανομής της Y can μπορεί να εκφρασθεί ως

$$F_Y(t) = P(F(T) \leq t) = P(T \leq F^{-1}(t)) = \int_0^{F^{-1}(t)} g_B(x) dx.$$

Υπολογίζοντας την πρώτη παράγωγο της τελευταίας συνάρτησης και αντικαθιστώντας αυτήν στην παραπάνω ισότητα, προκύπτει ότι

$$E(\ln Y) = = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 t^{a-1} (1-t)^{b-1} \ln t dt. \quad (2.1)$$

Δεδομένου ότι

$$f(t) = \frac{1}{B(a,b)} t^{a-1} (1-t)^{b-1}, \quad 0 \leq t \leq 1$$

είναι η γνωστή συνάρτηση πυκνότητας της κατανομής Beta με παραμέτρους a, b , τότε

$$\frac{1}{B(a,b)} \int_0^1 t^{a-1} (1-t)^{b-1} dt = 1$$

ή ισοδύναμα

$$\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 t^{a-1} (1-t)^{b-1} dt = 1.$$

Συνεπώς,

$$\frac{\Gamma(a+b)}{\Gamma(b)} \int_0^1 t^{a-1} (1-t)^{b-1} dt = \Gamma(a)$$

οπότε, υπολογίζοντας την πρώτη παράγωγο ως προς a , συμπεραίνουμε ότι

$$\frac{\Gamma'(a+b)}{\Gamma(b)} \int_0^1 t^{a-1} (1-t)^{b-1} dt + \frac{\Gamma(a+b)}{\Gamma(b)} \int_0^1 (t^{a-1} \ln t) (1-t)^{b-1} dt = \Gamma'(a). \quad (2.2)$$

Αν λάβουμε υπόψιν μας ότι η ροπή πρώτης τάξης της τυχαίας μεταβλητής $\ln F(T)$, εκφράζεται μέσω της (2.1), τότε η ισότητα (2.2) μπορεί να γραφεί ως

$$\frac{\Gamma'(a+b)\Gamma(a)}{\Gamma(a+b)} + \Gamma(a)E(\ln F(T)) = \Gamma'(a)$$

και η απόδειξη ολοκληρώνεται.

(ii) Η απόδειξη του μέρους (ii) προκύπτει ανάλογα.

Πρόταση 2.2. Αν $T = X_F^{(a,b)}$ είναι μία beta-generated τυχαία μεταβλητή με αθροιστική συνάρτηση κατανομής G_B , τότε

$$(i) V(\ln(F(T))) = \psi'(a) - \psi'(a+b)$$

$$(ii) V(\ln(1-F(T))) = \psi'(b) - \psi'(a+b).$$

Απόδειξη. (i) Παραγωγίζοντας ως προς a την (2.2), έχουμε:

$$\begin{aligned} & \frac{\Gamma''(a+b)}{\Gamma(b)} \int_0^1 t^{a-1} (1-t)^{b-1} dt + \frac{\Gamma'(a+b)}{\Gamma(b)} \int_0^1 (t^{a-1} \ln t)(1-t)^{b-1} dt \\ & + \frac{\Gamma'(a+b)}{\Gamma(b)} \int_0^1 t^{a-1} \ln t (1-t)^{b-2} dt + \frac{\Gamma(a+b)}{\Gamma(b)} \int_0^1 t^{a-1} (\ln t)^2 (1-t)^{b-2} dt = \Gamma''(a). \end{aligned} \quad (2.3)$$

Συνεπώς, έπειτα από αλγεβρικές πράξεις, προκύπτει ότι

$$E(\ln^2 F(T)) = \frac{\Gamma''(a)}{\Gamma(a)} - \frac{\Gamma''(a+b)}{\Gamma(a+b)} - 2 \frac{\Gamma'(a+b)}{\Gamma(a+b)} E(\ln F(T)).$$

Συνεπώς,

$$\begin{aligned} V(\ln F(T)) &= E(\ln^2 F(T)) - (E(\ln F(T)))^2 \\ &= \psi'(a) + (\psi(a))^2 - (\psi'(a+b) + (\psi(a+b))^2) \\ &\quad - 2\psi(a+b)(\psi(a) - \psi(a+b)) - (\psi(a) - \psi(a+b))^2 \end{aligned}$$

και η απόδειξη ολοκληρώθηκε

(ii) Η απόδειξη του μέρους (ii) προκύπτει ανάλογα.

Για παράδειγμα, αν T ακολουθεί την beta-εκθετική κατανομή με παραμέτρους a, b, λ (βλ. Nadarajah & Kotz (2006)), τότε η αθροιστική συνάρτηση κατανομής της αρχικής δίνεται ως

$$F(t) = 1 - e^{-\lambda t}, \quad t > 0$$

ή ισοδύναμα μπορούμε να γράψουμε ότι

$$\ln(1-F(t)) = -\lambda t.$$

Συνεπώς, με τη βοήθεια της Πρότασης 2.1, προκύπτει ότι

$$-\lambda E(T) = \psi(b) - \psi(a+b)$$

ή

$$E(T) = \frac{\psi(a+b) - \psi(b)}{\lambda}.$$

Πρόσθετα, χρησιμοποιώντας την Πρόταση 2.2, προκύπτει ότι

$$\lambda^2 V(T) = \psi'(b) - \psi'(a+b)$$

ή ισοδύναμα

$$V(T) = \frac{\psi'(b) - \psi'(a+b)}{\lambda^2}.$$

3. ΡΥΘΜΟΣ ΑΠΟΤΥΧΙΑΣ ΜΙΑΣ ΒΕΤΑ-GENERATED ΤΥΧΑΙΑΣ ΜΕΤΑΒΛΗΤΗΣ

Στην παράγραφο αυτή θα μελετήσουμε το ρυθμό αποτυχίας μιας τυχαίας μεταβλητής που ακολουθεί beta-generated κατανομή. Πιο συγκεκριμένα, θα θεμελιώσουμε μία μέθοδο που υπολογίζει το ρυθμό αποτυχίας μιας beta-generated κατανομής, όταν ο ρυθμός αποτυχίας της αρχικής κατανομής είναι γνωστός.

Πρόταση 3.1. Έστω

$$r(x) = \frac{f(x)}{1-F(x)} = \frac{d}{dx}(-\ln(1-F(x)))$$

ο ρυθμός αποτυχίας της αρχικής κατανομής. Τότε ο ρυθμός αποτυχίας της $X_F^{(a,b)}$

$$r_B(t) = r_F(t; a, b) = \frac{g_B(t)}{1-G_B(t)} = \frac{d}{dt}(-\ln(1-G_B(t)))$$

μπορεί να εκφρασθεί ως ακολούθως

$$r_B(t) = \frac{r(t)}{h(F(t); a, b)} \quad (3.1)$$

όπου

$$h(x; a, b) = \int_0^1 v^{b-1} \left(v + \frac{1-v}{x}\right)^{a-1} dv. \quad (3.2)$$

Απόδειξη. Χρησιμοποιώντας την προφανή ταυτότητα

$$I_x(a, b) = 1 - I_{1-x}(b, a),$$

προκύπτει ότι

$$\bar{G}_B(t) = 1 - G_B(t) = 1 - I_{F(t)}(a, b) = I_{\bar{F}(t)}(b, a)$$

και ο ρυθμός αποτυχίας $r_B(t)$ παίρνει την ακόλουθη μορφή

$$r_B(t) = \frac{g_B(t)}{\bar{G}_B(t)} = \frac{(F(t))^{a-1} (1-F(t))^{b-1} f(t)}{\int_0^{\bar{F}(t)} s^{b-1} (1-s)^{a-1} ds} = \frac{(F(t))^{a-1} (\bar{F}(t))^b}{\int_0^{\bar{F}(t)} s^{b-1} (1-s)^{a-1} ds} \cdot r(t).$$

Το ολοκλήρωμα στον παρονομαστή, εφαρμόζοντας το μετασχηματισμό $s = \bar{F}(t)v$, εκφράζεται ως

$$\int_0^{\bar{F}(t)} s^{b-1} (1-s)^{a-1} ds = (\bar{F}(t))^b \int_0^1 v^{b-1} (1-\bar{F}(t)v)^{a-1} dv$$

$$= (\bar{F}(t))^b (F(t))^{a-1} \int_0^1 v^{b-1} \left(\frac{1-\bar{F}(t)v}{F(t)}\right)^{a-1} dv$$

Συνεπώς,

$$r_B(t) = \frac{r(t)}{\int_0^1 v^{b-1} \left(\frac{1-v+F(t)v}{F(t)}\right)^{a-1} dv} = \frac{r(t)}{\int_0^1 v^{b-1} \left(v + \frac{1-v}{F(t)}\right)^{a-1} dv},$$

και η απόδειξη ολοκληρώνεται.

Στη συνέχεια, θα εξετάσουμε αν η συμπεριφορά του ρυθμού αποτυχίας της αρχικής κατανομής διατηρείται στην αντίστοιχη beta-generated κατανομή.

Πρόταση 3.2. Η μονοτονία του ρυθμού αποτυχίας της κατανομής της τυχαίας μεταβλητής $T = X_F^{(a,b)}$ εξαρτάται μόνο από το a και τη μονοτονία του ρυθμού αποτυχίας της F . Πιο συγκεκριμένα,

- (i) αν $a > 1$ και η τυχαία μεταβλητή X έχει αύξων ρυθμό αποτυχίας ($F \in IFR$), τότε η $X_F^{(a,b)}$ έχει αύξων ρυθμό αποτυχίας,
- (ii) αν $0 < a < 1$ και η τυχαία μεταβλητή X έχει φθίνων ρυθμό αποτυχίας ($F \in DFR$), τότε η $X_F^{(a,b)}$ έχει φθίνων ρυθμό αποτυχίας
- (iii) αν $a = 1$ τότε

$$r_B(t) = br(t).$$

Απόδειξη. Είναι σαφές ότι η συνάρτηση $h(x; a, b)$, όπως αυτή ορίστηκε στην (3.2), είναι φθίνουσα όταν $a > 1$. Δεδομένου ότι η F είναι αύξουσα, ο παρονομαστής της (3.1) θα είναι αύξουσα συνάρτηση ως προς t αν $a > 1$ και φθίνουσα αν $0 < a < 1$.

- (i) Αν $a > 1$ και $F \in IFR$, ο αριθμητής της (3.1) είναι αύξουσα συνάρτηση ως προς t και ο παρονομαστής είναι φθίνουσα συνάρτηση ως προς t . Συνεπώς ο ρυθμός αποτυχίας $r_B(t)$ είναι αύξουσα συνάρτηση ως προς t , που αποδεικνύει ότι $X_F^{(a,b)} \in IFR$.
- (ii) Αν $0 < a < 1$ και $F \in DFR$, ο αριθμητής της (3.1) είναι φθίνουσα συνάρτηση ως προς t και ο παρονομαστής είναι αύξουσα συνάρτηση ως προς t . Συνεπώς, ο ρυθμός αποτυχίας $r_B(t)$ είναι φθίνουσα συνάρτηση ως προς t , που αποδεικνύει ότι $X_F^{(a,b)} \in DFR$.
- (iii) Για $a = 1$ προκύπτει ότι

$$h(x; 1, b) = \int_0^1 v^{b-1} dv = 1/b$$

για όλα τα $x \in IR$ and και το αποτέλεσμα προκύπτει άμεσα με τη βοήθεια

της σχέσης (3.1).

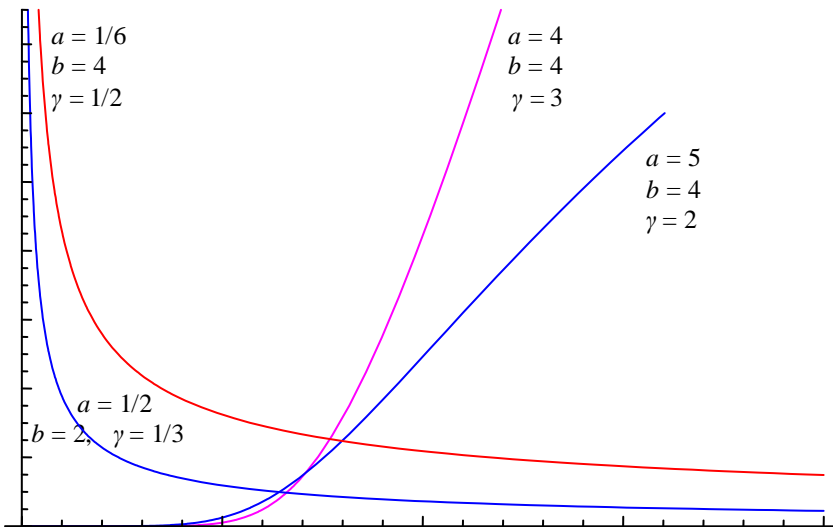
Για παράδειγμα, αν $F(t) = 1 - \exp(-\lambda t^\gamma)$, $t > 0$ (κατανομή Weibull), γνωρίζουμε ότι η F ανήκει στην κλάση κατανομών *IFR* για $\gamma > 1$ και στην *DFR* για $0 < \gamma < 1$. Συνεπώς, η συνάρτηση πυκνότητας πιθανότητας της beta-Weibull είναι:

$$g_B(t) = \frac{1}{B(a,b)} \lambda \gamma t^{\gamma-1} (1 - e^{-\lambda t^\gamma})^{a-1} e^{-\lambda(b-1)t - \lambda t^\gamma}, \quad t > 0$$

Αποτελεί μία οικογένεια κατανομών με τέσσερις παραμέτρους που ανήκει

- α. στην κλάση *IFR*, για $\gamma > 1$ και $a > 1$,
- β. στην κλάση *DFR*, για $0 < \gamma < 1$ και $0 < a < 1$.

Το Σχήμα 3.1 παρουσιάζει τη μορφή του ρυθμού αποτυχίας της beta-Weibull κατανομής για $\lambda = 1$ και για διάφορες τιμές των παραμέτρων a, b, γ που οδηγούν σε *IFR/DFR* κατανομές.



Σχήμα 3.1. Ρυθμός αποτυχίας της beta-Weibull κατανομής με $\lambda = 1$ και διάφορες επιλογές των παραμέτρων a, b, γ που καταλήγουν σε μονότονες συναρτήσεις ρυθμού αποτυχίας.

ABSTRACT

In the present article we study a new family of distributions, which was first introduced by Jones (2004) and is related to the probability density function of order statistics drawn from a random sample. More specifically, we prove general results for the monotonicity of the failure rate of the aforementioned class of distributions (known as beta-generated), while closed formulas for the evaluation of the moments of random variables, related to the beta-generated distributions, are established.

ΑΝΑΦΟΡΕΣ

- Akinsete, A., Famoye, F. & Lee, C. (2008). The beta-Pareto distribution, *Statistics*, **42**, 547-563.
- Jones, M. C. (2004). Families of distributions arising from distributions of order statistics, *Test*, **13**, 1-43.
- Nadarajah, S. & Kotz, S. (2006). The beta exponential distribution, *Reliability Engineering and System Safety*, **91**, 689-697.
- Zografos, K. (2008). On some beta generated distributions and their maximum entropy characterization: the beta-Weibull distribution, In *Advances in Inequalities from Probability Theory & Statistics* (Eds. Barnett, N. S. & Dragomir, S. S.), 203-225.



ΔΕΙΓΜΑΤΟΛΗΨΙΑ ΣΤΟ ΔΙΑΔΙΚΤΥΟ: ΔΥΝΑΤΟΤΗΤΕΣ ΑΝΑΚΤΗΣΗΣ ΣΥΓΚΕΚΡΙΜΕΝΗΣ ΠΛΗΡΟΦΟΡΙΑΣ

Νικόλαος Φαρμάκης¹, Γεώργιος Μακρής²

Τμήμα Μαθηματικών, ΑΠΘ

¹farmakis@math.auth.gr, ²geormak@hotmail.com

ΠΕΡΙΛΗΨΗ

Η ανάπτυξη του παγκόσμιου ιστού είχε ως αποτέλεσμα την δημιουργία μιας τεράστιας αποθήκης δεδομένων και πληροφοριών, με ελεύθερη πρόσβαση στην πληροφορία αυτή.

Στην παρούσα εργασία γίνεται έρευνα για το χρόνο που καταναλώνει ένας χρήστης για την αναζήτηση πληροφορίας μέσω του διαδικτύου. Μαθητές Λυκείου στο πλαίσιο του μαθήματος Πληροφορικής αναζητούν πληροφορίες με χρήση μηχανών αναζήτησης για συγκεκριμένα θέματα και με βάση συγκεκριμένο δειγματοληπτικό σχέδιο. Καταγράφονται οι χρόνοι που χρειάστηκαν και οι ενέργειες που έγιναν από τους χρήστες για την εύρεση συγκεκριμένης πληροφορίας. Διερευνάται *το αν και το κατά πόσο*, με τις υπάρχουσες τεχνολογίες, οι μηχανές αναζήτησης είναι σε θέση να απαντήσουν σε ερωτήσεις του τύπου:

«Πότε κτίστηκε ο Παρθενώνας ;»,

«Ο κύριος Γεώργιος Μακρής σε ποιο Λύκειο διδάσκει;», κλπ.

Σε κάθε ερώτηση σημειώθηκε ο χρόνος που διέρρευσε μέχρι να βρεθεί η απάντηση. Αυτή ήταν η 1^η παράμετρος που μελετήθηκε ως συνεχής τυχαία μεταβλητή (τ.μ.). Επίσης μετρήθηκαν τα βήματα (χτυπήματα, clicks) που χρειάστηκαν. Αυτή ήταν μία άλλη τ.μ. που προσομοίαζε με συνεχή τ.μ. (τουλάχιστο όταν έχουμε πάνω από 8 επίπεδα). Για τις τ. μ. επιχειρήθηκε εκτίμηση για την μορφή της συνάρτησης πυκνότητας πιθανότητας μέσω μιας μεθόδου που στηρίζεται στην έννοια του Συντελεστή Μεταβλητότητας.

Το βασικά ερωτήματα που προσπαθεί να αντιμετωπίσει η έρευνά μας είναι:

- 1ο) Πόσο χρόνο κάνει να βρεθεί απάντηση στο διαδίκτυο σε ερωτήσεις όπως οι παραπάνω;
- 2ο) Πόσες κινήσεις χρειάζεται;
- 3ο) Μπορεί να υπάρξει άμεσα απάντηση;

Λέξεις Κλειδιά: Διαδίκτυο, Πληροφορία, Δειγματοληψία, Μηχανή Αναζήτησης.

1. ΕΙΣΑΓΩΓΗ

Η ανάπτυξη του παγκόσμιου ιστού είχε ως αποτέλεσμα την δημιουργία μιας τεράστιας αποθήκης δεδομένων και πληροφοριών. Η πρόσβαση στην πληροφορία

αυτή είναι ελεύθερη για τον κάθε χρήστη, τουλάχιστον σε πολύ μεγάλο βαθμό. Φυσικά η αναζήτηση της πληροφορίας γίνεται μέσω των διαφόρων μηχανών αναζήτησης, όπως είναι οι google, bing, yahoo, κλπ. Οι χρήστες αφιερώνουν αρκετό χρόνο για την αναζήτηση της πληροφορίας που θέλουν. Έτσι ανακύπτει το ερώτημα αν και κατά πόσο ο χρήστης πετυχαίνει να πάρει άμεσα την απάντηση στην ερώτησή του ή την παίρνει έμμεσα. Επίσης ενδιαφέρει και μία εκτίμηση της πιθανότητας να έχουμε περιπτώσεις άμεσης απάντησης σε ερωτήματα όπως και σε τι είδους ερωτήματα είναι μη μηδενική αυτή η πιθανότητα άμεσης απάντησης. Στην παρούσα εργασία προσπαθούμε να απαντήσουμε σε σχετικά με τα παραπάνω ερωτήματα μέσα από δειγματοληψία.

Ένα δείγμα $n=50$ μαθητών Γενικού Λυκείου εξετάστηκε καθώς προσπαθούσε να βρει πληροφορίες μέσω του παγκόσμιου Ιστού, χρησιμοποιώντας μηχανές αναζήτησης όπως: Google, Bing, Yahoo, κλπ, Kolaczyk (2009). Προσπαθούν να βρουν απαντήσεις στις παρακάτω πέντε ερωτήσεις:

1. Ποια είναι η διεύθυνση του τμήματος Μαθηματικών του ΑΠΘ;
2. Ποιο είναι το e-mail του καθηγητή σας του μαθήματος Πληροφορικής;
3. Ποια χρονιά κτίστηκε ο Παρθενώνας;
4. Είναι παντρεμένος ή όχι ο καθηγητής σας του μαθήματος Πληροφορικής ;
5. Πόσα παιδιά έχει ο Πρωθυπουργός της χώρας ;

Για κάθε μία από τις παραπάνω πέντε ερωτήσεις καταγράφονται δύο χαρακτηριστικά και καταγράφονται οι τιμές για τις δύο τυχαίες μεταβλητές T and S :

T = ο χρόνος που χρειάστηκε για να βρεθεί η πληροφορία. Τον χρόνο τον μετράμε σε δευτερόλεπτα (seconds). Το όριο των 1.800 δευτερόλεπτων (=30 λεπτών) έχει μπει όταν ο χρήστης δεν μπορέσει να βρει την πληροφορία που αναζητά.

S = τα βήματα (κλικ) που γίνονται / πραγματοποιούνται μέχρι ο χρήστης βρει την πληροφορία που θέλει. Δεν υπάρχει όριο στα χτυπήματα που κάνει ο χρήστης.

Σύμφωνα με τα παραπάνω παίρνουμε 10 τυχαίες μεταβλητές οι οποίες συμβολίζονται με $T_1, T_2, T_3, T_4, T_5, S_1, S_2, S_3, S_4, S_5$ με 50 τιμές για την κάθε μία. Οι πρώτες 5 τ.μ. $T_i, i=1,2,3,4,5$ αναπαριστούν τα δευτερόλεπτα. Οι τελευταίες 5 τ.μ. $S_i, i=1,2,3,4,5$ αναπαριστούν τα βήματα.

Πίνακας 1 : Οι τιμές των μεταβλητών που καταγράφηκαν

	T_1	T_2	T_3	T_4	T_5	S_1	S_2	S_3	S_4	S_5
1	19	1800	30	78	317	1	25	3	5	25
2	87	1800	28	80	460	4	30	3	9	30
3	27	1800	25	87	502	2	27	2	8	27
4	19	1800	32	294	164	1	22	3	30	15

5	15	1800	24	164	337	1	35	2	12	27
6	105	1800	22	85	323	5	38	3	7	24
7	108	1800	45	71	297	6	25	5	5	24
8	78	1800	41	123	267	4	40	3	11	30
9	77	1800	21	145	313	5	18	2	10	15
10	68	1800	28	512	284	5	20	2	45	21
11	112	1800	35	71	272	9	27	3	5	18
12	91	1800	58	85	80	6	50	6	7	5
13	95	1800	37	95	385	5	21	4	10	24
14	47	1800	36	87	161	5	24	3	9	12
15	59	1800	25	84	135	4	25	2	8	11
16	98	1800	27	89	168	7	22	5	8	11
17	93	1800	28	250	176	9	21	2	24	11
18	34	1800	41	245	361	2	50	5	29	24
19	106	1800	35	120	222	8	27	3	15	12
20	69	1800	32	142	58	9	31	3	12	5
21	22	1800	41	76	190	2	33	4	8	12
22	45	1800	42	57	175	7	21	4	6	14
23	20	1800	32	123	415	1	22	4	8	34
24	30	1800	40	96	481	3	42	4	5	38
25	71	1800	25	145	291	9	41	2	12	31
26	102	1800	28	101	521	12	40	2	8	42
27	40	1800	31	83	175	5	41	2	7	12
28	96	1800	37	124	232	10	27	2	11	19
29	20	1800	70	142	342	2	28	7	12	25
30	71	1800	32	100	141	5	25	3	8	9
31	60	1800	41	87	70	5	21	4	7	5
32	49	1800	42	88	271	6	22	3	7	21
33	62	1800	40	123	160	7	40	5	11	18
34	17	1800	31	409	415	1	41	2	35	45
35	60	1800	38	312	218	5	42	7	25	18

36	41	1800	31	100	162	3	35	3	8	12
37	30	1800	22	98	549	2	38	2	7	27
38	79	1800	25	72	522	8	31	2	6	25
39	107	1800	23	86	94	9	27	2	8	18
40	80	1800	21	146	276	7	23	1	12	18
41	33	1800	38	148	276	4	24	3	13	24
42	86	1800	47	107	164	9	47	5	9	22
43	21	1800	30	91	296	2	41	2	8	18
44	32	1800	22	62	265	3	25	1	5	19
45	70	1800	24	68	367	7	24	2	4	21
46	53	1800	36	103	500	7	29	3	12	42
47	34	1800	35	85	161	6	21	3	9	15
48	107	1800	32	82	382	8	22	3	9	29
49	95	1800	33	123	331	8	25	4	12	21
50	36	1800	38	79	237	5	48	4	7	21

Κανείς έφηβος δεν μπόρεσε να απαντήσει την δεύτερη ερώτηση ($T_2=1800$).

Εξετάστηκε ο τύπος των παραπάνω τυχαίων μεταβλητών. Για κάθε τυχαία μεταβλητή $T_1, T_2, T_3, T_4, T_5, S_1, S_2, S_3, S_4, S_5$ ελέγχονται οι τιμές κάποιων παραμέτρων όπως: Μέση τιμή, Διακύμανση, Τυπική απόκλιση, Συντελεστής Μεταβλητότητας. Τέλος θα προσπαθήσουμε να εκτιμήσουμε τον τύπο της συνάρτησης πυκνότητας πιθανότητας (σ.π.π., pdf) των ανωτέρω τ.μ. Για την τ.μ. T_2 το όριο των 1800 δευτερολέπτων έχει καταγραφεί οπότε δεν υπάρχει λόγος να γίνουν οποιοδήποτε υπολογισμοί.

2. ΕΞΕΤΑΖΟΝΤΑΣ ΤΗΝ ΠΕΡΙΠΤΩΣΗ T_1

Συχνά από τη φύση της ερώτησης προκύπτει ότι η συνάρτηση πυκνότητας πιθανότητας (σ.π.π) είναι συμμετρική (A-1 & A-2 στο Παράρτημα).

Ας ξεκινήσουμε με την τυχαία μεταβλητή T_1 .

Βλέπουμε ότι η μικρότερη και η μεγαλύτερη τιμή της T_1 είναι $T_{min}=15$ και $T_{max}=112$. Έτσι αντιμετωπίζουμε ένα εύρος τιμών $R=112-15=98$. Αυτό μας επιτρέπει να διαμερίσουμε το διάστημα των τιμών T_1 σε 7 υποδιαστήματα πλάτους $W=14$. Υποθέτουμε ότι τα διαστήματα αυτά είναι κλειστά από τα αριστερά και ανοικτά από τα δεξιά. Συμβολίζουμε με n_i τον αριθμό των στοιχείων που αντιστοιχούν σε οποιαδήποτε από τις επτά κλάσεις, βλ. Εικόνα 1. Έχουμε, επίσης, σε αυτή την εικόνα, το πολύγωνο συχνότητας και ψάχνουμε για την συνάρτηση σ.π.π.

Οι υπολογισμοί των παραμέτρων της T_1 έδωσαν τις τιμές του παρακάτω Πίνακα 2:

Πίνακας 2: Οι τιμές των παραμέτρων της T_1

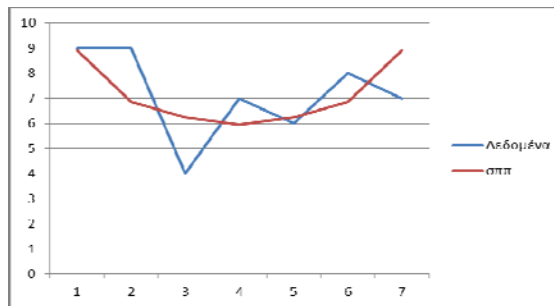
<i>Mean X</i>	62.32	<i>k</i>	-0.1775
<i>Var X</i>	885.1200	<i>h</i>	0.01674633
<i>Std X</i>	29.795	$X^2_{(6;0.05)}=$	12.5916
<i>CoV X</i>	0.4774	Sample-$X^2=$	2.2773 < $X^2_{(6;0.05)}$
λ	2,5720		

Εφαρμόζοντας τη μέθοδο που περιγράφεται στην εργασία Farmakis (2003) η αντίστοιχη συνάρτηση πυκνότητας πιθανότητας είναι η:

$$f(x) = \begin{cases} 0.01674633 \cdot (x-a)^{-0.1775}, & x \in [15, 64] \\ 0.01674633 \cdot (b-x)^{-0.1775}, & x \in [64, 113] \\ 0, & x \notin [15, 113] \end{cases} \quad (2.1)$$

Από την (2.1) υπολογίσαμε θεωρητικές τιμές για την T_1 και τις συγκρίναμε με τις παρατηρούμενες $n_i, i=1,2,\dots,7$, δοκιμασία X^2 , για τον έλεγχο «καλής προσαρμογής». Προέκυψε τιμή $X^2=2.2773 < X^2_{(6;0.05)}$ οπότε δεν μπορούμε να απορρίψουμε την H_0 , ότι η μέθοδος δίνει σωστή μορφή για την σ.π.π. Σημειώνεται ότι η μέθοδος δίνει πολυωνυμική μορφή (εύκολα προσβάσιμη και εύκολη στη διαχείριση) για τις σ.π.π.

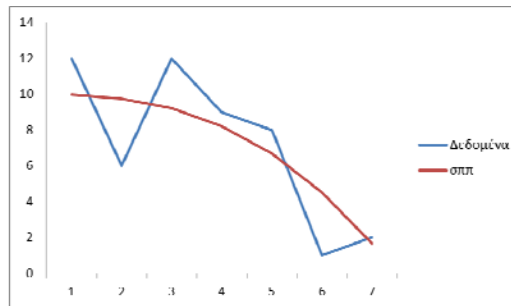
Εικόνα 1 : Το πολύγωνο συχνοτήτων του πλήθους των μαθητών υποδεικνύει ότι έχουμε μία συμμετρική σ.π.π. (η λεία καμπύλη)



3. ΕΞΕΤΑΖΟΝΤΑΣ ΤΗΝ ΠΕΡΙΠΤΩΣΗ T_6 - ΜΙΑ ΦΘΙΝΟΥΣΑ σ.π.π.

Θα παρουσιαστεί μία σ.π.π. με φθίνουσα συμπεριφορά. Πρόκειται για την T_6 που απεικονίζει τις μέσες επιδόσεις των 50 φοιτητών ως προς όλες τις παραμέτρους $T_m, m=1,3,4,5$ που αντιμετωπίζουν επιτυχώς. Έχουμε $T_{min}=63$ και $T_{max}=224$, δηλ εύρος τιμών $R=161$ και τις ταξινομούμε σε 7 ομάδες με εύρος $W=23$ sec η κάθε μία τους. Στην Εικόνα 2 δίνονται πολύγωνο πληθών και η τελική μορφή της σ.π.π. με την λειαινούσα μορφή, που είναι φθίνουσα.

Εικόνα 2 : Το πολύγωνο συχνοτήτων δείχνει ότι πρόκειται για μία φθίνουσα σ.π.π.



Τα παραπάνω δεδομένα δίνουν τιμές των παραμέτρων στον Πίνακα 3:

Πίνακας 3: Οι τιμές των παραμέτρων της T_6

<i>Mean X</i>	60,26	<i>k</i>	2,51253
<i>Var X</i>	1460,47	<i>h</i>	0,00868326
<i>Std X</i>	38,22	$X^2_{(6;0.05)}=$	12,5916
<i>CoV X</i>	0,6342	Sample-$X^2=$	5,846793⁽¹⁾ < $X^2_{(6;0.05)}$
<i>λ</i>	2,48637		

(1) Σημείωση: Η τιμή της X^2 -score είναι 5,847 δηλ. η υπόθεση H_0 δεν απορρίπτεται. Στο ίδιο συμπέρασμα φτάνουμε και αν συμπτύξουμε τις κλάσεις έτσι ώστε οι αναμενόμενες συχνότητες να είναι μεγαλύτερες ή ίσες του 5.

Η σ.π.π. δίνεται από τον τύπο (3.1)

$$f(x) = \begin{cases} 0,00868326 \left\{ 1 - \left(\frac{x}{161} \right)^{2,51253} \right\}, & x \in [0,161] \\ 0, & x \notin [0,161] \end{cases} \quad (3.1)$$

Ολοκληρώνοντας αυτή τη συνάρτηση πήραμε θεωρητικές τιμές συχνοτήτων και τις συγκρίναμε με τις παρατηρηθείσες (12, 6, 12, 9, 8, 1,2). Η δοκιμασία X^2 απέδωσε την H_0 ισχύουσα, όπου H_0 : «Η (3.1) περιγράφει θεωρητικά την κατανομή της T_6 με βάση τα δεδομένα του δείγματος σε σ.σ.α=0.05».

4. ΕΞΕΤΑΣΗ ΤΩΝ ΑΛΛΩΝ ΠΕΡΙΠΤΩΣΕΩΝ ΤΩΝ T_m και S_m $m=1, \dots, 6$

Αρχικά πρέπει να δηλώσουμε ότι δεν έχουμε ασχοληθεί με τις τυχαίες μεταβλητές T_2 και S_2 . Οι τυχαίες μεταβλητές αυτές έχουν να κάνουν σχέση με την δεύτερη ερώτηση (η πρώτη αφορά τον χρόνο και η δεύτερη τα κλικ) την οποία δεν απάντησαν τελικά οι έφηβοι.

Για τις υπόλοιπες τυχαίες μεταβλητές είμαστε σε θέση να δώσουμε μία εκτίμηση για το είδος της σ.π.π. προσδιορίσαμε και συμβολίζουμε στον παρακάτω πίνακα 4. Συμβολίζεται με: Σ= συμμετρική, Φ= φθίνουσα, Χ= δεν έγινε έλεγχος.

ΠΙΝΑΚΑΣ 4 : Είδος της σ.π.π. για την κάθε μεταβλητή

τμ	T_1	T_2	T_3	T_4	T_5	S_1	S_2	S_3	S_4	S_5	T_6	S_6
σππ	Σ	Χ	Φ	Φ	Σ	Σ	Χ	Σ	Φ	Σ	Φ	Φ

Η S_6 απεικονίζει τις μέσες επιδόσεις των 50 φοιτητών ως προς όλες τις παραμέτρους $S_m, m=1, \dots, 5$.

ΣΥΜΠΕΡΑΣΜΑΤΑ

Είναι ξεκάθαρο ότι οι T_1, T_4, T_5, S_4, T_6 περνούν το κριτήριο του X^2 -τεστ και ο τύπος της σ.π.π. τους αναφέρεται στον πίνακα 4. Πολλές φορές χρησιμοποιείται ο ένας ή ο άλλος τύπος φθίνουσας σ.π.π. από τις δύο μορφές που αναφέρονται στο Παράρτημα παρακάτω. Ακολουθούμε τους περιορισμούς που αναφέρονται στην μορφή απόδοσης του εκθέτη $k(\lambda)$, βλ. (Α-5), (Α-6), (Α-7), (Α-8).

Στα αρχικά ερωτήματα που τέθηκαν στην εισαγωγή απαντάμε μετά την έρευνα αυτή:

1^ο) Ο χρόνος κρίνεται αρκετά μακρύς ώσπου να έχουμε την απάντηση, γιατί χρειάζεται περιπλάνηση στο διαδίκτυο και πέρασμα από αρκετές σελίδες.

2^ο) Υπάρχουν πολλές (και ίσως ανυπέβλητες) δυσκολίες για να βρει κανένας απάντηση σχετικά με την διεύθυνση ηλεκτρονικού ταχυδρομείου κάποιου προσώπου.

3^ο) Η ανάπτυξη της «*ΣΗΜΑΝΤΙΚΗΣ στο ΔΙΑΔΙΚΤΥΟ*» (Semantic Web) μάλλον θα δώσει απάντηση για το άμεσο ή μη της απάντησης σε κλασικά ερωτήματα του τύπου που αναλύσαμε στην παρούσα εργασία.

ΠΑΡΑΡΤΗΜΑ

Η ιδέα του συντελεστή μεταβλητότητας (C_v) είναι πολύ γνωστή και χρησιμοποιείται ευρέως από πολλούς επιστήμονες (Βιολόγους, Κοινωνιολόγους, Οικονομολόγους, Μαθηματικούς, κ.α..) για πολλά χρόνια. Πρόσφατα μία νέα χρήση προτείνεται από τον Φαρμάκη (2003). Προτάθηκε η χρήση του συντελεστή μεταβλητότητας (C_v) ως ένα εργαλείο για να προσδιοριστεί η συνάρτηση πυκνότητας πιθανότητας για μία τ. μ. X , η οποία βασίζεται σε ένα δείγμα μεγέθους n . Αυτή η συνάρτηση είναι πολυωνυμικού τύπου και προϋποθέτει ότι η τ.μ. είναι συνεχής και συμμετρική. Η σ.π.π. μπορεί να είναι επίσης αύξουσα ή φθίνουσα συνάρτηση, Φαρμάκης (2010). Πρόκειται να δώσουμε κάποιους ορισμούς και περιγραφές για τον ρόλο του συντελεστή μεταβλητότητας (C_v) για τον προσδιορισμό της σ.π.π. μιας τυχαίας μεταβλητής X από ένα δείγμα.

Ορισμός A1: Για μία τ.μ. X η οποία παίρνει θετικές τιμές ο συντελεστής μεταβλητότητας δίνεται από τον τύπο $Cv = \frac{\sigma}{EX}$ και μπορεί να εκτιμηθεί από ένα

δείγμα από τον τύπο $\hat{Cv} = \frac{s}{\bar{x}}$.

Ορισμός A2: Μία συνάρτηση f η οποία ορίζεται στο διάστημα $[a, b]$ του συνόλου των πραγματικών αριθμών \mathbf{R} είναι συμμετρική με άξονα συμμετρίας την κατακόρυφη γραμμή (ε) $x = \frac{a+b}{2}$, εάν ισχύει: $f\left(\frac{a+b}{2} + x\right) = f\left(\frac{a+b}{2} - x\right)$, $x \in [0, \frac{b-a}{2}]$.

Η πολυωνυμικού τύπου σ.π.π. της τυχαίας μεταβλητής X που εκτιμάται, δίνεται από τον ακόλουθο τύπο:

$$f(x) = \begin{cases} h \cdot (x-a)^k, & x \in [a, \frac{a+b}{2}] \\ h \cdot (b-x)^k, & x \in [\frac{a+b}{2}, b], \quad k \neq -1, \quad h = \frac{2^k(k+1)}{(b-a)^{k+1}} \\ 0, & x \notin [a, b] \end{cases} \quad (\text{A-1})$$

Στην εργασία Φαρμάκης (2003) αποδεικνύεται ότι η παράμετρος k δίνεται από τον τύπο:

$$k = \frac{-5 + \sqrt{1+8 \cdot \lambda}}{2} > -1, \quad \lambda = \left(Cv \cdot \frac{b+a}{b-a}\right)^{-2}, \quad \lambda > 0. \quad (\text{A-2})$$

Στην εργασία Φαρμάκης (2010) για την περίπτωση της αύξουσας σ.π.π. τυχαίας μεταβλητής X δίνεται από τον παρακάτω τύπο:

$$f(x) = \begin{cases} h \cdot \left(\frac{x}{b}\right)^k, & x \in [0, b], \quad k > -1, \quad h = \frac{k+1}{b} \\ 0, & x \notin [0, b] \end{cases} \quad (\text{A-3})$$

και υπολογίζοντας την τιμή της παραμέτρου k από τον τύπο :

$$k = -2 + \sqrt{1+\lambda}, \quad \lambda = Cv^{-2}, \quad \lambda > 0. \quad (\text{A-4})$$

Αν η σ.π.π. είναι φθίνουσα τότε μπορεί να οριστεί από τους τύπους (A-5) and (A-7):

$$f(x) = \begin{cases} h \cdot \left(1 - \frac{x}{b}\right)^k, & x \in [0, b], \quad k > -1, \quad h = \frac{k+1}{b} \\ 0, & x \notin [0, b] \end{cases} \quad (\text{A-5})$$

$$\text{με} \quad k = \frac{3-\lambda}{\lambda-1} = -1 + \frac{2}{\lambda-1}, \quad \lambda = Cv^{-2}, \quad (\text{A-6})$$

$$\text{ή} \quad f(x) = \begin{cases} h \cdot \left(1 - \left(\frac{x}{b}\right)^k\right), & x \in [0, b], \\ 0, & x \notin [0, b] \end{cases} \quad (\text{A-7})$$

$$\text{με} \quad k = -2 + \sqrt{3 \cdot \frac{\lambda+1}{3-\lambda}}, \quad \lambda = Cv^{-2}, \quad h = \frac{k+1}{k \cdot b}. \quad (\text{A-8})$$

Η συνάρτηση (A-5) ονομάζεται φθίνουσα τύπου I και η συνάρτηση (A-7) είναι η φθίνουσα τύπου II τουλάχιστον στην παρούσα δημοσίευση. Το χ^2 -τεστ χρησιμοποιείται για να ελεγχτεί εάν η τιμή της παραμέτρου k που προέκυψε από το δείγμα είναι ικανή να δώσει μία συνάρτηση πυκνότητας πιθανότητας με καλύτερη προσαρμογή στα δεδομένα, Φαρμάκης (2003), (2010). Επίσης η δημιουργία κλάσεων είναι απαραίτητη για την παρουσίαση των δεδομένων και την εκτίμηση των παραμέτρων της τ.μ. X όπως η μέση τιμή, διακύμανση, τυπική απόκλιση, συντελεστής μεταβλητότητας.

ABSTRACT

A sample of High School students is searching on the Web trying to get information and answers to some simple questions. The question will be of the type:

“When PARTHENON was built?”,

“In which High School is Mr. G. Makris teaching?”, etc., etc.

Two random variables are studied:

(a) The time passed until an answer was reached and

(b) The hits (clicks) used for reaching the answer.

Basically the next questions are faced and answered:

1st) How much time we need to find out answers to question like the above one.

2nd) How many movements (clicks) we need to get the answers we are looking for?

3rd) Can we get answers directly?

Ευχαριστίες: Ευχαριστούμε τους κριτές της εργασίας για τις εύστοχες παρατηρήσεις και τις διορθώσεις τους.

ΑΝΑΦΟΡΕΣ

N. Farmakis, “*Introduction to Sampling*“, Edit. A. & P. Christodoulidi Publications, Thessaloniki, 2009 (in Greek).

N. Farmakis, “Estimation of Coefficient of Variation: Scaling of Symmetric Continuous Distributions”, *Statistics in Transition*, Vol. 6, No 1, pp 83-96, 2003.

N. Farmakis, “Coefficient of Variation: Connecting Sampling with some Increasing Distribution Models”, *Proceedings of Stochastic Modeling Techniques & Data Analysis (SMTDA)*, Chania, Crete, June 2010.

E. Kolaczyk, “*Statistical Analysis of Network Data*” Springer Science +Business Media, LLC 2009.



ΣΧΕΔΙΑΣΜΟΙ ΕΠΑΝΑΛΑΜΒΑΝΟΜΕΝΩΝ ΜΕΤΡΗΣΕΩΝ, ΙΚΑΝΕΣ ΚΑΙ ΑΝΑΓΚΑΙΕΣ ΣΥΝΘΗΚΕΣ ΒΕΛΤΙΣΤΟΠΟΙΗΣΗΣ

Μιλτιάδης Χαλκιάς¹, Στρατής Κουνιάς²

¹ Τμήμα Διοίκησης Επιχειρήσεων, ΤΕΙ Πειραιά, mchalik@teipir.gr

² Τμήμα Μαθηματικών, Πανεπιστήμιο Αθηνών, skounias@math.uoa.gr

ΠΕΡΙΛΗΨΗ

Εξετάζονται σχεδιασμοί επαναλαμβανόμενων μετρήσεων με t αγωγές, n πειραματικές μονάδες και p περιόδους. Το μοντέλο είναι με ασυσχέτιστες παρατηρήσεις, τα σφάλματα έχουν σταθερή διασπορά ενώ οι προς εκτίμηση παράμετροι είναι : (i) Η άμεση επίδραση των αγωγών (direct effects) (ii) Η μεταφερόμενη επίδραση των αγωγών (residual effects). Στην συγκεκριμένη εργασία, (a) διευκρινίζεται η διαφορά των καθολικά βέλτιστων και Φ-βέλτιστων σχεδιασμών, (b) επεκτείνονται οι αναγκαίες συνθήκες των Cheng and Wu (1980), για να περιλάβουν και την περίπτωση $n=2 \bmod 4$, p άρτιος, (στην περίπτωση αυτή οι εκτιμήτριες των παραμέτρων στον βέλτιστο σχεδιασμό είναι συσχετισμένες), (c) δείχνεται ότι οι συγκεκριμένες συνθήκες είναι επίσης αναγκαίες για τη Φ-βελτιστοποίηση της εκτίμησης άμεσων και μεταφερόμενων επιδράσεων.

Λέξεις κλειδιά: Καθολικά βέλτιστος, Φ-βέλτιστος, κυριαρχία, συνάρτηση πληροφορίας.

1. ΕΙΣΑΓΩΓΗ

Στους σχεδιασμούς διασταύρωσης ή σχεδιασμούς επαναλαμβανόμενων μετρήσεων σε κάθε μία από τις n πειραματικές μονάδες, οι οποίες στο εξής θα ονομάζονται μονάδες, εφαρμόζεται μία από τις t αγωγές σε κάθε μια από τις p διαδοχικές περιόδους.

Στους σχεδιασμούς αυτούς μας ενδιαφέρει η εκτίμηση της άμεσης επίδρασης κάθε αγωγής (direct effect) και της μεταφερόμενης επίδρασης κάθε αγωγής (residual effect). Μεταφερόμενη επίδραση είναι η απομένουσα επίδραση της αγωγής που ασκήθηκε την προηγούμενη περίοδο από αυτή που εξετάζουμε.

Χρησιμοποιούμε το μοντέλο των Hedayat and Afsarinejad (1975, 1978). Οι ορισμοί ομοιομορφίας ως προς τις περιόδους και τις πειραματικές μονάδες καθώς και

του πλήρους ισορροπημένου σχεδιασμού, δόθηκαν στις εργασίες των Hedayat και Afsarinejad (1978) και στην εργασία των Cheng and Wu (1980). Ακόμα η έννοια του καθολικά βέλτιστου σχεδιασμού δόθηκε από τον Kiefer (1975). Με το θέμα αυτό έχουν ασχοληθεί πολλοί ερευνητές, Kunert (1984), Kushner (1997a, 1997b) κ.ά.

Συμβολίζουμε με $RMD(t, n, p)$ τους σχεδιασμούς επαναλαμβανόμενων μετρήσεων με t αγωγές, n μονάδες και p περιόδους.

Ορισμός 1.1 Ένας σχεδιασμός είναι ομοιόμορφος ως προς τις μονάδες αν κάθε μονάδα έχει το ίδιο πλήθος (i) αγωγών A και B , όταν το πλήθος p των περιόδων είναι άρτιος (ii) αγωγών A και B στις πρώτες $p-1$ περιόδους, όταν ο p είναι περιττός. •

Ορισμός 1.2 Ένας σχεδιασμός είναι ομοιόμορφος ως προς τις περιόδους αν σε κάθε περίοδο υπάρχει ίσο πλήθος των A και B . •

Ορισμός 1.3 Ένας σχεδιασμός είναι ισχυρά ισορροπημένος αν έχει συνολικά το ίδιο πλήθος των ζευγών AA, AB, BA, BB . •

Ένας σχεδιασμός προσδιορίζεται αν δίνεται ποια αγωγή εφαρμόζεται σε κάθε μονάδα και σε κάθε περίοδο.

Οι Cheng και Wu (1980) έδωσαν τις επόμενες αναγκαίες συνθήκες για να είναι ένας σχεδιασμός καθολικά βέλτιστος για την περίπτωση που εξετάζουμε.

i) Έστω d^* ένας ομοιόμορφος και πλήρως ισορροπημένος σχεδιασμός. Η ικανή συνθήκη ώστε ο d^* να είναι καθολικά βέλτιστος για την εκτίμηση των άμεσων και μεταφερόμενων επιδράσεων είναι, $d^* \in RMD(t, n = \lambda_1 t, p = \lambda_2 t)$.

ii) Έστω d^* πλήρως ισορροπημένος, ο οποίος είναι ομοιόμορφος ως προς τις περιόδους και ομοιόμορφος ως προς τις μονάδες στις πρώτες $p-1$ περιόδους. Αν $d^* \in RMD(t, n = \lambda_1 t, p = \lambda_2 t + 1)$, $\lambda_1, \lambda_2 \geq 1$, τότε είναι καθολικά βέλτιστος για τη διαφορά των άμεσων και επαναλαμβανόμενων επιδράσεων διότι οι επιδράσεις αυτές δεν είναι εκτιμήσιμες.

Αν d είναι ομοιόμορφος ως προς τις μονάδες, τότε σε κάθε μονάδα έχουμε το ίδιο πλήθος αγωγών A και B , έτσι όταν p είναι άρτιος το αποτέλεσμα (i) των Cheng και Wu εφαρμόζεται. Με το ίδιο σκεπτικό το αποτέλεσμα (ii) των Cheng και Wu εφαρμόζεται όταν p είναι περιττός.

2. ΚΥΡΙΑΡΧΙΑ ΚΑΙ ΒΕΛΤΙΣΤΟΠΟΙΗΣΗ

Ορισμός 2.1 Το διάνυσμα $\mathbf{x} = (x_1, \dots, x_n)'$ κυριαρχείται (majorized) από το $\mathbf{y} = (y_1, \dots, y_n)'$, (συμβολισμός $\mathbf{x} \prec \mathbf{y}$), αν $\mathbf{x} = \mathbf{S}\mathbf{y}$ όπου \mathbf{S} είναι ένας διπλά στοχαστικός πίνακας, που σημαίνει ότι έχει μόνο μη αρνητικά στοιχεία και το άθροισμα των στοιχείων των γραμμών και των στηλών του είναι ίσο με 1. •

Αν $\mathbf{x} \prec \mathbf{y} \Leftrightarrow \sum_{i=1}^n g(x_i) \geq \sum_{i=1}^n g(y_i)$ για κάθε κοίλη συνάρτηση, ενώ για κυρτές συναρτήσεις η ανισότητα αντιστρέφεται (Marshall and Olkin (1979), p.11).

Ορισμός 2.2 Αν \mathbf{C} και \mathbf{D} είναι συμμετρικοί $k \times k$ πίνακες και ο \mathbf{C} ανήκει στο διανυσματικό χώρο των πινάκων της μορφής \mathbf{PDP}' όπου \mathbf{P} είναι $k \times k$ μεταθετικός πίνακας τότε ο πίνακας \mathbf{C} κυριαρχείται από τον \mathbf{D} και συμβολίζεται με $\mathbf{C} \prec \mathbf{D}$, αυτό σημαίνει,

$$\mathbf{C} = \sum_i a_i \mathbf{P}_i \mathbf{D} \mathbf{P}_i', \sum_i a_i = 1, a_i > 0.$$

Ορισμός 2.3 Η συνάρτηση $\varphi(\mathbf{C})$ ονομάζεται συνάρτηση πληροφορίας αν $\varphi(\mathbf{C}) \rightarrow R$ για κάθε $\mathbf{C} \in nnd(k)$ όταν ικανοποιούνται οι ακόλουθες συνθήκες:

$$(i) \mathbf{C} \geq \mathbf{D} \Rightarrow \varphi(\mathbf{C}) \geq \varphi(\mathbf{D})$$

$$(ii) \varphi(p\mathbf{C} + (1-p)\mathbf{D}) \geq p\varphi(\mathbf{C}) + (1-p)\varphi(\mathbf{D}) \quad \forall 0 < p < 1, \mathbf{C}, \mathbf{D} \geq \mathbf{0}$$

$$(iii) \varphi(d\mathbf{C}) = d\varphi(\mathbf{C}), \quad \forall d > 0, \mathbf{C} \geq \mathbf{0} \Rightarrow \varphi(\mathbf{0}) = 0 \Rightarrow \varphi(\mathbf{C}) \geq 0, \quad \forall \mathbf{C} \geq \mathbf{0}$$

$$(iv) \varphi(\mathbf{C}) = \varphi(\mathbf{PCP}') \quad \forall \mathbf{C} \in nnd(k), \mathbf{P} \in perm(k)$$

(2.1)

Έτσι οι συναρτήσεις πληροφορίας είναι κοίλες και αύξουσες. •

Ο συμβολισμός $nnd(k)$ χρησιμοποιείται για $k \times k$ μη αρνητικά ορισμένους πίνακες, με $perm(k)$ συμβολίζεται ο $k \times k$ μεταθετικός πίνακας.

Ορισμός 2.4 Αν $\sigma^2 \mathbf{V}_d$ είναι ο πίνακας διασποράς των εκτιμητών των παραμέτρων, τότε ο $\mathbf{Q}_d = \mathbf{V}_d^{-1}$ ονομάζεται πίνακας πληροφορίας. •

Ορισμός 2.5 Ένας σχεδιασμός d^* με πίνακα πληροφορίας \mathbf{C}_d^* είναι καθολικά βέλτιστος στην κλάση των σχεδιασμών F αν μεγιστοποιεί την ποσότητα $\varphi(\mathbf{C}_d)$, $d \in F$, για κάθε συνάρτηση πληροφορίας φ . •

Για καθολικά βέλτιστους σχεδιασμούς προτιμάμε να δουλεύουμε χρησιμοποιώντας πίνακες πληροφορίας και αύξουσες κοίλες συναρτήσεις $\varphi(\mathbf{C})$ αντί με πίνακες διασποράς και φθίνουσες κυρτές $g(\mathbf{V})$. Μία κλάση συναρτήσεων πληροφορίας για θετικά ορισμένους $k \times k$ πίνακες \mathbf{Q} είναι (Pukelsheim (1993)),

$$\varphi_s(\mathbf{Q}) = \begin{cases} \lambda_{\max}(\mathbf{Q}) & \text{αν } s = 1 \\ \left(\frac{1}{k} \text{tr}(\mathbf{Q}^s) \right)^{1/s} & \text{αν } s \neq 0, \infty \\ (\det \mathbf{Q})^{1/k} & \text{αν } s = 0 \\ \lambda_{\min}(\mathbf{Q}) & \text{αν } s = -\infty \end{cases}$$

Για $s \leq 1$, $\varphi_s(\mathbf{Q})$ είναι συνάρτηση πληροφορίας, για $s = -1$ έχουμε A-βελτιστοποίηση, για $s = 0$ D-βελτιστοποίηση και για $s = -\infty$ E-βελτιστοποίηση.

Η MV βελτιστοποίηση, που σημαίνει την ελαχιστοποίηση της μέγιστης διασποράς των υπό εκτίμηση παραμέτρων, δεν περιλαμβάνεται στην παραπάνω κλάση.

Ορισμός 2.6. Ένας σχεδιασμός d^* με πίνακα πληροφορίας $C_d^*(k)$, με ιδιοτιμές $\lambda_1, \lambda_2, \dots, \lambda_k$, είναι Φ-βέλτιστος, στην κλάση των σχεδιασμών F αν ελαχιστοποιεί την ποσότητα $\sum_{i=1}^k \varphi(\lambda_i)$ για όλες τις συνεχείς, φθίνουσες και κυρτές συναρτήσεις $\varphi(\lambda)$.

Σημειώνουμε ότι MV- βελτιστοποίηση περιλαμβάνεται στους καθολικά βέλτιστους (universally optimal) σχεδιασμούς αλλά όχι στους Φ-βέλτιστους.

Σε ένα πρόβλημα σχεδιασμών αναζητούμε καθολικά βέλτιστους σχεδιασμούς, αν δεν υπάρχουν, τότε αναζητούμε Φ- βέλτιστους σχεδιασμούς και τέλος αν δεν υπάρχουν τότε αναζητούμε A-,D-,E- ή MV- βέλτιστους σχεδιασμούς.

3. ΤΟ ΜΟΝΤΕΛΟ

Το μοντέλο είναι:

$$Y_{ij} = \tau_{d(i,j)} + \pi_j + \delta_{d(i,j-1)} + \gamma_i + e_{ij} \quad (3.1)$$

Το $i = 1, \dots, n$, αναφέρεται στη μονάδα το $j = 1, \dots, p$ στην περίοδο, $\tau_{d(ij)} \in \{\tau_A, \tau_B\}$ είναι η άμεση επίδραση για την εφαρμοζόμενη αγωγή, την j περίοδο στην i μονάδα, $\delta_{d(i,j-1)} \in \{\delta_A, \delta_B\}$ είναι η μεταφερόμενη επίδραση της αγωγής που ασκήθηκε την $(j-1)$ περίοδο, π_j η επίδραση της περιόδου j και γ_i η επίδραση της μονάδας i .

Τα σφάλματα e_{ij} είναι ανεξάρτητα μεταξύ των περιόδων κάθε μονάδας και ανάμεσα στις μονάδες, έχουν μέση τιμή 0 και σταθερή διασπορά.

Το μοντέλο (3.1) σε διανυσματική μορφή γράφεται,

$$\mathbf{Y} = \tau_A \boldsymbol{\tau}_A + \tau_B \boldsymbol{\tau}_B + \delta_A \boldsymbol{\delta}_A + \delta_B \boldsymbol{\delta}_B + \pi_1 \boldsymbol{\pi}_1 + \dots + \pi_p \boldsymbol{\pi}_p + \gamma_1 \boldsymbol{\gamma}_1 + \dots + \gamma_n \boldsymbol{\gamma}_n + \mathbf{e} \quad (3.2)$$

όπου $\mathbf{Y}, \boldsymbol{\tau}_A, \boldsymbol{\tau}_B, \boldsymbol{\delta}_A, \boldsymbol{\delta}_B, \boldsymbol{\pi}_1, \dots, \boldsymbol{\pi}_p, \boldsymbol{\gamma}_1, \dots, \boldsymbol{\gamma}_n, \mathbf{e}$ είναι $pn \times 1$ διανύσματα, το $\boldsymbol{\tau}_A$ ($\boldsymbol{\tau}_B$) έχει 1 όταν εφαρμόζεται η αγωγή A(B) και 0 αλλού, ενώ $\boldsymbol{\tau}_A + \boldsymbol{\tau}_B = \mathbf{1}_{pn}$, $\boldsymbol{\delta}_A$ ($\boldsymbol{\delta}_B$) έχει 0 στην πρώτη περίοδο, 1 όταν η αγωγή A(B) εφαρμόστηκε στην προηγούμενη περίοδο και 0 αλλού, $\boldsymbol{\pi}_i$ έχει 1 στην i περίοδο και 0 αλλού, έτσι $\boldsymbol{\delta}_A + \boldsymbol{\delta}_B + \boldsymbol{\pi}_1 = \mathbf{1}_{pn}$, επίσης $\boldsymbol{\pi}_1 + \dots + \boldsymbol{\pi}_p = \mathbf{1}_{pn}$. Ακόμα $\boldsymbol{\gamma}_i, i = 1, \dots, n$ έχει 1 όταν εμφανίζεται η i πειραματική μονάδα και 0 αλλού, οπότε $\boldsymbol{\gamma}_1 + \dots + \boldsymbol{\gamma}_n = \mathbf{1}_{pn}$.

Κρατώντας μόνο τα γραμμικά ανεξάρτητα διανύσματα (Kounias and Chalikias (2008)) το μοντέλο (3.2) μετασχηματίζεται στο:

$$\mathbf{E}(\mathbf{Y}) = \tau(\boldsymbol{\tau}_A - \boldsymbol{\tau}_B) + \delta(\boldsymbol{\delta}_A - \boldsymbol{\delta}_B) + \tilde{\pi}_1 \boldsymbol{\pi}_1 + \dots + \tilde{\pi}_p \boldsymbol{\pi}_p + \tilde{\gamma}_1 \boldsymbol{\gamma}_1 + \dots + \tilde{\gamma}_n \boldsymbol{\gamma}_n \quad (3.3)$$

όπου $\tau = (\tau_A - \tau_B) / 2$, $\delta = (\delta_A - \delta_B) / 2$

Το μοντέλο (3.1) γράφεται σε διανυσματική μορφή: $\mathbf{Y} = \mathbf{X}\mathbf{b} + \mathbf{e}$, όπου \mathbf{Y} είναι $pn \times 1$, ο πίνακας σχεδιασμού \mathbf{X} είναι $(pn) \times s$, \mathbf{b} είναι $s \times 1$, \mathbf{e} είναι $(pn) \times 1$ διάνυσμα, s είναι το πλήθος των άγνωστων παραμέτρων. Αν ενδιαφερόμαστε για κάποιες και όχι

για όλες τις παραμέτρους γράφουμε $\mathbf{b}' = (\mathbf{b}'_1, \mathbf{b}'_2)$, όπου \mathbf{b}_1 είναι οι r παραμέτρους που μας ενδιαφέρουν και \mathbf{b}_2 είναι οι $s-r$ παραμέτρους που υπολείπονται. Η μέθοδος ελαχίστων τετραγώνων μας δίνει την εκτίμηση των παραμέτρων που μας ενδιαφέρουν,

$$\mathbf{X}'_1(\mathbf{I}_{pn} - \mathbf{P}(\mathbf{X}_2))\mathbf{X}_1\hat{\mathbf{b}}_1 = \mathbf{X}'_1(\mathbf{I}_{pn} - \mathbf{P}(\mathbf{X}_2))\mathbf{Y}$$

όπου $\mathbf{X} = (\mathbf{X}_1 | \mathbf{X}_2)$ είναι $(pn) \times s$ πίνακας, \mathbf{X}_1 είναι $(pn) \times r$ ενώ $\mathbf{P}(\mathbf{X}_2) = \mathbf{X}_2(\mathbf{X}'_2\mathbf{X}_2)^{-1}\mathbf{X}'_2$ είναι ο $(pn) \times (pn)$ πίνακας ορθής προβολής στον διανυσματικό χώρο των στηλών του \mathbf{X}_2 . Ο πίνακας διασποράς των εκτιμημένων παραμέτρων είναι,

$$\text{var}(\hat{\mathbf{b}}_1) = \sigma^2(\mathbf{X}'_1(\mathbf{I}_{pn} - \mathbf{P}(\mathbf{X}_2))\mathbf{X}_1)^{-1} = \sigma^2\mathbf{Q}^{-1} \quad (3.4)$$

Όπου $\mathbf{Q} = \mathbf{X}'_1(\mathbf{I}_{pn} - \mathbf{P}(\mathbf{X}_2))\mathbf{X}_1$ είναι ο πίνακας πληροφορίας των παραμέτρων που μας ενδιαφέρουν ενώ στην περίπτωση μιας παραμέτρου ο \mathbf{Q} είναι η απόσταση του διανύσματος \mathbf{X}_1 από τον γραμμικό χώρο $L(\mathbf{X}_2)$ των στηλών του \mathbf{X}_2 .

Με $m(A)_j, m(B)_j$ συμβολίζουμε το πλήθος των A, B αντιστοίχως στην j περίοδο σε όλες τις μονάδες. Με $n(AB)$ το πλήθος των εμφανίσεων του ζεύγους AB σε όλες τις μονάδες ενώ με ανάλογο τρόπο ορίζονται $n(AA), n(BA), n(BB)$. Ακόμα με $n(A)_i$ συμβολίζουμε το πλήθος των A στην i μονάδα $i=1, \dots, n$ και με $\tilde{n}(A)_i$ συμβολίζουμε το πλήθος των A στην i μονάδα στις $p-1$ πρώτες περιόδους. Τέλος με $n(A)$ συμβολίζουμε το πλήθος των A σε όλες τις μονάδες ενώ με ανάλογο τρόπο έχουμε $n(B)_i, \tilde{n}(B)_i, n(B)$.

Οι παράμετροι που μας ενδιαφέρουν είναι, $\tau = (\tau_A - \tau_B)/2, \delta = (\delta_A - \delta_B)/2$, τότε από την σχέση (3.3) έχουμε $\mathbf{X}_1 = \{(\boldsymbol{\tau}_A - \boldsymbol{\tau}_B), (\boldsymbol{\delta}_A - \boldsymbol{\delta}_B)\}$, $\mathbf{X}_2 = \{\boldsymbol{\pi}_1, \dots, \boldsymbol{\pi}_p, \boldsymbol{\gamma}_1, \dots, \boldsymbol{\gamma}_n\}$.

Θεώρημα 3.1 Αν το μοντέλο της (3.3) είναι σωστό και $\mathbf{Q} = \mathbf{X}'_1\mathbf{X}_1 - \mathbf{X}'_1\mathbf{X}_2(\mathbf{X}'_2\mathbf{X}_2)^{-1}\mathbf{X}'_2\mathbf{X}_1$, τότε

$$\mathbf{Q} = \mathbf{E} - \frac{1}{n}\mathbf{C} - \frac{1}{p}\mathbf{D} \quad (3.5)$$

Όπου

$$\mathbf{E} = \begin{bmatrix} np & n(AA) + n(AB) + n(BA) + n(BB) \\ n(AA) + n(AB) + n(BA) + n(BB) & n(p-1) \end{bmatrix}$$

$$\mathbf{C} = \begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix}, \mathbf{D} = \begin{bmatrix} d_{11} & d_{12} \\ d_{21} & d_{22} \end{bmatrix}, \quad (3.6)$$

$$c_{11} = \sum_{i=1}^p (m(A)_i - m(B)_i)^2, c_{22} = \sum_{i=1}^{p-1} (m(A)_i - m(B)_i)^2$$

$$c_{12} = c_{21} = \sum_{i=1}^{p-1} (m(A)_i - m(B)_i)(m(A)_{i+1} - m(B)_{i+1}), \quad (3.7)$$

$$d_{11} = \sum_{i=1}^n \left((n(A)_i - n(B)_i) - \frac{1}{n}(n(A) - n(B)) \right)^2, d_{22} = \sum_{i=1}^n \left((\tilde{n}(A)_i - \tilde{n}(B)_i) - \frac{1}{n}(\tilde{n}(A) - \tilde{n}(B)) \right)^2$$

$$d_{12} = d_{21} = \sum_{i=1}^n \left((n(A)_i - n(B)_i) - \frac{1}{n}(n(A) - n(B)) \right) \left((\tilde{n}(A)_i - \tilde{n}(B)_i) - \frac{1}{n}(\tilde{n}(A) - \tilde{n}(B)) \right)$$

Απόδειξη. Γράφουμε $\mathbf{X}_2 = (\mathbf{Z} | \mathbf{W})$, όπου $\mathbf{Z} = (\boldsymbol{\pi}_1, \dots, \boldsymbol{\pi}_p)$ και $\mathbf{W} = (\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \dots, \boldsymbol{\gamma}_n)$, τότε ο γενικευμένος αντίστροφος του block πίνακα δίνεται από την σχέση (Searle 1971, p.27),

$$\begin{aligned} (\mathbf{X}'_2 \mathbf{X}_2)^- &= \begin{bmatrix} \mathbf{Z}'\mathbf{Z} & \mathbf{Z}'\mathbf{W} \\ \mathbf{W}'\mathbf{Z} & \mathbf{W}'\mathbf{W} \end{bmatrix}^- = \begin{bmatrix} (\mathbf{Z}'\mathbf{Z})^- & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \\ &+ \begin{bmatrix} -(\mathbf{Z}'\mathbf{Z})^- \mathbf{Z}'\mathbf{W} \\ \mathbf{I}_n \end{bmatrix} \begin{bmatrix} \mathbf{W}'(\mathbf{I}_n - \mathbf{P}(\mathbf{Z}))\mathbf{W} \\ \mathbf{0} \end{bmatrix} \begin{bmatrix} -\mathbf{W}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^- & \mathbf{I}_n \end{bmatrix} \end{aligned} \quad (3.8)$$

Έτσι $\mathbf{P}(\mathbf{X}_2) = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^- \mathbf{Z}' + (\mathbf{W} - \mathbf{P}(\mathbf{Z})\mathbf{W}) \begin{bmatrix} \mathbf{W}'(\mathbf{I}_n - \mathbf{P}(\mathbf{Z}))\mathbf{W} \\ \mathbf{0} \end{bmatrix} (\mathbf{W}' - \mathbf{W}'\mathbf{P}(\mathbf{Z}))$, τότε

$$\begin{aligned} \mathbf{Z}'\mathbf{Z} &= n\mathbf{I}_p, \mathbf{W}'\mathbf{W} = p\mathbf{I}_n, \mathbf{Z}'\mathbf{W} = \mathbf{J}_{p,n}, \mathbf{W}'(\mathbf{I}_n - \mathbf{P}(\mathbf{Z}))\mathbf{W} = p(\mathbf{I}_n - \frac{1}{n}\mathbf{J}_{n,n}) \\ \begin{bmatrix} \mathbf{W}'(\mathbf{I}_n - \mathbf{P}(\mathbf{Z}))\mathbf{W} \\ \mathbf{0} \end{bmatrix} &= \frac{1}{p}(\mathbf{I}_n - \frac{1}{n}\mathbf{J}_{n,n}), \end{aligned} \quad (3.9)$$

$$(\mathbf{X}'_2 \mathbf{X}_2)^- = \begin{bmatrix} \frac{1}{n}\mathbf{I}_p & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} + \frac{1}{p} \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_n - \frac{1}{n}\mathbf{J}_{n,n} \end{bmatrix} = \begin{bmatrix} \frac{1}{n}\mathbf{I}_p & \mathbf{0} \\ \mathbf{0} & \frac{1}{p}(\mathbf{I}_n - \frac{1}{n}\mathbf{J}_{n,n}) \end{bmatrix} \quad (3.10)$$

$$\mathbf{X}'_1 \mathbf{X}_2 (\mathbf{X}'_2 \mathbf{X}_2)^- \mathbf{X}'_2 \mathbf{X}_1 = \begin{bmatrix} \mathbf{X}'_1 \mathbf{Z} & \mathbf{X}'_1 \mathbf{W} \end{bmatrix} \begin{bmatrix} \frac{1}{n}\mathbf{I}_p & \mathbf{0} \\ \mathbf{0} & \frac{1}{p}(\mathbf{I}_n - \frac{1}{n}\mathbf{J}_{n,n}) \end{bmatrix} \begin{bmatrix} \mathbf{Z}'\mathbf{X}_1 \\ \mathbf{W}'\mathbf{X}_1 \end{bmatrix} \quad (3.11)$$

$$\mathbf{X}'_1 \mathbf{Z} = \begin{bmatrix} (\boldsymbol{\tau}_A - \boldsymbol{\tau}_B)' \\ (\boldsymbol{\delta}_A - \boldsymbol{\delta}_B)' \end{bmatrix} (\boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \dots, \boldsymbol{\pi}_p) = \begin{bmatrix} m(A)_1 - m(B)_1 & m(A)_2 - m(B)_2 & \dots & m(A)_p - m(B)_p \\ 0 & m(A)_1 - m(B)_1 & \dots & m(A)_{p-1} - m(B)_{p-1} \end{bmatrix}$$

$$\mathbf{X}'_1 \mathbf{W} = \begin{bmatrix} (\boldsymbol{\tau}_A - \boldsymbol{\tau}_B)' \\ (\boldsymbol{\delta}_A - \boldsymbol{\delta}_B)' \end{bmatrix} (\gamma_1, \dots, \gamma_n) = \begin{bmatrix} n(A)_1 - n(B)_1, & \dots, & n(A)_n - n(B)_n \\ \tilde{n}(A)_1 - \tilde{n}(B)_1, & \dots, & \tilde{n}(A)_n - \tilde{n}(B)_n \end{bmatrix}$$

Έτσι

$$\mathbf{X}'_1 \mathbf{X}_2 (\mathbf{X}'_2 \mathbf{X}_2)^{-1} \mathbf{X}'_2 \mathbf{X}_1 = \frac{1}{n} \mathbf{C} + \frac{1}{p} \mathbf{D} \quad (3.12)$$

$$\text{Ακόμα } \mathbf{X}'_1 \mathbf{X}_1 = \begin{bmatrix} (\boldsymbol{\tau}_A - \boldsymbol{\tau}_B)' \\ (\boldsymbol{\delta}_A - \boldsymbol{\delta}_B)' \end{bmatrix} \begin{bmatrix} \boldsymbol{\tau}_A - \boldsymbol{\tau}_B & \boldsymbol{\delta}_A - \boldsymbol{\delta}_B \end{bmatrix}$$

$$= \begin{bmatrix} n(A) + 0 + 0 + n(B) & n(AA) - n(AB) - n(BA) + n(BB) \\ n(AA) - n(AB) - n(BA) + n(BB) & \tilde{n}(A) + 0 + 0 + \tilde{n}(B) \end{bmatrix} = \mathbf{E}$$

Επομένως $\mathbf{Q} = \mathbf{X}'_1 \mathbf{X}_1 - \mathbf{X}'_1 \mathbf{X}_2 (\mathbf{X}'_2 \mathbf{X}_2)^{-1} \mathbf{X}'_2 \mathbf{X}_1$ είναι όπως στο θεώρημα 3.1. •

Θεώρημα 3.2 Οι τρεις 2×2 πίνακες $\mathbf{E}, \mathbf{C}, \mathbf{D}$ της σχέσης (3.6) είναι μη αρνητικά ορισμένοι και,

- (i) Για n άρτιο, $\mathbf{C} \geq \mathbf{0}$ με ισότητα μόνο αν ο σχεδιασμός είναι ομοιόμορφος ως προς τις περιόδους.
- (ii) Για n άρτιο, p άρτιο, $\begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix} + \begin{bmatrix} d_{11} & d_{12} \\ d_{21} & d_{22} \end{bmatrix} \geq \begin{bmatrix} 0 & 0 \\ 0 & n \end{bmatrix}$, με ισότητα μόνο αν ο σχεδιασμός είναι ομοιόμορφος ως προς τις περιόδους και τις μονάδες.
- (iii) Για n άρτιο, p περιττό $\begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix} + \begin{bmatrix} d_{11} & d_{12} \\ d_{21} & d_{22} \end{bmatrix} \geq \begin{bmatrix} n & 0 \\ 0 & 0 \end{bmatrix}$ με ισότητα μόνο αν ο σχεδιασμός είναι ομοιόμορφος ως προς τις περιόδους και τις μονάδες στις $p-1$ πρώτες περιόδους.

Απόδειξη. (i) Από τον ορισμό της σχέσης (3.6) ο πίνακας $\mathbf{X}'_1 \mathbf{Z} \mathbf{Z}' \mathbf{X}_1 = \mathbf{C}$ είναι μη αρνητικά ορισμένος και $c_{11} = 0$, μόνο όταν $m(A)_i = m(B)_i, i = 1, 2, \dots, p$ και τότε $c_{12} = c_{21} = 0, c_{22} = 0$, αυτό συμβαίνει μόνο όταν ο σχεδιασμός είναι ομοιόμορφος ως προς τις περιόδους και ακόμα $n(A) = n(B)$ και $\tilde{n}(A) = \tilde{n}(B)$.

(ii) Με τον ίδιο τρόπο ο πίνακας \mathbf{D} είναι μη αρνητικά ορισμένος και $d_{11} = 0$ μόνο αν $n(A)_i = n(B)_i, i = 1, 2, \dots, n$, αυτό συμβαίνει μόνο αν ο σχεδιασμός είναι ομοιόμορφος ως προς τις μονάδες, για p άρτιο. Εφόσον $\tilde{n}(A) = \tilde{n}(B)$, τότε $\tilde{n}(A)_i - \tilde{n}(B)_i = \pm 1 \quad i = 1, 2, \dots, n$, προκύπτει το αποτέλεσμα που ακολουθεί. Σε

αυτή την περίπτωση $\varphi(\mathbf{Q}) \leq \varphi\left(\mathbf{E} - \frac{1}{p} \begin{bmatrix} 0 & 0 \\ 0 & n \end{bmatrix}\right)$ για κάθε συνάρτηση πληροφορίας

$\varphi(\mathbf{Q})$.

(iii) Ομοιόμορφος στις μονάδες συνεπάγεται $d_{22} = 0$ μόνο αν $\tilde{n}(A)_i - \tilde{n}(B)_i = 0$, $i = 1, 2, \dots, n$, τότε $n(A) - n(B) = \pm 1$, γι' αυτό για κάθε συνάρτηση πληροφορίας έχουμε $\varphi(\mathbf{Q}) \leq \varphi(\mathbf{E} - \frac{1}{p} \begin{bmatrix} n & 0 \\ 0 & 0 \end{bmatrix})$.

Λήμμα 3.1 Αν ένας σχεδιασμός επαναλαμβανόμενων μετρήσεων είναι ομοιόμορφος ως προς τις περιόδους τότε ο n είναι άρτιος και $n(AA) = n(BB), n(AB) = n(BA)$.

Απόδειξη. Έστω $m(AA, i)$ το πλήθος των AA που εμφανίζονται στις περιόδους $i, i+1$, με τον ίδιο τρόπο ορίζονται $m(AB, i), m(BA, i), m(BB, i), i = 1, 2, \dots, p-1$, τότε

$$m(A)_i = m(AA, i) + m(AB, i), \quad m(A)_{i+1} = m(AA, i) + m(BA, i), \quad i = 1, 2, \dots, p-1 \quad (3.14)$$

Αλλά $m(A)_i = m(A)_{i+1}, i = 1, 2, \dots, p-1$, έτσι από τις (3.14), $m(AB, i) = m(BA, i)$.

Επίσης $m(B)_{i+1} = m(AB, i) + m(BB, i), i = 1, 2, \dots, p-1$, τότε, από (3.14), $m(AA, i) = m(BB, i)$.

Οπότε $n(AA) = \sum_{i=1}^{p-1} m(AA, i) = \sum_{i=1}^{p-1} m(BB, i) = n(BB)$, ομοίως $n(AB) = n(BA)$.

Θεώρημα 3.3 Αν το μοντέλο που δίνεται στην (3.3) είναι σωστό, τότε οι αναγκαίες και ικανές συνθήκες για έναν σχεδιασμό d^* να είναι Φ-βέλτιστος για την εκτίμηση των άμεσων και επαναλαμβανόμενων είναι:

(i) Αν $n = 0 \pmod{4}$ και p άρτιος, τότε ο d^* είναι ομοιόμορφος ως προς τις περιόδους και ως προς τις μονάδες και ισχυρά ισορροπημένος.

(ii) Αν n είναι άρτιος και p περιττός, τότε ο d^* είναι ομοιόμορφος ως προς τις περιόδους ως προς τις μονάδες στις πρώτες $p-1$ περιόδους και ισχυρά ισορροπημένος.

(iii) Αν $n = 2 \pmod{4}$ και p άρτιος, τότε d^* είναι ομοιόμορφος ως προς τις περιόδους, ως προς τις μονάδες και ακόμα ισχύει:

$$n(AA) - n(AB) - n(BA) + n(BB) = 2(n(AA) - n(AB)) = \pm 2$$

Απόδειξη. (i) θεωρώντας τους θετικά ορισμένους πίνακες,

$$\mathbf{Q} = \begin{bmatrix} np & n(AA) - n(AB) - n(BA) + n(BB) \\ n(AA) - n(AB) - n(BA) + n(BB) & n(p-1) - (n/p) \end{bmatrix} \quad (3.15)$$

$$Q^* = \begin{bmatrix} np & 0 \\ 0 & n(p-1) - (n/p) \end{bmatrix} \quad (3.16)$$

Αν λ_1, λ_2 είναι οι ιδιοτιμές του Q και μ_1, μ_2 οι ιδιοτιμές του Q^* , τότε $\lambda_1 + \lambda_2 = \mu_1 + \mu_2 = np + n(p-1) - (n/p)$ και $\lambda_1 \lambda_2 \leq \mu_1 \mu_2 = np((n(p-1) - (n/p)))$, γι' αυτό $(\mu_1, \mu_2) \prec (\lambda_1, \lambda_2)$, αυτό σημαίνει ότι το (μ_1, μ_2) κυριαρχείται από το (λ_1, λ_2) και ισχύει $g(\mu_1) + g(\mu_2) \leq g(\lambda_1) + g(\lambda_2)$ για όλες τις συνεχείς φθίνουσες κυρτές συναρτήσεις $g(x)$, οπότε ο d^* είναι Φ -βέλτιστος.

(ii) Στην περίπτωση που ο Q έχει $np - (n/p)$ αντί για np και $n(p-1)$ αντί για $n(p-1) - (n/p)$ η απόδειξη είναι όπως πριν και ο Φ -βέλτιστος σχεδιασμός έχει πίνακα πληροφορίας,

$$Q^* = \begin{bmatrix} np - (n/p) & 0 \\ 0 & n(p-1) \end{bmatrix} \quad (3.17)$$

(iii) Στην περίπτωση αυτή δεν μπορούμε να έχουμε $n(AA)=n(AB)=n(BA)=n(BB)$ γιατί $n(p-1)$ δεν είναι πολλαπλάσιο του 4, ο Φ -βέλτιστος σχεδιασμός έχει πίνακα πληροφορίας:

$$Q^* = \begin{bmatrix} np & \pm 2 \\ \pm 2 & n(p-1) - (n/p) \end{bmatrix} \quad (3.18)$$

ABSTRACT

Repeated Measurement Designs, with $t=2$ treatments, n (experimental) units and p periods are examined. The model is with uncorrelated observations following a continuous distribution with constant variance and the parameters of interest are (i) the difference of direct effects, (ii) the difference of residual effects. (a) The difference of Universal optimality and Φ -optimality is clarified, (b) The sufficient conditions of Cheng and Wu (1980) are extended to include the case $n \equiv 0 \pmod{2}$, p even, (c) It is shown that these conditions are also necessary for Φ -optimality for estimating direct and residual effects.

ΑΝΑΦΟΡΕΣ

Cheng, C. S. & Wu, C. F. (1980), Balanced repeated measurements designs. *Annals of Statistics*, **11**, 29-50. Correction (1983) **11**, 349.

Hedayat, A & Afsarinejad, K. (1975), Repeated measurements designs I. *A survey of Statistical Design and Linear Models*, Srivastava, J.N. (ed.), pp. 229-242.

Hedayat, A & Afsarinejad, K. (1978), "Repeated measurements designs II". *Annals of Statistics*, **18**, 1805-1816.

Kiefer, J. (1975), Construction and optimality of generalized Youden squares. *A survey of Statistical Designs and Linear Models*. J. N. Srivastava, editor, North-Holland, Amsterdam, 333-353.

Kunert, J. (1984), Optimality of balanced uniform measurements designs. *Annals of Statistics*, **12**, 1006-1017.

Kushner, H. B. (1997a), Optimality and efficiency of the two treatment repeated measurement design. *Biometrika*, **84**, 455-468. Corrigendum: *Biometrika* (1999), **86**, 234.

Kushner, H. B. (1997b), Optimal Repeated Measurement Designs: The Linear Optimality equations, *The Annals of Statistics*, **25**, 2328-2344.

Kounias, S. & Chalikias, M. (2008), Estimability of Parameters in a Linear Model *Statistics and Probability Letters*, vol. **28** issue 15, 2437-2439.

Marshall A.W. & Olkin I. (1979), *Inequalities: Theory of majorization and its applications*, Academic Press.

Pukelsheim F. (1993), *Optimal design of experiments*. John Wiley and sons, inq.

εργασίες

στα αγγλικά



BRAIN DRAIN AND BRAIN GAIN WITHIN EU27: CURRENT PICTURE AND POSSIBLE REASONS.

Yiannis C. Bassiakos

Department of Economics, University of Athens

ybassiak@econ.uoa.gr

ABSTRACT

In 2009-2010 the EC conducted a study on the mobility of European Researchers (MORE). Part of the study was the mobility of researchers in HEIs. The main finding of the MORE study is that about 56% of the University researchers have worked for, at least, three months in a country different than the one where they received their Ph.D. The MORE study treats mobility as a positive attribute fostering cross-fertilization between the members of European Union. Another side of mobility though, is brain drain or brain gain, depending on one's point of view. It is common knowledge that Greece has a sizable brain drain. It belongs to the "donor nations", so to speak. Some other members of the EU are "receiver" nations. In the present study an attempt is made to classify the 27 members of EU into "donors" and "receivers" and examine the reasons that explain such a classification. The explanatory variables examined are the per capita GDP (2009), the number of Universities in the 2009 Shanghai top 500 per 100000 students in each country, and % unemployment in each country. The initial results indicate that the per capita GDP and unemployment rate in the country of origin are positively related to the probability of brain drain and Eastern European countries and Greece are the main victims of the brain drain within the Union. The high reputation of the HEIs in the "receiving" country on the other hand has a positive impact on the brain drain. In layman's terms researchers from poorer countries immigrate to countries with more reputable HEIs and higher per capita GDP.

Keywords: brain drain, mobility, University researchers

INTRODUCTION

Brain drain refers to the emigration of skilled and professional personnel from developing countries to advanced industrial nations (Miyagiwa, 1991). The issue of brain drain is a continuous problem in world economic and sociology literature. It has also become controversial with respect to the impact the emigration of well educated and talented people has in the country of origin and its economy. According to Portes, (1976), "... seized the issue of brain drain as another manifestation of the process by which countries at the center of the capitalist economy extract surplus and resources

from those in the periphery. Emigration is, from this standpoint, the human equivalent and reflection of international arrangements through which weaker economies are subordinated to the interests and priorities of the stronger ones". A mathematical modeling of the emigration process and driving forces appears in Bhagwati & Rodriquez (1975).

In later years, a new approach is appearing in the literature. According to this approach, brain drain has also positive effects: the prospect of migration drives people to improve their skills via education and thus the overall level of skills and expertise rises in the country of origin, helping its development and innovation, (Agrawal et al., 2001, Beine et al., 2001, Stark 2004, Stark et al., 1997).

These contradictory approaches raise the question of who benefits the most from highly educated professionals migration. It is logical to assume that the higher benefits go to the countries which accept the immigrants, since they utilize the abilities and talents of the professionals, without having to pay for their education. Whereas the countries which send the immigrants have a net loss, since the education they have paid for does not produce its full benefit within the country itself.

By explaining the forces behind this type of immigration, one can take a step towards resolving the conflict: if poor economic conditions in the country of origin drive immigration, then it is more likely that such immigration exacerbates its economic problems. The concept of cross-fertilization of knowledge included in the Lisbon Treaty as a goal, will be better served by a bidirectional and equivalent flow of "brains", not by a unidirectional and unequal one.

METHODS

In the present study, an attempt is made to quantify and explain the brain drain within EU27. The data come from a study on the mobility of European Researchers conducted in 2009-2010 on behalf of the EC (MORE project). This was part of a greater effort by EC to implement the Lisbon Treaty and measure its effects. Part of the MORE study was the mobility of researchers in HEIs. The main finding of the MORE study is that about 56% of the University researchers have worked for, at least, three months in a country different than the one where they received their Ph.D. The MORE study treats mobility as a positive attribute, fostering scientific cross-fertilization between the members of European Union.

The focus will be on a subset of data, concerning the outward (from the country of birth) mobility of HEI personnel. It is common knowledge, for example, that Greece has a sizable brain drain. It belongs to the "donor nations", so to speak. Some other members of the EU are "receiver" nations. In the present study a person is defined as "immigrant" if he or she was born in a country different from the one he or she is working at present. The probability of being immigrant is the main dependent variable of the study. As explanatory variables, per capita GDP (2009), number of Universities in the 2009 Shanghai top 500 per 100000 students, and % unemployment, are examined along with dummy variables for the country of origin

and the country of current employment. The relation between the dependent and the explanatory variables will be ascertained using stepwise logistic regression modeling the probability of migration using the following equation:

$\log\left(\frac{P}{1-P}\right) = \beta_1 \cdot X_1 + L + \beta_n \cdot X_n$ In this equation, $P = P(Y = 1)$, where $Y=1$ means “works in a country different from the one he/she was born in” and the X_i represent the 30 explanatory variables, which are 27 binary for the country of origin (e.g. 1 = born in Greece, 0 = born elsewhere) and three continuous variables (difference in $\ln(\text{per person GDP})$), difference in the number of HEI’s in the top 500 Shanghai index per 100000 students and unemployment in the country of origin). Consequently, when a factor achieves statistical significance, the impact measured represents the effect of the presence of the factor against its absence for each combination of all other factors. For example, if Greece (= born in Greece) is statistically significant, it represents the additional impact of “born in Greece” at the mean level of the three continuous variables. Solving the above equation for P , after estimating the β_i s, we obtain an estimate of the proportion of probability of the event attributable to each factor i , when all other factors are held at any fixed level. The model used contains no constant term, since in multiple application over different variable sets the no intercept model had much better fit. The model above is the initial model fitted. The procedure removes the non significant explanatory variables one at a time, until all variables in the model are statistically significant. The data will be presented as mean \pm standard deviation or percents, as applicable.

RESULTS

Immigration patterns of HEI personnel between the EU27 states are dictated by a variety of reasons. Table 1 presents the picture for the top five countries of origin (i.e. with the absolutely highest negative difference) and the top five destination countries (i.e. with the highest positive difference). Countries with very small number of data should be ignored. What becomes evident from this table is that Greece (33%), Bulgaria (23.5%), Slovakia (17.5%), Romania (13%) and Hungary (12.7%) are the biggest “exporters”, while the United Kingdom (9.5%), Sweden (6.1%), Denmark (7.3%), the Netherlands (14.1%) and France (17.6%) are the biggest “importers” of HEI personnel. The overall immigration percent is 12.7%. It is worth noting that France and the Netherlands, while having a higher than average percentage of migration, also have a much higher percentage of researchers migrating towards them (22.5% and 20.4%, respectively). On the other hand Hungary and Romania do not have a higher than average percentage of migration (12.7% and 13%, respectively), but have no researchers migrating towards them.

Table 1. Counts and percents of HEI personnel send and received by the top five EU27 states (in each category: exporters and importers).

	SEND				RECEIVE				difference
	NO		YES		NO		YES		
	Count	%	Count	%	Count	%	Count	%	
United Kingdom	341	90.5	36	9.5	341	70.3	144	29.7	20.1
Sweden	230	93.9	15	6.1	230	77.7	66	22.3	16.2
Denmark	127	92.7	10	7.3	127	84.7	23	15.3	8.0
Netherlands	183	85.9	30	14.1	183	79.6	47	20.4	6.4
France	155	82.4	33	17.6	155	77.5	45	22.5	4.9
Hungary	48	87.3	7	12.7	48	100.0	0	0.0	-12.7
Romania	94	87.0	14	13.0	94	100.0	0	0.0	-13.0
Slovakia	47	82.5	10	17.5	47	95.9	2	4.1	-13.5
Bulgaria	52	76.5	16	23.5	52	100.0	0	0.0	-23.5
Greece	63	67.0	31	33.0	63	94.0	4	6.0	-27.0

In an attempt to explain this picture, the probability of immigration was modeled on 27 dummy variables corresponding to each one of the EU27 states, as well as three continuous variables, the difference of natural logarithms of the per person GDP (Eurostat Statistical Books, 2010 and EU-27 population data), of the country of origin minus the country of destination, the number of Universities in the 2009 Shanghai top 500 per 100000 students in the country of destination, and % unemployment (ages 14 to 25) in the country of destination. The relevant values for 2009 are presented in Table 2.

Table 2: Per capita GDP, Unemployment rate (ages 14-25), and University score in all EU-27 states.

COUNTRY	per capita GDP	Unemployment rate (ages 14-25)	University score
Austria	33140	4.8	2.7
Belgium	31366	7.9	1.75
Bulgaria	4454	6.8	0
Czech Republic	13108	6.7	0
Cyprus	21267	5.3	0.3
Denmark	40442	6	1.7
Estonia	10243	13.8	0
Finland	32099	8.2	1.6
France	29628	9.5	1
Germany	29378	7.5	1.75
Greece	21091	9.5	0.35

Hungary	9280	10	0.4
Ireland	36751	11.9	1.55
Italy	25329	7.8	1
Latvia	8300	17.1	0
Lithuania	7956	13.7	0
Luxembourg	76505	5.4	0
Malta	13810	6.9	0
Netherlands	34588	3.4	2
Poland	8131	8.2	0.1
Portugal	15422	9.6	0.5
Romania	5390	6.9	0
Slovakia	11702	12	0
Slovenia	17169	5.9	0.8
Spain	22937	18	0.6
Sweden	31675	8.3	2.7
United Kingdom	25436	7.6	1.7

The results of the (stepwise) logistic regression through the origin are presented in Table 3 below. The model fit of the logistic regression is satisfactory: Nagelkerke's R^2 is equal to 0.697 and, while the Hosmer & Lemeshow statistic is large (and statistically significant, p -value < 0.0005), this is a result of some of the cells of the classification table having very small expected values. The correct classification rate is 88.2%, indicating a rather good fit.

Table 3.

	B (slope)	Standard error of B	significance	Exp(B)
Germany	-1.956	0.179	0.00000000	0.141
UK	-1.841	0.223	0.00000000	0.159
Denmark	-1.425	0.385	0.00021623	0.241
Austria	-1.388	0.318	0.00001246	0.249
Sweden	-1.324	0.349	0.00014678	0.266
Hungary	-1.255	0.592	0.03404039	0.285
ln(per capita GDP) difference	-1.248	0.226	0.00000003	0.287
Finland	-1.198	0.421	0.00439222	0.302
France	-1.095	0.219	0.00000059	0.335
Portugal	-0.994	0.405	0.01403139	0.370
Czech Republic	-0.930	0.432	0.03133109	0.394
Belgium	-0.659	0.290	0.02322235	0.517

Unemployment (ages 15-24, destination)	-0.096	0.004	0.00000000	0.908
Greece	0.519	0.247	0.03571477	1.680
University rating (destination)	0.671	0.093	0.00000000	1.957
Cyprus	2.402	0.951	0.01152436	11.050

The logistic model results have some interesting points: Taking under consideration per person GDP, HEI quality and unemployment, Greece and Cyprus are net “exporters”, while Germany, the United Kingdom, Denmark, Austria, Sweden, Hungary, Finland, France, Portugal, the Czech Republic and Belgium are “importers”. Germany, Hungary, Finland, Portugal, the Czech Republic and Belgium have “changed sides”, in the presence of the additional explanatory variables. The GDP differences, unemployment rates in the 14-25 age group and HEI quality: Higher HEI quality in the country of destination increases the probability of emigration, while higher GDP and unemployment rate in the country of destination decrease the probability of emigration.

CONCLUSIONS

The clear impact of the difference in per capita GDP and destination country unemployment rate in the migration rate, tends to support the claim of the researchers that brain drain has a negative impact in the countries of origin. In order to draw more solid conclusions on this subject one should conduct a study using time series type data and compare the rate of increase of the per capita GDP between “importer” and “exporter” states, controlling for variables known to affect GDP, other than tertiary education. An interesting observation is the importance of the quality of the HEIs: High quality educational institutions attract students from other countries. This observation emphasizes the importance of investing in higher education since such investment not only benefits the “investor” via the knowledge production and application mechanisms, but also can bring a direct profit from the attraction of foreign students to the universities.

Acknowledgements: The author wishes to express his gratitude to the anonymous referees for their valuable and insightful comments, which helped improve the present work.

ΠΕΡΙΛΗΨΗ

Στο διάστημα 2009-2010 η ΕΕ διεξήγαγε μελέτη σχετικά με την κινητικότητα των ευρωπαίων ερευνητών (MORE). Μέρος της μελέτης MORE ήταν η κινητικότητα των ερευνητών στα ΑΕΙ. Το κύριο εύρημα της μελέτης MORE είναι ότι περίπου το 56% των ερευνητών του Πανεπιστημίου έχουν εργαστεί επί, τουλάχιστον, τρεις μήνες σε χώρα διαφορετική όα αυτή όπου έλαβαν το διδακτορικό τους. Η μελέτη MORE αντιμετωπίζει την κινητικότητα ως θετικό χαρακτηριστικό για την προώθηση της μεταφοράς γνώσης μεταξύ των μελών της Ευρωπαϊκής Ένωσης. Μια άλλη πλευρά της κινητικότητας εν τούτοις, είναι η διαρροή επιστημονικού κεφαλαίου ή το κέρδος επιστημονικού κεφαλαίου, ανάλογα με την οπτική γωνία του καθενός. Είναι γνωστό ότι η Ελλάδα έχει μια αρκετά μεγάλη διαρροή επιστημονικού κεφαλαίου. Ανήκει στα «έθνη-δότες», να το πω έτσι. Κάποια άλλα μέλη της ΕΕ είναι «υποδοχείς» επιστημονικού κεφαλαίου. Στην παρούσα μελέτη γίνεται μια προσπάθεια να εντοπιστούν τα

έθνη «υποδοχείς» και τα έθνη «δότες» της ΕΕ και να εξετασθούν οι λόγοι που ερμηνεύουν το χαρακτηρισμό αυτό. Οι ερμηνευτικές μεταβλητές που εξετάζονται είναι το κατά κεφαλήν ΑΕΠ (2009), ο αριθμός των Πανεπιστημίων στα κορυφαία 500 ανά 100.000 φοιτητές σε κάθε χώρα, (σύμφωνα με την κατάταξη του 2009, κατάλογος Σαγκάης), και το ποσοστό ανεργίας στις ηλικίες 15-24 ετών σε κάθε χώρα. Η διαφορά του κατά κεφαλήν ΑΕΠ (χώρας προέλευσης – χώρας υποδοχής) και ποσοστό ανεργίας στη χώρα υποδοχής σχετίζονται αρνητικά με την πιθανότητα της διαρροής εγκεφάλων από τη χώρα προέλευσης. Η μεγάλη φήμη των ΑΕΙ στη χώρα υποδοχής από την άλλη πλευρά έχει θετικό αντίκτυπο στη «διαρροή επιστημονικού κεφαλαίου».

Acknowledgments: The author wishes to thank the DG Research of the European Commission for the permission to use the data set which was developed as a part of the project “Study on the mobility and career paths of EU researchers (contract no. RTD/C/C4/2008/0203349). The project was funded by the DG Research of the European Commission.

REFERENCES

- Agrawal A., Kapur D., McHale J. and Oettl A., (2011), *Brain Drain or Brain Bank? The impact of skilled emigration on poor-country innovation*, Journal of Urban Economics, Vol. 69, pp. 43-55.
- Beine M., Docquier F. and Rapoport H., (2001), *Brain Drain and Economic growth: theory and evidence*, Journal of Development Economics, Vol. 64, pp. 275-289.
- Bhagwati J., Rodriguez C., Welfare-(1975), *Theoretical Analyses of the Brain Drain*, Journal of Development Economics, Vol. 2, pp. 195-221.
- Eurostat Statistical Books: European Economic Statistics, 2010 Edition.
http://epp.eurostat.ec.europa.eu/cache/ITY_OFFPUB/KS-GK-10-001/EN/KS-GK-10-001-EN.PDF
- EU-27 population data
http://epp.eurostat.ec.europa.eu/portal/page/portal/population/data/main_tables
- Miyagiwa K., (1991), *Scale Economies in Education and the Brain Drain Problem*, International Economic Review, Vol. 32, No. 3, pp. 743-759.
- Portes A. (1976), *Determinants of the Brain Drain*, International Migration Review, Vol. 10, No. 4, pp. 489-508.
- Stark O. (2004), *Rethinking the Brain Drain*, World Development, Vol. 32, No. 1, pp. 15-22.
- Stark O., Helmenstein C. and Prskawetz A., (1997), *A brain gain with a brain drain*, Economics Letters, Vol. 60, pp. 363-367.
- Stark O., Helmenstein C. and Prskawetz A., (1998), *Human capital depletion, human capital formation, and Migration: a blessing or a “curse”?* Economics Letters, Vol. 60, pp. 363-367.
- Vidal J. P., (1998), *The Effect of Emigration on Human Capital Formation*, Journal of Population Economics, Vol. 11, pp. 589-600.



DO THE GOALS PER GAME IN 2010 WORLD CUP OF SOCCER FOLLOW THE POISSON DISTRIBUTION?

Vassilios C. Hombas
University of Athens, Greece
c.hombas@gmail.com

SUMMARY

The soccer match dataset containing information on 64 fixtures of the 2010 World Cup is discussed. A “goodness of fit” is carried out in order to find whether the data on the number of goals per game fit a Poisson distribution.

Keywords: *Teaching, World Cup, FIFA.*

1. INTRODUCTION

The World Cup of Soccer takes place every four years. The one in 2010 was organized by South Africa from 11 June to 10 July 2010. The first round contains 48 matches in 8 groups of six and the matches are selected at random. The 2nd round contains 16 teams in one group of 8 matches. The quarterfinal contains 8 teams in a group of 4 matches. The 3rd play-off contains only two matches and it is for the 3rd position. Finally, the final match is for the champion.

In World Cup 2010, Spain was the champion. We consider that the rules of soccer are known and we don't describe them here.

Our investigation is based on FIFA (Federation International de Football Associations) [1], World Cup 2010 results and the goals list given at the end of the article. From this we created Table 1. The mean of goals per game scored by the 32 teams in the group matches, that is 64 matches yield 157 goals; hence $\bar{x} = 2.4531$.

Table 1. Observed distribution and Poisson expected frequencies

Observed frequency		Poisson distribution with $\lambda = \bar{x}$	
Goals / game	Observed Frequencies	Probabilities	Expected Frequencies
(i)	(f _i)	(P _i)	(e _i)
0	6	0.086058	5.51
1	16	0.211110	13.51
2	13	0.258937	16.57
3	15	0.211732	13.55

4	6	0.129850	8.31
5	5	0.063707	4.07
6	1	0.026046	1.66
7	1	0.009127	0.58
≥ 8	1	0.002799	0.18
157			

Our aim is to test whether the goals per game may be looked upon as values of a random variable having a Poisson distribution. This is a standard goodness of fit problem.

A difficulty arises when λ is unknown. This can be circumvented by means of a theoretical result which permits the replacement of λ by its maximum likelihood estimate $\bar{x} = \hat{\lambda} = 2.4531$, provided one degree of freedom is deducted from the estimated parameter (see Freund (1971, p. 337)).

A second problem has been raised now: Are the 64 random realizations independent, when games are played at different venues, at different times of the week or days? This is quite possible. Here we suppose that our random realizations are independent, though in reality the variables are rarely independent.

2. Discussion of Independence

A problem with this assumption would arise if there are reasons to believe that there is some sort of strategic interaction between these matches, i.e. if the score in one match affects the score of another. In practice, great care has to be taken to avoid erroneously applying the Poisson distribution. A good way to understand how such interaction can be achieved is to review the regulations of the competitions. We can see all regulations in the Reference [1]. The possibility of observing strategic interactions between fixtures is stronger during the first round when the league scoring system – three points for a win, one point for a draw and none for a defeat – is applied.

The scoring system or ranking clearly creates the prerequisites for strategic interaction between the matches. If one of the two teams is not motivated to play full speed, the other team's success depends on a particular score. The probability of that score's realization is higher than it would have been if the game was considered in isolation. One may argue that the strategic interactions have more to do with who wins a match, but less so for the random variable that we consider here – the number of goals.

Another factor that may cast doubt on the identical distribution of the random trials is team identity. Football is a cultural phenomenon that reflects thorough differences in the social fabric across countries and regions. An example is the Latin American football and, most notably, Brazilian football is slower but much more technical than the football played in Europe.

3. THE TEST

There are many approaches to test whether a set of observed frequencies is compatible with a set of expected frequencies. A very good approach for this case is to apply the Pearson χ^2 statistic :

$$\chi^2 = \sum_{i=1}^k (f_i - e_i)^2 / e_i \quad (1)$$

which follows a χ^2 (chi-square) distribution with $k-1$ degrees of freedom and using the right tail of the χ^2 distribution as a critical region; i.e. reject the null hypothesis, if the value χ^2 of χ^2 is greater than $\chi_{\alpha, k-1}^2$ ($\chi^2 > \chi_{\alpha, k-1}^2$) where $\chi_{\alpha, k-1}^2$ is the $1-\alpha$ quantile of the chi-square distribution with $k-1$ degrees of freedom.

Here $\chi^2 = 6.593$, $k=9$, λ is the only estimated parameter, therefore the degrees of freedom would be $k-1-1=7$.

For $\alpha=0.05$ the critical point is $\chi_{0.05, 7}^2 = 14.067$. Hence, since $\chi^2 = 6.593 < \chi_{0.05, 7}^2 = 14.067$, there is not enough evidence in the data to reject the null hypothesis. Consequently the Poisson distribution provides a good fit.

4. CONCLUDING REMARKS

I believe that this genuine example is very useful in teaching χ^2 the goodness of fit test. This dataset is appropriate for an introductory statistics course; it can be used in various ways and for various purposes. I have found this data set to be rich in examples, particularly in constructing random variables.

Acknowledgements: I thank Professors Takis Papaioannou and Theofilos Kakoullos for corrections and their helpful suggestions.

Περίληψη

Συζητούνται τα δεδομένα των 64 ποδοσφαιρικών παιχνιδιών του Παγκοσμίου Κυπέλλου του 2010. Χρησιμοποιείται ο «έλεγχος καλής προσαρμογής χ^2 » και συμπεραίνεται ότι ο αριθμός των τερμάτων (goals) των παιχνιδιών ακολουθεί την κατανομή Poisson.

REFERENCES

- [1] FIFA website: <http://www.fifa.com/worldcup/matches/index.html>
UEFA Euro 2004: website: <http://www.uefa.com/>
- [2] Freund, J. (1971). Mathematical Statistics 2nd Ed. Prentice/Hall, New Jersey.

FIFA WORLD CUP 2010 RESULTS AND GOALS

<u>A/A</u>			<u>RESULT</u>		<u>No of Goals</u>
1	Final	Holland	0-1 AET	Spain	1
2	Third Place	Uruguay	2-3	Germany	5
3	Semifinal	Spain	1-0	Germany	1
4	Semifinal	Holland	3-2	Uruguay	5
5	Quarterfinal	Paraguay	0-1	Spain	1
6	Quarterfinal	Argentina	0-4	Germany	4
7	Quarterfinal	Uruguay	1-1/p.4-2	Ghana	6
8	Quarterfinal	Holland	2-1	Brazil	3
9	Round of 16	Spain	1-0	Portugal	1
10	Round of 16	Paraguay	0-0/p.5-3	Japan	8
11	Round of 16	Brazil	3-0	Chile	3
12	Round of 16	Holland	2-1	Slovakia	3
13	Round of 16	Argentina	3-1	Mexico	4
14	Round of 16	Germany	4-1	England	5
15	Round of 16	USA	1-2 AET	Ghana	3
16	Round of 16	Uruguay	2-1	South Korea	3
17	WC Group H	Chile	1-2	Spain	3
18	WC Group H	Switzerland	0-0	Honduras	0
19	WC Group G	North Korea	0-3	Ivory Coast	3
20	WC Group G	Portugal	0-0	Brazil	0
21	WC Group E	Cameroon	1-2	Holland	3
22	WC Group E	Denmark	1-3	Japan	4
23	WC Group F	Paraguay	0-0	New Zealand	0
24	WC Group F	Slovakia	3-2	Italy	5
25	WC Group D	Australia	2-1	Serbia	3
26	WC Group D	Ghana	0-1	Germany	1
27	WC Group C	Slovenia	0-1	England	1
28	WC Group C	USA	1-0	Algeria	1
29	WC Group B	Greece	0-2	Argentina	2
30	WC Group B	Nigeria	2-2	South Korea	4
31	WC Group A	Mexico	0-1	Uruguay	1
32	WC Group A	South Africa	2-1	France	3
33	WC Group H	Spain	2-0	Honduras	2
34	WC Group H	Chile	1-0	Switzerland	1
35	WC Group G	Portugal	7-0	North Korea	7
36	WC Group G	Brazil	3-1	Ivory Coast	4
37	WC Group F	Italy	1-1	New Zealand	2
38	WC Group F	Slovakia	0-2	Paraguay	2
39	WC Group E	Cameroon	1-2	Denmark	3
40	WC Group D	Ghana	1-1	Australia	2
41	WC Group E	Holland	1-0	Japan	1
42	WC Group C	England	0-0	Algeria	0
43	WC Group C	Slovenia	2-2	USA	4
44	WC Group D	Germany	0-1	Serbia	1
45	WC Group A	France	0-2	Mexico	2
46	WC Group B	Greece	2-1	Nigeria	3
47	WC Group B	Argentina	4-1	South Korea	5
48	WC Group A	South Africa	0-3	Uruguay	3
49	WC Group H	Spain	0-1	Switzerland	1
50	WC Group H	Honduras	0-1	Chile	1
51	WC Group G	Brazil	2-1	North Korea	3
52	WC Group G	Ivory Coast	0-0	Portugal	0
53	WC Group F	New Zealand	1-1	Slovakia	2
54	WC Group F	Italy	1-1	Paraguay	2
55	WC Group E	Japan	1-0	Cameroon	1
56	WC Group E	Holland	2-0	Denmark	2
57	WC Group D	Germany	4-0	Australia	2
58	WC Group D	Serbia	0-1	Ghana	1
59	WC Group C	Algeria	0-1	Slovenia	1
60	WC Group C	England	1-1	USA	2
61	WC Group B	Argentina	1-0	Nigeria	3

62	WC Group B	South Korea	2-0	Greece	2
63	WC Group A	Uruguay	0-0	France	0
64	WC Group A	South Africa	1-1	Mexico	2

Remark:

Rows 8 and 11 after the result, have a letter (p). This means these results are decided after penalties kicks. Penalties kick if the match finishes at 30 minutes extra time. The penalties are kicked one by one with the first kick decided by tossing a coin. Row 16 has an abbreviation AET. This means "At Extra Time". Using penalties the score was 1-1, 5-5 for rows 8 and 11 respectively. Using AET the score was 0-1, 1-2 for rows 1 and 16 respectively.



ADAPTIVE PHASE II DESIGN FOR EVALUATING SURVIVAL PROBABILITIES

S. Pouloupoulou¹, D. Karlis¹, U. Dafni²

¹Athens University of Economics and Business, Athens, Greece

{spoulopo, karlis}@aueb.gr

²University of Athens, Athens, Greece

udafni@nurs.uoa.gr

ABSTRACT

The suggested Phase II multi-stage designs (Lin et al., 1996, Case and Morgan, 2003, Jennison and Turnbull 2000) for monitoring survival probabilities focus on testing the null hypothesis $H_0 : S(x) = S_0(x)$ versus the alternative $H_A : S(x) > S_0(x)$ at significance level α and power $1 - \beta$ when $S(x) = S_A(x)$ for chosen targeted level $S_A(x)$, where x denotes the survival time of interest and $S(\cdot)$ denotes the survival function. In early phase clinical trials we often face the uncertainty of whether we have made the right choice for a specific target level $S_A(x)$. If we are too optimistic we will reject a potentially promising therapy. If we are too conservative, we will require more patients than necessary.

For that purpose, an adaptive multi-stage procedure for evaluating survival probability is proposed, in which by using the information of the previous stages, adjusts the target survival level in each stage, while the type I error rate is controlled. Also, this hypothesis creates confusion among clinicians and, in fact in many statisticians, since rejecting H_0 and specifying $S(x) = S_A(x)$ in power calculations cannot be considered equivalent that the survival probability of the particular therapy is greater than $S_A(x)$. To address this issue, in the proposed class of designs we test two hypotheses sequentially in each stage, one for testing if the therapy has unacceptably low therapeutic activity and the other for testing if the therapy has acceptable enough activity to warrant further study. For the first stage we use an optimistic target level $S_A(x)$ and for every other stage except from the last we test a less optimistic target level ($S_i(x) < S_A(x)$). For the last stage k we test only one hypothesis that corresponds to the non-interesting level $S_0(x)$.

During the design stage, estimates of survival probabilities can be obtained by making the assumption that expected survival time follows an exponential distribution or variants. Under specific upper bounds for the error levels, and the assumption that the accrual rate is fixed, we develop designs that minimize the expected study length (*ESL*).

Keywords: adaptive designs; clinical trials; phase II; multi-stage; sequential testing; survival probabilities;

1. INTRODUCTION

The primary goal of a phase II clinical trial is to determine whether a new regimen has sufficient activity to warrant further study. The traditional phase II cancer trials use response rate as the primary endpoint (Fleming (1982), Simon, (1989)), which is defined by whether the treatment has shrunk the tumour to a pre-specified point at a given time point. Recently, biotechnology has led to the discovery of many promising anticancer agents, which can prevent the growth of the tumor but may not kill the tumour cells and lead to shrinkage of the tumour. This leads to increased interest in using time-to-event endpoints such as overall survival or progression-free survival for phase II evaluation.

A similar design strategy with standard phase II trial designs (Fleming (1982), Simon, (1989)) could be used by substituting survival or progression-free survival with response rate. In this design, if one wanted to focus on survival at a specific time point x , he could compare survival rate of those patients who were enrolled at least x -years prior to interim analysis with historical rates. If we don't manage to observe all the patients for the full period, a simple proportion of those who survived the specific time period would be a biased estimate of the survival rate, and omitting those patients with potential follow-up time less than x is inefficient.

Thus, another approach was proposed using the Kaplan-Meier or Nelson-Aalen estimates of the survival probability at x -years during an interim analysis, once some patients have been enrolled for at least x -years. Jennison and Turnbull, (2000) suggest monitoring such a trial using Kaplan-Meier estimates and the spending function approach originally described by Lan and DeMets (1983). Lin et al (1996) have further developed group sequential procedures for monitoring the survival probability using Nelson-Aalen estimates during the interim analysis. Where, Case and Morgan (2003) by applied results presented by Lin et al (1996) develop optimal two-stage designs for evaluating survival probabilities that minimize either the expected duration of accrual or the expected study length under H_0 .

The suggested Phase II multi-stage designs (Lin et al., 1996, Case and Morgan, 2003, Jennison and Turnbull 2000) for monitoring survival probabilities at a specific time point x , focus on testing the hypothesis: $H_0 : S(x) = S_0(x)$ versus $H_A : S(x) > S_0(x)$ at significance level α . Based on hypothesis testing framework, the value $S_A(x)$ is used as the value of the survival probability that needs to be detected as active with high probability (power). The targeted level $S_A(x)$ is pre-specified so that $S_0(x)$ is chosen to represent the survival probability achievable with standard treatment and $S_A(x)$ is chosen so that the difference $S_A(x) - S_0(x)$ represents a targeted improvement with the new treatment. These designs allow early termination of the trial in the interim analysis, only if a treatment shows unacceptably low therapeutic activity, and allow acceptance of the agent only at the last stage. Also, it should be clear, that rejecting H_0 and specifying $S(x) = S_A(x)$

in power calculations cannot be considered equivalent that the survival probability of the particular therapy is greater than $S_A(x)$. This fact creates confusion among clinicians and, in fact in many statisticians, since the conclusion that a particular therapy is effective enough to warrant further study is based only on rejecting the hypothesis $S(x) = S_0(x)$ and not on not rejecting the hypothesis $S(x) \geq S_A(x)$. This may introduce ambiguity in the evaluation of type I and II errors and the choice of the appropriate practical decision at the end of the study.

In early phase clinical trials, we often face the uncertainty of whether we have made the right choice for a specific target survival level $S_A(x)$. If the expected survival level is too optimistic we will reject a potentially promising therapy. Otherwise, if the expected survival level is too conservative, more patients would be required. A natural way to resolve this dilemma is to choose $S_A(x)$ with some flexibility. For that purpose, we propose an adaptive multi-stage procedure that using the information from the previous stages, adjust the target survival level in each stage, while the type I error rate is controlled. In the proposed class of designs we adopt the sequentially testing (TST) procedure within each stage for evaluating survival probabilities proposed by Poulopoulou et al. (2009), to address the common misconception that rejection of the null hypothesis provides definitive evidence for the veracity of the alternative hypothesis and the resulting confusion surrounding the appropriate decision at the end of the trial. We test two hypotheses sequentially in each stage, one for testing if the therapy has unacceptably low therapeutic activity and the other to test if the therapy has acceptable enough activity to warrant further study. For the first stage we use an optimistic target level $S_A(x)$ and for every other stage except from the last we test a less optimistic target level ($S_i(x) < S_A(x)$). For the last stage k we test only one hypothesis that corresponds to the non-interesting level $S_0(x)$.

In Section 2, we provide details about sequential analysis of survival function and in Section 3, the characteristics of the proposed adaptive hypothesis phase II design for survival probabilities. In Section 4, we derive optimal proposed designs by minimizing *ESL* subject to constraints on the errors, assuming variety survival distributions. Finally, concluding remarks and extensions of the proposed methods are presented in Section 5.

2. SEQUENTIAL ANALYSIS OF SURVIVAL FUNCTION

Due to ethical concerns, in time to event phase II designs, patients are entering in k stages with sample size n_1, n_2, \dots, n_k , respectively or corresponding stages duration t_1, t_2, \dots, t_k , respectively. Assume $n = \sum_{i=1}^k n_i$ patients are accrued to a trial at times Y_1, \dots, Y_n . Let T_1, \dots, T_n denote the failure times and C_1, \dots, C_n the censoring times since study entry. When the data are examine at calendar time t , we observe for each individual i either a failure time or a censoring time and an indicator specifying which. That is, we observe the time $X_i(t) = \min [T_i, C_i, \max(0, (t - Y_i))]$ and the failure indicator $\Delta_i(t) = I$

$\{T_i \leq \min[C_i, \max(0, (t - Y_i))]\}$. The Nelson-Aalen estimate of the cumulative hazard function $\Lambda(x; t)$ at time t based on the data available at the analysis is:

$$\widehat{\Lambda}(x; t) = \sum_{\{i: X_i(t) \leq x\}} \frac{\Delta_i(t)}{R_i(t)},$$

where $R_i(t) = \sum_{j=1}^n I\{X_j(t) \geq X_i(t)\}$. For a fixed t it is well-known that the process $n^{1/2} \left\{ \widehat{\Lambda}(x; t) - \Lambda(x) \right\}$, $x < t$, converges weakly to a zero-mean Gaussian process with independent increments, which assuming constant accrual has variance function (Lin et al. (1996)):

$$\sigma^2(x; t) = \int_0^x \frac{\lambda(u) du}{S(u) f(u) \min \left[\frac{t-u}{\min\{t, MDA\}}, 1 \right]}$$

where $\lambda(u)$ denotes the hazard function and $f(u)$ is the probability function of loss to follow up.

At different interim looks ($t_l \neq t_m$), Lin et al. (1996) show that the two parameter process $n^{1/2} \left\{ \widehat{\Lambda}(x; t) - \Lambda(x) \right\}$, $x < t$, converges weakly to a zero-mean Gaussian process $W(x; t)$ with $\text{cov} \{W(x; t_l), W(x; t_m)\} = \sigma^2 \{x, \max(t_l, t_m)\}$. Due to the limited number of patients who are eligible for a given phase II study as well as to the generally large number of treatment regimens awaiting phase II trials, the sample size n for a given trial is kept relatively small and frequently fixed in advance. This fact leads to same slight derivation of Nelson-Aalen estimator from the normality (Bie et al, 1987). A considerable improvement can be obtained by applying one transformation of Nelson-Aalen estimator, like logarithm transformation (Bie et al, 1987). Therefore, in order to design and analyze sequential trials, the joint distribution of $\log(\widehat{\Lambda}(x; t))$ calculated over the course of the study is multivariate normal with mean $\boldsymbol{\mu} = \log(\Lambda(x))$ and covariance matrix $\boldsymbol{\Sigma}$, where covariates between $\widehat{\Lambda}(x; t_l)$ and $\widehat{\Lambda}(x; t_m)$ given by

$$\sigma^2 \{x, \max(t_l, t_m)\} / \widehat{\Lambda} \{x, \max(t_l, t_m)\}$$

divided by the total sample size until this time period.

3. PROPOSED ADAPTIVE MULTI-STAGE DESIGNS

In the proposed multi-stage design for evaluating survival probability under the two sequential testing procedure within each stage (Poulopoulou et al., 2009), starting with an optimistic target survival level $S_A(x)$ in the first stage we adjust the target survival level in each stage, by using the information of the previous stages, so that we test a less optimistic target survival level. More specific, for each stage except the last ($i=1, \dots, k-1$): We first test if the survival probability of x failure time of interest, is at most equal to an uninteresting level $S_0(x)$ ($H_{01i} :$

$S(x) \leq S_0(x)$ versus $H_{A1i} : S(x) > S_0(x)$). If we reject this null hypotheses H_{01i} , then we proceed to test the second hypotheses regarding to the specific targeted survival level $S_{Ai}(x)$ ($H_{02i} : S(x) \geq S_{Ai}(x)$ versus $H_{A2i} : S(x) < S_{Ai}(x)$), where $S_A(x) = S_{A1}(x) > S_{A2}(x) > \dots > S_{A_{k-1}}(x) > S_0(x)$. At the last stage k , having not terminate early the study, we test only one hypothesis that corresponds to the non-interesting level $S_0(x)$ ($H_{01k} : S(x) \leq S_0(x)$ versus $H_{A1k} : S(x) > S_0(x)$).

An interim analysis could be implemented any time after x -years and before MSL, where MSL denote the maximum study length ($MSL = \sum_{i=1}^k t_i + x$). When the accrual duration of each intermediate stage is complete, a decision is made either to terminate the trial or to continue to the next stage according to the following decision rule: For each stage except the last, $i = 1, \dots, k - 1$, H_{01i} and H_{02i} denote the hypotheses H_{01} and H_{02} respectively at the i^{th} stage.

- If $\log(\widehat{\Lambda}(x; t_i)) \geq u_i$, stop sampling and not reject $H_{01i} : S(x) \leq S_0(x)$, i.e., consider the therapy ineffective.
- If $\log(\widehat{\Lambda}(x; t_i)) < u_i$, proceed with testing the second hypothesis $H_{02i} : S(x) \geq S_i(x)$:
 - If $\log(\widehat{\Lambda}(x; t_i)) \leq \ell_i$, stop sampling and not reject $H_{02i} : S(x) \geq S_i(x)$, i.e., consider the therapy effective.
 - If $\log(\widehat{\Lambda}(x; t_i)) > \ell_i$, continue to the next stage $i + 1$.

where ℓ_i and u_i denotes the cut-off points of interest. Note that $\ell_i < u_i$ for all i and $S_A(x) = S_{A1}(x) > S_{A2}(x) > \dots > S_{A_{k-1}}(x) > S_0(x)$. For the last stage k , when they past $x - years$ from the completion of the accrual duration we terminate the trial and:

- If $\log(\widehat{\Lambda}(x; t_k)) \geq u_k$, do not reject $H_{01k} : S(x) \leq S_0(x)$, i.e., conclude the therapy ineffective.
- If $\log(\widehat{\Lambda}(x; t_k)) < u_k$, reject $H_{01k} : S(x) \leq S_0(x)$, i.e., conclude the therapy is effective enough to warrant farther study.

In adaptive designs, like the proposed design, the stopping probabilities at each stage are important. In fact, they used to calculate the average sample number (ASN) and the error probabilities of the trial (Chang (2007, p. 64)). Therefore, based on the fact that the joint distribution of $\log(\widehat{\Lambda}(x; t_i))$ ($i = 1, 2, \dots, k$) is a multivariate normal with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ determined by the hypothesized values ($S_0(x)$ and $S_A(x)$), there are two types of stopping probabilities: the efficacy stopping probability ($ESP(x; \boldsymbol{\mu})$, which is probability of stopping to claim efficacy, by not rejecting H_{02i} in each stage except the last ($i = 1, 2, \dots, k - 1$) or by rejecting H_{01k} in the last stage k , and the futility stopping probability ($FSP(x; \boldsymbol{\mu})$, which is the probability of stopping to claim inefficacy (not rejecting H_{01i} for $i = 1, 2, \dots, k$). Based on these probabilities we can calculate the error probabilities and the expected study length (ESL) of the total procedure. The futility error probability (FEP), which is the probability of erroneously reject the therapy as ineffective, is given by: $FEP(x; \boldsymbol{\mu}_A, \boldsymbol{\Sigma}_A) = Pr(G_1 \geq u_1) + \sum_{i=2}^k Pr(\ell_1 < W_1 < u_1, \dots, \ell_{i-1} < W_{i-1} < u_{i-1}, W_i \geq u_i)$. Where

$G_1 \sim N(\log(\Lambda_A(x)); \sigma_A^2(x; t_1)/(n_1 \cdot \Lambda_A(x)))$ and (W_1, W_2, \dots, W_i) with $i = 2, \dots, k$ is multivariate normal with mean $\boldsymbol{\mu}_A$ and covariance matrix $\boldsymbol{\Sigma}_A$. While the efficacy error probability (*EEP*), which is the probability of erroneously accept the therapy as effective, is given by: $EEP(x; \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) = P(G_1 \leq \ell_1) + \sum_{i=2}^{k-1} P(\ell_1 < W_1 < u_1, \dots, \ell_{i-1} < W_{i-1} < u_{i-1}, W_i \leq \ell_i) + P(\ell_1 < W_1 < u_1, \dots, \ell_{k-1} < W_{k-1} < u_{k-1}, W_k \leq u_k)$. Where $G_1 \sim N(\log(\Lambda_0(x)); \sigma_0^2(x; t_1)/(n_1 \cdot \Lambda_0(x)))$ and (W_1, W_2, \dots, W_i) with $i = 2, \dots, k$ is multivariate normal with mean $\boldsymbol{\mu}_0$ and covariance matrix $\boldsymbol{\Sigma}_0$. Thus, the expected study length (*ESL*) of the total procedure is given by: $ESL(x; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{i=1}^k T_i \cdot [FEP(x; \boldsymbol{\mu}, \boldsymbol{\Sigma}) + EEP(x; \boldsymbol{\mu}, \boldsymbol{\Sigma})]$, where $T_i = \sum_{j=1}^i t_j$ for $i = 1, \dots, k-1$ and $T_k = \sum_{j=1}^k t_j + x$.

4. OPTIMAL ADAPTIVE DESIGNS

Case and Morgan (2003) develop designs that minimize either the expected duration of accrual or the expected study length under H_0 . They choose optimal designs under the null hypothesis since they want to minimize the number of patients treated with an ineffective regimen. Following Case and Morgan (2003), we define optimal adaptive multi-stage phase II designs that satisfy the constraint for the errors and minimize the expected study length under the global H_{01} (denoted as $ESL(S_0)$). For each pair $S_0(x)$ and $S_A(x)$ with predetermined accrual period for each stage, the optimal designs that minimize the $ESL(S_0)$ are selected among all designs satisfying the constraints for the error probabilities and minimizing the total error levels by using the simulation annealing procedure (Fouskakis and Draper (2002)).

Table 1 provides optimal adaptive three-stage design parameters for various hypothesized values of one year survival rate. We examine these hypothesized values at the end of second, the third and the fourth year which is the end of the study, assuming that expected failure time follows a Weibull distribution with shape parameter $\theta = 1$ and rate parameter λ determined by the hypothesized values ($S_0(x)$ and $S_A(x)$), no lost to follow up and constant accrual rate 24 person per year. The decision cut-off points ℓ_i and u_i for each stage with fixed maximum duration of accrual ($MDA = 3$ -years) and the corresponding $ESL(S_0)$ are calculated, for futility and efficacy error probability at most equal to 0.10. In these initial results, as adaptive efficacy level $S_{A_i}(x)$ in intermediate stage i ($i = 2, \dots, k-1$), we use the average of futility $S_0(x)$ and efficacy $S_{A_{i-1}}(x)$ of the previous stage.

5. CONCLUDING REMARKS

In this paper an alternative adaptive hypothesis design, based on the sequential testing of hypotheses within each stage, is proposed for phase II trials. This procedure handle the problem of great uncertainty about the treatment efficacy that arised in early phases of clinical trials. In the proposed method starting with an optimistic targeted survival level in the first stage, we adjust the targeted

Table 1: Optimal adaptive designs that minimize ESL under the global H_{01}

$S_0(x)$	$S_A(x)$	ℓ_1	u_1	ℓ_2	u_2	u_3	FEP	EEP	ESL(μ_0, Σ_0)
0.10	0.25	0.52	0.61	0.53	0.55	0.54	0.099	0.099	2.08
0.20	0.35	0.16	0.55	0.17	0.54	0.18	0.094	0.093	3.00
0.30	0.45	-0.13	0.27	-0.12	0.26	-0.11	0.099	0.096	3.03
0.40	0.55	-0.42	0.02	-0.41	0.00	-0.40	0.099	0.094	3.09
0.50	0.65	-0.72	-0.25	-0.70	-0.26	-0.69	0.099	0.096	3.14
0.10	0.30	0.47	0.51	0.48	0.50	0.49	0.073	0.063	2.02
0.20	0.40	0.16	0.50	0.18	0.49	0.19	0.099	0.096	2.80
0.30	0.50	-0.13	0.23	-0.12	0.22	-0.11	0.094	0.096	2.88
0.40	0.60	-0.42	-0.05	-0.41	-0.06	-0.40	0.106	0.094	2.85 ¹
0.50	0.70	-0.72	-0.32	-0.71	-0.33	-0.70	0.108	0.096	2.87 ¹

¹FEP \leq 0.11

survival level in the intermediate stages by using the information of the previous stages. An important feature of the propose adaptive design is the ability to detect a moderate gain in the survival probability, when the true survival probability is lower than the initial expectation but still worthy.

In ongoing research we will work on the ability to adjust the sample size n_i in each stage (t_i respectively), in order to reduce the error levels and the total study duration.

ACKNOWLEDGEMENTS

This research has been co-financed by the European Union (European Social Fund - ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) - Research Funding Program: Heracleitus II. Investing in knowledge society through the European Social Fund.

ΠΕΡΙΛΗΨΗ

Οι προτεινόμενοι σχεδιασμοί πολλαπλών σταδίων Φάσης II (Lin et al., 1996, Case and Morgan, 2003, Jennison and Turnbull 2000) για την παρακολούθηση των πιθανοτήτων επιβίωσης, επικεντρώνονται στον έλεγχο της μηδενικής υπόθεσης $H_0 : S(x) = S_0(x)$ έναντι της εναλλακτικής $H_A : S(x) > S_0(x)$ σε επίπεδο σημαντικότητας α και ισχύ $1 - \beta$ η οποία υπολογίζεται σε επιλεγμένο επίπεδο επιβίωσης - στόχο $S_A(x)$. Όπου x είναι ο χρόνος επιβίωσης που μας ενδιαφέρει και $S(\cdot)$ η συνάρτηση επιβίωσης. Στα πρωταρχικά στάδια των κλινικών δοκιμών αντιμετωπίζουμε συχνά την αβεβαιότητα του κατά πόσον έ-

χουμε κάνει σωστή επιλογή του συγκεκριμένου επίπεδου επιβίωσης - στόχο $S_A(x)$. Εάν το αναμενόμενο επίπεδο επιβίωσης είναι πολύ αισιόδοξο θα απορριφθεί μια ελπιδοφόρα θεραπεία. Διαφορετικά, εάν το αναμενόμενο επίπεδο επιβίωσης είναι πολύ συντηρητικό, θα απαιτηθούν περισσότεροι ασθενείς από ότι είναι απαραίτητοι.

Για αυτό το σκοπό, προτείνουμε μια στρατηγική προσαρμοσίμων σχεδιασμών πολλαπλών σταδίων για την εκτίμηση της πιθανότητας επιβίωσης, κατά την οποία χρησιμοποιώντας την πληροφορία των προηγούμενων σταδίων, προσαρμόζουμε το επίπεδο επιβίωσης-στόχο $S_A(x)$ σε κάθε στάδιο, ενώ το σφάλμα τύπου I παραμένει σταθερό. Επίσης, η υπόθεση αυτή δημιουργεί σύγχυση μεταξύ κλινικών ιατρών, και στην πραγματικότητα και μεταξύ στατιστικών, εφόσον η απορρίψη της H_0 και ο προσδιορισμός της ισχύος στο σημείο $S(x) = S_A(x)$, δεν μπορεί να θεωρηθούν ισοδύναμα με το ότι η πιθανότητα επιβίωσης της συγκεκριμένης θεραπείας είναι μεγαλύτερη από $S_A(x)$. Για να αντιμετωπιστεί το πρόβλημα αυτό, στην προτεινόμενη κατηγορία σχεδιασμών ελέγχουμε δύο υποθέσεις ακολουθιακά σε κάθε στάδιο, μια για να ελέγξουμε το αν η θεραπεία έχει μη αποδεκτά χαμηλή θεραπευτική αξία και την άλλη για να ελέγξουμε το αν η θεραπεία είναι αρκετά αποτελεσματική ώστε να απαιτείται περαιτέρω μελέτη. Για το πρώτο στάδιο χρησιμοποιούμε ένα αισιόδοξο επίπεδο στόχο $S_A(x)$ και σε κάθε άλλο στάδιο εκτός του τελευταίου ελέγχουμε ένα λιγότερο αισιόδοξο επίπεδο στόχο ($S_i(x) < S_A(x)$). Στο τελικό στάδιο k ελέγχουμε μόνο την μια υπόθεση που αντιστοιχεί στο μη ενδιαφέρον επίπεδο $S_0(x)$.

Κατά το στάδιο του σχεδιασμού, εκτιμήσεις των πιθανοτήτων επιβίωσης μπορούν να ληφθούν κάνοντας την υπόθεση ότι ο αναμενόμενος χρόνος επιβίωσης ακολουθεί μια εκθετική κατανομή ή παραλλαγές της. Κάτω από συγκεκριμένα όρια για τα επίπεδα των σφαλμάτων και την υπόθεση σταθερού λόγου συσώρευσης, αναπτύσσουμε βέλτιστους σχεδιασμούς οι οποίοι ελαχιστοποιούν την αναμενόμενη διάρκεια της μελέτης.

REFERENCES

- Bie, O., Borgan, O., and Liestol, K. (1987). Confidence Intervals and confidence bands for the cumulative hazard rate function and their small sample properties. *Scandinavian Journal of Statistics* **14**(3), 221-233.
- Case, DL., Morgan, TM. (2003). Design of phase II cancer trials evaluating survival probabilities. *BMC Medical Research Methodology* 3:6,1-12.
- Chang, M. (2007). *Adaptive Design Theory and Implementation Using SAS and R*, New York: Chapman and Hall/CRC.
- Fleming, TR. (1982). One-Sample Multiple Testing Procedure for Phase II Clinical Trials. *Biometrics* **38**, 143-151.
- Fouskakis, D. and Draper, D. (2002). Stochastic optimization: a review. *International Statistical Review* **70**, 315-349.
- Jennison, C., and Turnbull, BW. (2000). Group sequential designs with applications to clinical trials. Boca Raton, Chapman & Hall/CRC, pp.259-277.
- Lan, KKG.,and DeMets, DL. (1983). Discrete sequential boundaries for clinical trials. *Biometrika*, **70**(3),659-663.
- Lin, DY., Shen, L., Ying, Z. and Breslow, NE. (1996). Group sequential designs for monitoring survival probabilities. *Biometrics*, **52**,1033-1042.
- Pouloupoulou, S., Karlis, D., Yiannoutsos, CT. and Dafni, U. (2009). Optimal multi-stage phase II design with sequential testing within each stage for evaluating response

rates and survival probabilities. *Proceedings of the 22th Conference of the Greek Statistical Institute* **22**, 323-332.

Simon, R. (1989). Optimal two-stage designs for phase II clinical trials. *Controlled Clinical Trials* **10**, 1-10.



TREATING THE REGRESSION MIS-SPECIFICATION PROBLEM THROUGH SPACE

ANALYZING EMPLOYMENT AND UNEMPLOYMENT EVIDENCE FROM GREECE

Pródromos-Ioánnis K. Prodromídís

Centre for Planning and Economic Research,
and Athens University of Economics and Business
pjprodr@kepe.gr

ABSTRACT

The article sets out to empirically explain the population shares of male and female employment, unemployment and non-participation in the labor force using the 2001 Greek Census data collected at the (disaggregated) municipal level. The six functions are analyzed within a seemingly unrelated regressions framework. To improve the quality of the estimated coefficients, the available regressors (demographic, educational etc.) are supplemented with information extracted from (a) relevant functional linkages displayed among localities and (b) the spatial patterns of the error term, which, apparently, reflect omitted (unknown) regressors. A number of micro-regions and localities are identified on the basis of (a) and (b), above; and spatial dummies constructed from them. The resulting econometric fits are generally superior to those relying on conventional territorial dummies such as regions (provinces) and prefectures (counties) drawn on the basis of historical or other criteria. Additionally, the econometric analysis yields a number of non-spatial results. By isolating the effects of the spatial and non-spatial factors driving economic activity we obtain a better understanding of the internal heterogeneity of the country. This, in turn, facilitates the formulation of better targeted policy interventions towards geographic areas exhibiting symptoms of workforce involvement and social cohesion that policy-makers find distressing. At the same time, the discovery of non-spatial results means that other interventions could be conducted at the national level. To the extent aspects of the missing variables are incorporated in the regression, the common misspecification problem ought to be ameliorated.

Keywords: Disaggregated data. Spatial patterns. Omitted regressors, Sub-regional workforce involvement.

1. INTRODUCTION

The article puts forward the proposition that it is feasible to capture aspects of the missing determinants along the spatial dimension, and construct spatial regressors associated with meaningful (and, possibly, statistically significant) coefficients. The

point is fleshed out through an empirical analysis of male and female employment, unemployment and non-participation in the labor force in Greece at the municipal level using disaggregated data from the *2001 Census*. The goal is to develop a procedure that: (a) Assesses the impact of non-spatial factors. (b) Identifies localities and clusters of localities (micro-regions) exhibiting symptoms of workforce involvement and social cohesion which policy-makers find distressing, such as low employment and/or high unemployment and non-participation in the labor force or exclusion from the labor market.

To that end, the article employs a parametric model that relies on: (i) Conventional determinants, namely demographic and educational composition, sector concentration, population density. (ii) Functional (travel-to-work) linkages observed among localities, obtained via an algorithm. (iii) Information extracted from the spatial patterns of the estimated residuals which, in all likelihood, stem from the missing explanatory variables. Insofar as aspects of the missing variables are introduced in the parametric model, the conventional regression misspecification problem ought to be ameliorated. If that is so, the procedure developed hereinafter may be of interest to a broader readership than the one concerned with (a) the design of employment or development policies at the sub-national level and (b) the proposition or application of such policies in Greece, in particular.

Up until now, all scholarly discussions regarding employment and unemployment at the sub-national level in Greece have been conducted either at the fairly aggregated regional and prefectural levels or on the basis of a very broad (binary) urban/rural-area distinction. Hence, the use of a narrower spatial grid is quite novel. Obviously, the new approach has the advantage of relying on (a) fewer assumptions (as regards the questions whether the territorial units are sufficiently homogeneous or constitute unified labor markets), (b) the entire population (unlike the *Labor Force Survey*'s reliance on representative samples of the regional populations), and (c) quite a few observations (1,034 municipalities). Compared to the much smaller number of observations involved in studies using regional and prefectural data (13 and 54 observations, respectively), the multitude of municipal observations permits the consideration of a wider set of factors in order to explain the workforce involvement of local populations. Additionally, it brings to light spatial patterns (some sub-regional, other cross-regional) which might otherwise go undetected. It goes without saying that the more detailed our description and understanding of (i) these patterns and (ii) the disparities in workforce involvement and social cohesion are, the higher the probability to formulate tailor-made, place-based economic development policies aiming to raise employment and reduce unemployment.

At the international level, a number of studies have already employed disaggregated data in order to econometrically analyze unemployment (e.g., Pehkonen and Tervo, 1998; Burgess and Profit, 2001; the sources cited therein). In the present article, as mentioned, we look at the employed and non-participating components of the populations(s), as well. It allows for a more thorough examination.

2. THE NEED TO DEAL WITH MISSING VARIABLES

One of the issues that trouble empirical analysts attempting to explain a dependent variable, Y , in terms of certain independent variables X and Z , is that Z may not be known or the data may be missing. Hence, in the case of a simple linear function of the sort,

$$Y = \alpha_1 + \beta_1 X + \gamma_1 Z + u, \quad (1)$$

the analyst may have little choice but to estimate:

$$Y = a_2 + b_2 X + v. \quad (2)$$

As a consequence: (i) $E(a_2) \neq \alpha_1$. (ii) If X and Z are correlated then $E(b_2) \neq \beta_1$. (iii) $\text{Var}(b_2) > \text{Var}(\beta_1)$. Hence, if one relies on data obtained from the whole population (as in this article) then he or she we may underrate or overrate the impact of the various factors; and if one relies on a sample then the result of the t-statistic is likely to be misleading. In particular, in the case of uncorrelated X and Z , the t-statistic will be underestimated. Hence, it may not be easy to confirm if b is significantly different from zero, i.e., if X is likely to affect Y . (Pindyck and Rubinfeld, 1991; Gujarati, 1995.)

In the pages that follow we demonstrate that in cross-sectional analyses it is possible to extract information relating to some aspect of the missing variable. Indeed, the information may be picked up along the spatial dimension. The rationale is as follows: If the features of Z are channeled in v , then to some or considerable extent, the spatial patterns of the missing variable may be inferred from the residuals. In particular, if we estimate

$$\hat{Y} = a + bX, \quad (3)$$

and obtain

$$\hat{u} = Y - \hat{Y}, \quad (4)$$

we may be able to consider categorical variables that capture the non-random concentrations and dispersions of \hat{u} . By inserting these dummies (i.e., this particular information) in the regression, and making X orthogonal to the spatial dummies, the misspecification problem ought to be ameliorated when the model is estimated again.

The approach may hold a number of practical advantages while evading some well known snags in the econometric treatment of space when the landscape is very much fragmented (e.g., in Greece, Indonesia, the Philippines). For instance, it is unlikely we can apply techniques relying on contiguity or distance weight matrices (e.g., Molho, 1995; Anselin et al. 2000) on Greece due to the country's irregular, very idiosyncratic terrain: (a) The contiguity matrix cannot handle the country's many islands. (It may also render inaccurate associations in the cases of continental localities that border each other along mountain-crests but, unlike other places, are not linked by road.) (b) The distance matrix mixes land with sea distances or assumes neatly decaying features when, in fact, some islands are well linked with some (but not all) nearby

islands or other islands and continental ports, whereas other islands are not linked with neighboring islands. (It may also render inaccurate associations in the cases of localities that are situated in some distance from each other but are well linked via the road-and-rail network, i.e., *much closer* than they appear on the map.)

At the same time, reliance on alternative dummies standing for conventional territorial partitions (drawn after geographic, historical, gerrymandering, geometric or other factors) is hard to defend if disparities are larger within rather than between the spatial units. By contrast, the dummies recovered in the manner proposed above, may turn out to be very relevant to the matter under consideration. Indeed, even if the underlying reason (or reasons) for the spatial patterns of the residuals (and, hence, for the construction of the dummies) is (or are) unclear at first, a study of their extent and other features improves our chances of deducing what the source(s) of the phenomenon or disparity may be. Consequently, we may design better targeted policy interventions in order to treat it. (In the case of the issue at hand: better targeted policies that treat undesirable symptoms of employment, unemployment, and non-participation.)

3. THE APPLICATION

The intention is to explain the male and female population shares of employed, unemployed and non-participants in the workforce across municipalities in terms of: population density, p , and its square (to capture the impact of the rate of change), population composition (namely, gender and age, a , as well as formal qualifications, e), industry-and-profession concentration combinations, c (e.g., unskilled labor in agriculture, skilled in manufacture, clerks in public administration etc.), on which we have data; knowing that several of the regressors employed in the literature are missing. These are: the wage (though it is partly explained by the available factors), non-labor income, culture, domestic technologies (on the labor supply side), the cost of non-labor inputs, level and price of output, producer technologies, and the objectives of public sector employers (on the labor demand side).

Hence, we attempt to extract some aspects of the omitted information from the spatial patterns of the residuals and to incorporate these aspects in the regression. Prior to doing so we make use of a delineation of the country's labor market or travel-to-work areas (TTWAs) estimated at the 15% commuting threshold by Prodromidis (2010a) in order to assign to the observations dummy variables that capture functional (inter-municipal commuting) linkages among localities. Then the dummies are calibrated further so as to incorporate the information extracted from the spatial patterns of the residuals (i.e., the information that seems to stem from the missing explanatory variables). The procedure for amalgamating the TTWA dummies into these, more sophisticated spatial dummies is described in the following steps: (I) Localities (observations) are grouped together (i.e., are assigned the same spatial dummy) on the basis of the TTWA groupings. (II) The six dependent variables, l , are explained in terms of the spatial dummies, d , and the available regressors, c, p, a, e . The latter are made orthogonal towards the spatial dummies and to each other in order to avoid

correlations. (In essence, instead of regressing l on d, c, p, a, e , we first weed-out the industry-and-profession combination dummies, c , that are highly or modestly correlated with the spatial arguments used; then regress p on d and c , predict π and estimate an orthogonal $\rho = p - \pi$; regress a on d, c , and ρ , predict \hat{a} , and estimate $\tilde{a} = a - \hat{a}$); and regress e on d, c, ρ, \tilde{a} , predict \hat{e}_y , and estimate an orthogonal $\tilde{e}_y = e_y - \hat{e}_y$. Thus, we may explain l in terms of d, c, ρ, \tilde{a} , and \tilde{e}_y .) (III) Six residuals are estimated for each observation. (IV) Neighboring TTWAs and TTWAs situated on opposite coasts (separated/linked by water) are banded together if their residuals are similar (i.e., if the all six residuals of each observation are similar.) As a result, in several cases the initial dummies are replaced by new (fewer) dummies. (V) Step II is repeated with the new spatial dummies and the other available regressors. (VI) The spatial dummies associated with statistically significant spatial effects at the 1% level in one or more functions, are retained. Hence, several nearby localities are re-zoned (grouped) on the basis of their residuals. As a result the degrees of freedom increase compared to step I. (VII) Steps II-VI are repeated with the new (broader) spatial formations instead of the initial TTWAs (where applicable).

The process yields forty spatial formations (micro-regions and outliers) drawn from relevant functional linkages and the missing variables information. In the empirical micro-regional analysis that follow, the rest of the country serves as the reference area.

4. THE EMPIRICAL ANALYSIS

We proceed by considering six seemingly unrelated regressions explained in terms of the available conventional explanatory variables and the spatial dummies. Understandably, one of the regressions (in this case, the regression regarding male abstention from the workforce) is estimated as the residual expression of the other five.

To drive home the point regarding the suitability of conventional spatial dummies (made in the last paragraph of Section 2) and to gain insights we run two additional variant specifications: one relying on the country's 13 regions (*peripheries*), and one relying on the country's 54 prefectures (*nomes*). As the said partitions reflect ancient, medieval or irrelevant factors, they provide a blurred picture of modern-day micro-reality. By contrast, the model constructed through the procedure outlined in Section 3 exhibits higher R^2 's (Table 1). It also possesses fewer regressors than the prefectural specification and makes fewer assumptions regarding the homogeneity of the territorial units involved. Hence, it may be preferable on statistical grounds.

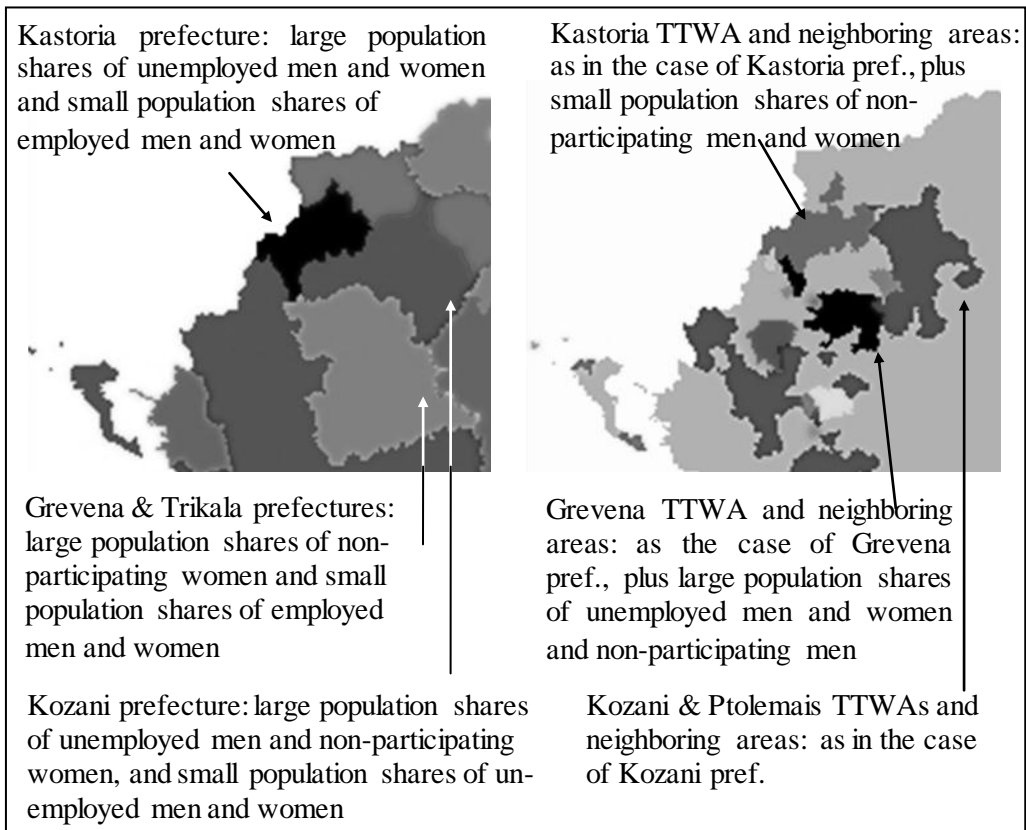
The estimated coefficients recovered from the three models via STATA are supplied in Prodromidis (2010b). To facilitate the reader, we project some of the spatial effects on a map using ArcGIS. According to the regional model, the communities in West Macedonia (pop. 295 thousand, situated in the centre of Figure 1) are inhabited by residents who exhibit small shares of employed men and women, and rather large population shares of unemployed and abstaining men and women compared to the reference population (the residents in the region of Attiki).

Table 1. The R²s associated with the three models

Seemingly unrelated equations explaining the share of

Models	employed females	employed males	unemployed females	unemployed males	non-participating females
Regional (R)	59,03%	68,76%	82,96%	30,37%	56,33%
Prefectural (P)	63,09%	70,39%	81,90%	42,06%	58,89%
Micro-regional (M)	69,42%	80,09%	85,60%	49,39%	67,25%

Figure 1. The spatial effects of the prefectural and micro-regional models projected on a map of north-western Greece



At the same time, the spatial coefficients recovered from the prefectural model suggest that only the residents in the communities situated in the prefecture of Kastoria (pop. 54 thousand) appear to exhibit rather small population shares of employed men and women and rather large shares of unemployed men and women compared to the reference population (the residents in the prefecture of Athens). The residents in the communities situated in the other prefectures of the region are

understood to exhibit somewhat different profiles. For instance, the residents in the communities situated in the prefecture of Florina (pop. 54 thousand) feature small population shares of unemployed women.

The spatial coefficients recovered from the micro-regional model vary even more: (a) In the communities situated on the plateaus of Kozani and Ptolemais (pop. 128 thousand) the population shares of employed men and women are rather small and the population shares of unemployed men and non-participating women rather large compared to the reference population. (b) In the communities situated on the plateau of Kastoria (pop. 51 thousand) the population shares of employed and non-participating men and women are small and the population shares of unemployed men and women rather large. (c) In the municipality of Siatista (pop. 6-7 thousand) the population share of unemployed men is even larger but the shares of unemployed and non-participating women and employed men and women are similar to those observed in the most of the country. (d) In other micro-regions of West Macedonia the profiles deviate in some other respect.

If the disparities that show up in the last model are masked in the regional and prefectural models, and if analysts go on relying on conventional territorial divisions, then they are likely to obtain misleading results regarding sub-regional heterogeneity. And if policy-makers act on advice derived from such results, then they are likely to make suboptimal use of the scarce resources earmarked for economic development and social cohesion interventions. For instance, given the smaller shares of abstaining people in area (b) compared to (a), the assumption of identical steps to reduce unemployment in both (a) and (b) is less likely to draw former or discouraged (job-seeking) residents into the local workforces in area (b) than in (a). *Ceteris paribus*, the policy will appear more successful or better managed in (b) than in (a), when in fact the two places need a different policy mix.

In view of our finding that micro-regions cross regional and prefectorial lines, it might make a lot of sense if we took a fresh look at the use of disaggregated data for the purpose of sub-national economic policy. For a very long time, national and EU territorial development policies have been devised, implemented, assessed on the basis of the regional and prefectorial framework.

Shifting our attention to the non-spatial effects, we come to Table 2 which supplies the signs of the estimated parameters and permits the inference of comments and recommendations regarding other, nation-wide policies. We list them as concisely as we can: The spreading out of industries relying on professions listed in the top rows appears to stimulate participation and reduce unemployment and abstention. At the same time, the spreading out of the two subsequent industry-and-profession combinations yields mixed results in male and female economic activity, but overall seems to bring about a rise in employment and reduction in unemployment. The modest urbanization of rural areas is likely to bring about a growth in local employment and a reduction in abstention among women. Demographic composition ought to be taken into account by policy-makers. If the aim is to boost employment e.

Table 2. The signs of the non-spatial coefficient supplied by the three models (R, P, M) Seemingly unrelated equations explaining the share (%) of

Independent variables (in %)	Employed			Unemployed			Non- particip.			
	Women		Men	Women		Men	Women			
	R	P	M	R	P	M	R	P	M	
<i>Municipal dummies featuring high concentrations of</i>										
1. Science & art professionals in trade, repairs	+	+	+	+	-	-	-	-	-	-
2. Science & art professionals in foreign organizat.	+	+	-	+	+	-	-	-	-	-
3. Plant/machine operators etc. in manufacture.	-	-	+	+	+	-	+	+	-	-
4. Skilled primary-sector workers in agriculture and related activities			+			-				-
<i>Population density (ranging from 0 to 36 thousand)</i>										
5. People per km ²	+	+	+	-	-	+	+	+	+	+
6. (People per km ²) squared	-	-	-	+	-	-	-	+	-	+
<i>Population composition (in %)</i>										
7. aged 0-4 years	+	+	+	+	+	+	-	-	-	+
8. aged 5-9 years	-	-	-	-	-	-	+	+	+	-
9. aged 10-19 years	+	+	+	-	-	-	-	-	-	+
10. women aged 20-64	+	+	+	-	-	-	+	+	+	+
11. men aged 20-64 (ref.)										
12. all aged 65-79 years	-	-	-	-	-	-	-	-	-	+
13. all aged 80+ years	+	+	+	-	-	-	+	+	+	+
<i>Formal qualifications (in %) women with</i>										
14. <9 years school	+	+	-	-	-	-	+	+	+	+
15. 9-12 years school	+	-	-	-	-	-	-	+	+	+
16. post-secondary school diploma (PSSD) or BA	+	+	+	+	+	-	-	-	+	+
<i>men with</i>										
17. <9 years school (ref.)										
18. 9-12 years school	-	-	-	-	-	-	-	-	-	+
19. PSSD or BA	-	-	-	-	-	-	+	+	+	-
<i>men and women with</i>										
20. postgraduate degree	-	-	-	+	-	-	+	+	+	-

Note: Regressors 1-4 take the value "1" when the location quotient (Isard, 1960) is equal to or greater than "3", and "0" otherwise. These regressors are not considered in the prefectural model due to their high or modest correlation with the spatial dummies. The highest *r* in the regional and micro-regional models is no more than 12,6%.

and reduce unemployment and non-participation: (a) Men with secondary school, post-secondary school, bachelor qualifications and women of all educational backgrounds will have to be persuaded of (if not enticed to) the benefits of participation. (ii) Women with secondary school qualification and, possibly, women with lesser qualifications, will have to upgrade their skills in order to be become more competitive in the market-place. (iii) The orientation procedure by which young men select their postgraduate subjects should improve.

Acknowledgement: I thank the conference participants for their helpful suggestions.

ΠΕΡΙΛΗΨΗ

Το άρθρο εξετάζει την κατανομή της απασχολήσεως, ανεργίας και μη-συμμετοχής στην αγορά εργασίας στην Ελλάδα σε επίπεδο δήμων-κοινοτήτων αξιοποιώντας τα απογραφικά στοιχεία του 2001. Ωστόσο, η απουσία στοιχείων για κάποιες ερμηνευτικές μεταβλητές ενδέχεται να καθιστά τις εκτιμήσεις των συντελεστών μεροληπτικές. Ως αντίδοτο επιδιώκεται η ενσωμάτωση στην παλινδρόμηση χωρικών ψευδομεταβλητών (ψ/μ) συνδεομένων με τις παραλειπόμενες μεταβλητές. Συγκεκριμένα, κατασκευάζονται ψ/μ για τις γειτνιάζουσες τοπικές αγορές εργασίας που εμφανίζουν τους αυτούς συνδυασμούς καταλοίπων. Έτσι (α) οργανώνονται οι οικονομικές ζώνες της χώρας σε ευρύτερες περιοχές που εκ των καταλοίπων φαίνεται να εμφανίζουν κοινά οικονομικά χαρακτηριστικά και (β) συνεκτιμάται ένα μέτρο των τοπικών ιδιοσυγκρασιών (δηλ., των ομοιοτήτων-διαφορών των τοπικών κοινωνιών) που αντανάκλα τις επιδράσεις των παραλειπομένων ερμηνευτικών μεταβλητών. Στον βαθμό που η ενσωμάτωση των χωρικών σχέσεων μέσω των συνηθισμένων τεχνικών δεν προσιδιάζει σε μια γεωγραφικά κατακερματισμένη χώρα όπως η Ελλάς, η εκτίμηση και συμπερίληψη χωρικών σχέσεων εκ των καταλοίπων ίσως συνιστά μια λύση. Επιπλέον, από την αντιπαραβολή των αποτελεσμάτων προς τα αποτελέσματα που βασίζονται στην περιφερειακή ή νομαρχιακή διαίρεση της χώρας διαπιστώνεται ότι το προτεινόμενο υπόδειγμα επιτρέπει την ανάληψη μιας καλύτερα στοχευμένης περιφερειακής πολιτικής. Ελπίζεται η παράθεση των ανωτέρω καθώς και των ευρημάτων και συμπερασμάτων που διατυπώνονται να συμβάλει στην καλύτερη κατανόηση και επίλυση του περιφερειακού προβλήματος της χώρας.

REFERENCES

- Anselin L., J. Cohen, D. Cook, W. Gorr and G. Tita. 2000. "Spatial Analyses of Crime." *Measurement and Analysis of Crime and Justice*. Edited by D. Duffee. Washington DC: National Institute of Justice, 213-262.
- Burgess S. and S. Profit. 2001. "Externalities in the matching of workers and firms in Britain." *Labour Economics*, 8: 313-33.
- Gujarati D. 1995. *Basic Econometrics*. New York: McGraw-Hill.
- Isard W. 1960. *Methods of Regional Analysis*. Cambridge MA: MIT Press.
- Molho I. 1995. "Spatial autocorrelation in British unemployment." *Journal of Regional Science*, 35: 641-658.

- Pehkonen J. and H. Tervo. 1998. "Persistence and Turnover in Regional Unemployment Disparities." *Regional Studies*, 32: 445-58.
- Pindyck R.S. and D.L.Rubinfeld. 1991. *Econometric Models and Economic Forecasts*. New York: McGraw-Hill.
- Prodromidis P.I. 2010a. "Identifying Spatial Labor Markets in Greece from the 2001 Travel-to-Work Patterns." *South-Eastern Europe Journal of Economics*, 8: 111-128.
- 2010b. "Analysing Local Employment and Unemployment in Greece under Conventional Zoning Regimes and Partitions extracted from the Data." *European Spatial Research and Policy*, 17: 61-91.



Forecasting Total Electron Content over Cyprus

E. Tsolaki, H. Haralambous, H. Papadopoulos

Frederick University, Cyprus
{eng.te, eng.hh, com.ph}@fit.ac.cy

ABSTRACT

The ionosphere is a region of the earth's upper atmosphere that ranges in height above the surface of the earth from approximately 50 km to 1000 km. It can affect the operation of communication, navigation, positioning and surveillance systems due to ionization that affects (delay) the propagation of radio waves. A direct measure of this delay is represented by Total Electron Content (TEC), which when multiplied by a factor that is a function of the signal frequency, it yields an estimate of this delay. Therefore the capability to predict TEC is crucial in providing mitigation measures to account for this delay and in applying possible mitigation techniques to reduce its detrimental effect.

This paper presents the development and comparison of two models for forecasting Total Electron Content (TEC) over Cyprus. In particular, an autoregressive (AR) time series and a neural network model are developed on the same set of data and then used for making short term forecasts of this important parameter.

Keywords: Total Electron Content, Autoregressive Moving Average Models, Neural Networks.

1. INTRODUCTION

The ionosphere is defined as a region of the earth's upper atmosphere where sufficient ionisation can exist to affect the propagation of radio waves. It ranges in height above the surface of the earth from approximately 50 km to 1000 km. The influence of this region on radio waves is accredited to the presence of free electrons. The impact of the ionosphere on communication, navigation, positioning and surveillance systems is determined by variations in its electron density profile and total electron content along the signal propagation path (Barclay, 2002 & Kersley et al., 2004). As a result satellite systems for communication and navigation, surveillance and control that are based on trans-ionospheric propagation may be affected by complex variations in the ionospheric structure in space and time leading to degradation of the accuracy, reliability and availability of their service. Total Electron Content (TEC) is an important parameter in trans-ionospheric links since when multiplied by a factor which is a function of the signal frequency, it yields an estimate of the delay imposed on the signal by the ionosphere due to

its dispersive nature (Goodman, 1992).

In this work we make a first attempt at the development of a forecasting model for TEC above Cyprus. The TEC data are based on processed RINEX files from the EUREF Permanent GPS Network station (NICO) at Nicosia, Cyprus and are freely available through Scripps Orbit and Permanent Array Center (SOPAC). Specifically we develop and compare an autoregressive moving average (ARMA) and a neural network model. Previous work with the same data was focused on long-term prediction of TEC for the case in which previous values of the parameter are not known (Haralambous et al., 2010 and Konstantinidis et al., 2010). Although there are around 60000 available hourly TEC measurements that were recorded above Cyprus from 1998 to 2009, we base the models developed in this paper on the data of year 2008. This is because we are mainly interested in stationary models where effects discussed in section 2 do not affect the data. In particular, we have chosen the data of year 2008 as it has less missing values than previous years and complete recordings for the 12 months. In sections 3 and 4 we propose time series and neural network models respectively for the data. Finally, in section 5, using the proposed models we make one step ahead predictions for one week.

2. MEASUREMENTS AND CHARACTERISTICS OF TEC

Dual-frequency GPS data recorded by GPS receivers enable an estimation of ionospheric variability because of the frequency dependent delay imposed on the signal due to the ionosphere (see Wild et al., 1989). By processing code and phase measurements on two frequencies in the L-band ($L_1 = 1575.42 \text{ MHz}$, $L_2 = 1227.60 \text{ MHz}$) it is possible to extract an estimate of the Total Electron Content (TEC) measured in total electron content units, ($1 \text{ TECU} = 10^{16} \text{ el m}^{-2}$). This is the total amount of electrons along a particular line of sight between the receiver and a GPS satellite in a column of 1 m^2 cross-sectional area and represents a typical quantitative parameter of interest to GPS users. TEC is therefore the integral of the electron density profile from the ground to an infinite height (practically the height of the satellite).

The analysis used in the present work to estimate TEC from GPS transmissions was carried out by means of the procedure developed by Ciraolo, 1993 and 2000. It can be shown (see Davis, 1990), that the GPS ionospheric delay (t_{ion}) is expressed in terms of TEC as

$$t_{ion} = \frac{40.3}{c f^2} TEC \quad (1)$$

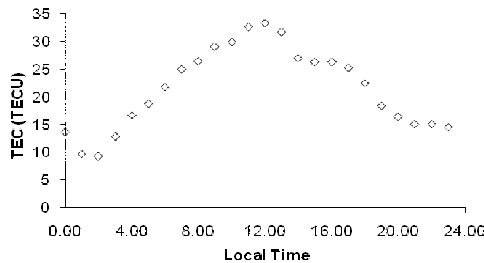
where

$$TEC = \int_{k_1}^{k_2} N(h) dh, \quad (2)$$

c is the velocity of light in free space, f is the frequency of radio wave propagating through the ionosphere (GPS radio signal) and $N(h)$ is the ionospheric electron density at the height h above the earth's surface.

The density of free electrons within the ionosphere and therefore TEC depend upon the strength of the solar ionizing radiation which is a function of time of day, season, geographical location and solar activity (Goodman, 1992 & Wild et al., 1989). Since solar activity has an impact on ionospheric dynamics which in turn influence the electron density of the ionosphere, TEC also exhibits variability on daily, seasonal and long-term time scales in response to the effect of solar radiation. It is also subject to abrupt variations due to enhancements of geomagnetic activity following extreme manifestations of solar activity disturbing the ionosphere from minutes to days on a local or global scale. The most profound solar effect on TEC is reflected on its daily variation as shown in figure 1. As it is clearly depicted, there is a strong dependency of TEC on local time which follows a sharp increase of TEC around sunrise and gradual decrease around sunset. This is attributed to the rapid increase in the production of electrons due to the photo-ionization process during the day and a more gradual decrease due to the recombination of ions and electrons during the night.

Figure 1: Diurnal variation of TEC graph



The long-term effect of solar activity on TEC follows an eleven-year cycle and is clearly shown in the upper figure 2 where noon values of TEC are plotted against time as well as against a modeled monthly mean sunspot number R which is a well established index of solar activity (see lower figure 2). We can observe a marked correlation of the mean level of TEC and modeled sunspot number. We note here that the gaps in the figure correspond to missing values.

There is also a seasonal component in the variability of TEC which can be attributed to the seasonal change in extreme ultraviolet (EUV) radiation from the sun. This can be clearly identified in figure 3 that plots noon values of TEC for high and low solar activity periods (years 2001 and 2008).

3. FITTING OF ARMA MODELS

As mentioned earlier we base our analysis on the data recorded in 2008. Solar activity in the particular year was low and so low variability in the data is ex-

Figure 2: Long-term TEC variation (upper figure) and solar activity variation (lower figure)

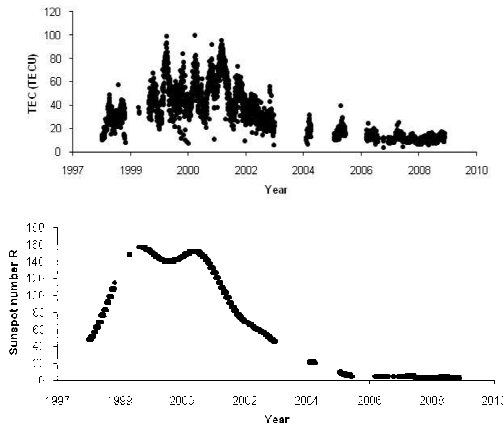
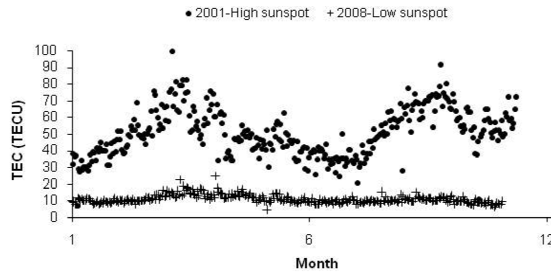


Figure 3: Seasonal variation of TEC at noon



pected. It is thus reasonable to consider stationary autoregressive moving average models ($ARMA(p, q)$) of the form

$$(1 + \alpha_1 B + \alpha_2 B^2 + \dots + \alpha_p B^p) X_t = (1 + \beta_1 B + \beta_2 B^2 + \dots + \beta_q B^q) e_t \quad (3)$$

where e_t is a purely random process and B is the backward shift operator defined by $B^k X_t = X_{t-k}$. Also to ensure stationarity of the above model we require that all roots of the polynomial $1 + \alpha_1 B + \alpha_2 B^2 + \dots + \alpha_p B^p$ lie outside the unit circle.

The first problem encountered in this data set was that of missing data. This was due to breakdowns of the machine and when this happened it lasted for several hours or even days. Therefore, to deal with this problem it seemed reasonable to remove all measurements of those particular days. Although missing data can affect the order and the predictability of the model we believe that in this case this was not an important problem as the amount of the data missing was very small. The total number of values analyzed was 7752. Before analyzing the data we first removed the sample mean.

As the particular data were recorded during a low solar activity period, only a 24 hour seasonality was expected to be present and this was removed before

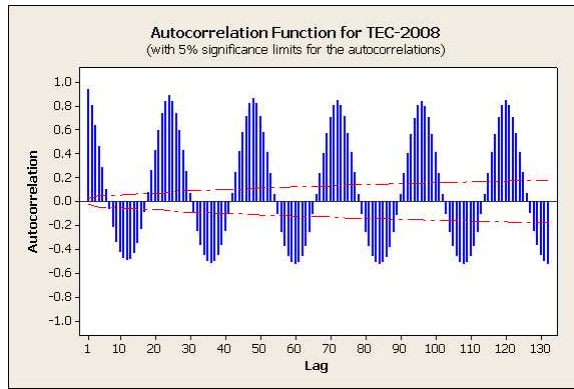
analyzing the data. The 24 hour seasonality can be seen in the estimated autocorrelation function shown in figure 4.

The autocovariance (R) and autocorrelation (r) functions are estimated by

$$\hat{R}(r) = \frac{1}{N} \sum_{r=1}^{N-|r|} X_t X_{t+|r|}, r = 0, \pm 1, \dots \quad (4)$$

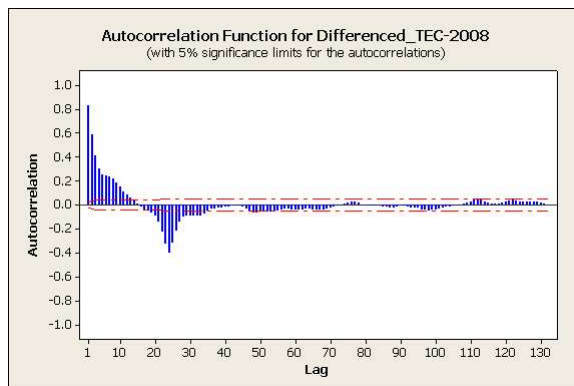
$$\hat{\rho}(r) = \frac{\hat{R}(r)}{\hat{R}(0)}. \quad (5)$$

Figure 4: Autocorrelation Function of TEC (2008)



It can easily be seen that there is a 24 hour seasonality that must be removed from the data before fitting the model. This is done by differencing the series at lag 24 and the autocorrelation function of the differenced data is estimated and shown in figure 5.

Figure 5: Autocorrelation function of differenced TEC (2008)



No other apparent seasonal trend is obvious in the data. This was tested with the KPSS (Kwiatkowski et al., 1992) test of stationarity which resulted in a p-value greater than 0.05. Thus, we proceed to fit an ARMA model to the data.

There are several criteria for order determination when an ARMA model is fit to data. A commonly used criterion is *Akaike's Information Criterion* (AIC) introduced by Akaike, 1969 and is the one used here to determine the order of the model to be fit. Also, in order to estimate the parameters of the model we follow the Hannan and Rissanen procedure, 1982. The optimum model found was an AR(30) model and the estimated parameters are shown in table 1.

Table 1: *Estimated coefficients of the AR(30) model*

Coef (Lag)	-1.1243 (1)	0.4250 (2)	-0.0913 (3)	0.0265 (4)	-0.0250 (5)	-0.0230 (6)
Coef (Lag)	-0.0230 (7)	-0.0072 (8)	-0.0019 (9)	0.0005 (10)	0.0040 (11)	0.0010 (12)
Coef (Lag)	-0.0105 (13)	-0.0056 (14)	0.0112 (15)	-0.0038 (16)	0.0016 (17)	0.0044 (18)
Coef (Lag)	0.0105 (19)	-0.0318 (20)	0.0319 (21)	-0.0017 (22)	-0.0252 (23)	0.4628 (24)
Coef (Lag)	-0.5114 (25)	0.1737 (26)	-0.0186 (27)	0.0039 (28)	-0.0097 (29)	-0.0180 (30)

In the next section we introduce neural networks. A neural network model is fit to the same data set and the two models are compared by predicting the data recorded the first week of 2009.

4. NEURAL NETWORKS

Neural networks (NNs) are based on the structure and function of biological nervous systems, such as the brain. A NN consists of a set of *units* (which represent neurons in biological nervous systems), also called *neurons* and a set of one-way *connections*, or *edges*, which transfer activity between units. Each one of these connections is associated with a weight, which determines the intensity of the signal transferred through it. The weights hold the processing ability of the network since learning is achieved by adjusting them in such a way that the output of the network gradually approaches the desired output for each training pattern presented.

The development of a network involves composing a suitable structure for the task at hand and choosing the learning algorithm that will be used for its training. The structure, or architecture, of a network defines the number of neurons it consists of and the way they are connected with each other. In this work we used a fully-connected feed-forward NN. The units in feed-forward networks are arranged in layers and connections can only exist between two adjacent layers; a unit in one layer cannot send a signal to any of the units in the same layer, or in any layer other than the next. In fully-connected NNs all units in each layer are connected to all units in the next layer. The choice of the number of hidden layers and units in each layer of a network is mainly performed experimentally.

Here a NN was trained to predict the next (differenced) TEC value using as inputs the previous k values. The recorded differenced values of 2008 were used for the NN development. This NN consisted of a single hidden layer with tangent sigmoid activations and a single output unit with a linear activation. It

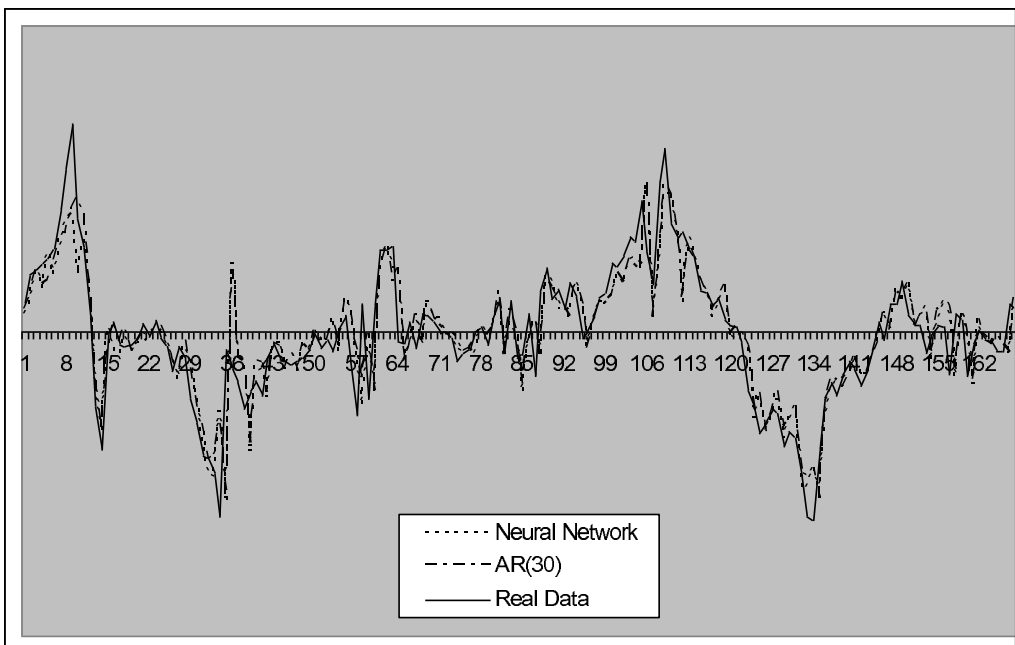
was trained with the Levenberg-Marquardt backpropagation algorithm with early stopping based on a validation set consisting of 20% of the available training data. The number of hidden units u of the network's hidden layer and the number k of previous TEC values used as its inputs were determined experimentally using the recorded values of 2008. This was done by evaluating different combinations of k and u following a 10-fold cross-validation process. Specifically, the available input-output patterns were split in 10 parts of almost equal size and the predictions of each part were made by training a neural network with the corresponding k and u combination on the remaining 9 parts. This allowed the evaluation of each combination of k and u values on all data available at the time. The best performance was obtained with $k = 28$ previous TEC values as inputs and $u = 10$ hidden units.

After determining the number of inputs and hidden units to use, a neural network was then trained on all available values (of 2008) and was evaluated on the values recorded in the first week of 2009.

5. FORECASTS

Figure 6 gives one step ahead predictions for one week (in hours) of the differenced TEC data for both the $AR(30)$ time series and the neural network models. The solid line represents the real values whereas the dashed lines represent the predicted values from the two models.

Figure 6: One step ahead predicted values for one week



As it can be seen both methods give very similar predictions. The normalized mean square error of the AR model predictions was 0.261 whereas the normalized mean square error of the neural network model predictions was 0.266 and thus both models exhibit similar performance. Both predictions are close to the true values and the pattern of the data is closely followed.

6. CONCLUSIONS

We have developed two forecasting models for TEC over Cyprus based on the data of year 2008. In particular an AR(30) time series and a neural network model for one step ahead prediction were developed and evaluated. Both models give promising results that encourage further study.

Future work to be carried out includes the statistical analysis of the entire 11 year dataset. Different periodicities as discussed in this paper must be taken into consideration and be removed from the data. Furthermore, as the variability of the data changes over the years due to solar activity, nonstationary time series models could be considered (see Priestley, 1981).

Περίληψη

Η ιονόσφαιρα είναι η περιοχή των άνω στρωμάτων της ατμόσφαιρας της γης που κυμαίνεται σε απόσταση από 50 χμ μέχρι 1000 χμ από την επιφάνειά της. Μπορεί να επηρεάσει την λειτουργία συστημάτων επικοινωνίας, πλοήγησης, προσδιορισμού θέσης και παρακολούθησης λόγω του φαινομένου του ιονισμού που επηρεάζει (καθυστερεί) την μετάδοση των ραδιοκυμάτων. Ένα άμεσο μέτρο της καθυστέρησης αυτή είναι το μέτρο της Ολικής Ηλεκτρονικής Περιεκτικότητας (TEC) το οποίο όταν πολλαπλασιαστεί με έναν όρο, που είναι συνάρτηση της συχνότητας του σήματος, δίνει μία εκτίμηση της καθυστέρησης. Έτσι η ικανότητα πρόβλεψης του μέτρου TEC είναι εξαιρετικά σημαντική για τη παροχή μέτρων μετρίασης αυτής της καθυστέρησης και την εφαρμογή τεχνικών για τη μείωση των αρνητικών της επιπτώσεων.

Η εργασία αυτή παρουσιάζει την ανάπτυξη και σύγκριση δύο μοντέλων για την πρόβλεψη του TEC πάνω από την Κύπρο. Συγκεκριμένα, αναπτύχθηκε ένα μοντέλο χρονοσειρών και ένα μοντέλο νευρωνικών δικτύων με βάση τα ίδια δεδομένα. Στη συνέχεια αυτά χρησιμοποιήθηκαν για να γίνουν βραχυπρόθεσμες προβλέψεις της σημαντικής αυτής παραμέτρου.

REFERENCES

- Akaike, H. (1969). Fitting autoregressive models for prediction. *Ann. Statist. Math.*, **21**, 243-247.
- Barclay, L. W. (2002). Ionospheric Effects and Communication Systems performance. *Keynote paper at the 10th Ionospheric Effects Symposium*, Washington DC.

- Ciraolo, L. (1993). Evaluation of GPS L2-L1 biases and related daily TEC profiles. *Proceedings of the GPS/Ionosphere Workshop*, Neustrelitz, 90-97.
- Ciraolo, L. (2000). Results and problems in GPS TEC evaluation. *Proceedings of the Radio Communications Research Units First GPS TEC Workshop*, Chilton, 47-60.
- Davis, K. (1990). *Ionospheric Radio*. Peter Peregrinus Ltd. London, UK.
- Goodman, J. (1992). *HF Communications, Science and Technology*, Nostrand Reinhold.
- Hannan, E. J. and Rissanen, J. (1982). Recursive estimation of mixed autoregressive -moving average order. *Biometrika*, **69**, 89-94. Correction (1983), **70**, 303.
- Haralambous, H., Vrionides, P., Economou L., Papadopoulos, H (2010). A local Total Electron Content Neural Network model over Cyprus. *Proceedings of the 4th International Symposium on Communications, Control and Signal Processing*, IEEE. DOI: 10.1109/ISCCSP.2010.5463306.
- Kersley, L., Malan, D., Pryse, S. E., Cander, L. R., Bamford, R. A., Belehaki, A., Leitingner, R., Radicella S. M., Mitchell C. N., Spencer P. S. J. (2004). Total electron content - A key parameter in propagation: measurement and use in ionospheric imaging. *Annals of Geophysics* **47**(2-3), 1067-1091.
- Konstantinidis, A., Haralambous, H., Agapitos, A., Papadopoulos, H., (2010). A GP-MOEA/D approach for modelling total electron content over Cyprus. *Engineering Intelligent Systems* **18**(3-4), 193-203. CRL Publishing.
- Kwiatkowski, D., Phillips, P. C. B., Schmidt, P. and Shin, Y. (1992). Testing the Null Hypothesis of Stationarity against the Alternative of a Unit Root. *Journal of Econometrics* **54**, 159-178.
- Priestley, M. B. (1981). *Spectral Analysis and Time Series*, 2 vols. Academic Press, London and New York.
- Wild, U., Beutler, G., Gurtner, W., Rothacher, M. (1989). Estimating the Ionosphere Using One or More Dual Frequency GPS Receivers, *Proceedings of the Fifth International Geodetic Symposium on Satellite Positioning, Las Cruces, New Mexico*, **2**, 724-736, March 13-17.

DIHEDRAL ANALYSIS OF SYMMETRY RANKINGS

M. Viana

University of Illinois at Chicago

viana@uic.edu

ABSTRACT

The methodology of dihedral analysis is illustrated in the study and interpretation of experimental data indexed by horizontal, vertical, and double reflections.

Keywords: Symmetry studies, symmetry orbits, orbit invariants.

1. INTRODUCTION

Figure 1 shows a set of 24 distinct flags used to illustrate an example of experimental results indexed by reflection symmetries.

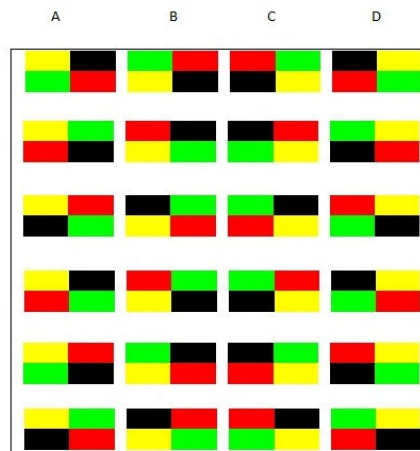


Figure 1: The 24 flags used in the flag ranking experiment

Fifteen subjects were asked to rank the flags in each row according to their preferences, so that each subject produced six complete rankings of four flags, one ranking for each row, similar to the rankings shown in (1), produced by three

subjects in the study.

1st	2nd	3rd	4th	1st	2nd	3rd	4th	1st	2nd	3rd	4th
A	D	B	C	D	A	C	B	A	D	B	C
A	D	B	C	C	B	D	A	D	A	B	C
C	B	D	A	B	C	D	A	C	A	D	B
C	B	A	D	C	D	B	A	C	A	B	D
D	C	A	B	C	D	B	A	A	B	D	C
A	B	C	D	B	C	A	D	A	D	B	C

(1)

2. SYMMETRY ORBITS

Each row in the set of flags shown in Figure 1 is a distinct symmetry orbit. To see this, note that each flag in column B can be obtained from the flag in column A by horizontal reflection, indicated here by h . Similarly, each flag in column D is the image of flag A by a vertical (v) reflection, whereas the flag in column C is the double reflection vh or hv of flag A. The double reflection, in this case, corresponds to a 180 deg rotation, or a point reflection symmetry (o), so that we write $vh = hv = o$. Flag A is, of course, its own unaltered image, the effect of the identity transformation on itself. On the other hand, no two flags in distinct orbits are image of each other by one of these symmetry transformations.

The resulting symmetries in the set $D_2 = \{1, v, h, o\}$, together with the operation of composition of bijective mappings, form a finite group and multiply according to the table shown in (2). That is, the product $\tau\sigma$ of two symmetries $\tau, \sigma \in D_2$ is a symmetry in D_2 ; the product is associative; $1\tau = \tau 1 = \tau$, so that 1 is the identity element and each $\tau \in D_2$ has a corresponding $\tau^{-1} \in D_2$ such that $\tau\tau^{-1} = \tau^{-1}\tau = 1$. Moreover, D_2 is a commutative group. That is, $\tau\sigma = \sigma\tau$ for all $\tau, \sigma \in D_2$.

D_2	1	v	h	o
1	1	v	h	o
v	v	1	o	h
h	h	o	1	v
o	o	h	v	1

(2)

The data shown in (3) are the joint frequency counts for the first-choice (rows) and second-choice flags (columns) in the experiment. For example, in 12 rankings flag A was the first choice and flag D was the second choice. We will write

$(A, D) = 12$ to indicate the frequency count associated with the (A, D) transition.

	A	B	C	D	total
A	0	2	3	12	17
B	4	0	14	4	22
C	6	11	0	10	27
D	16	1	7	0	24
total	26	14	24	26	90

(3)

We now observe that there were 14 (B, C) transitions and 11 (C, B) transitions, both of which are effected by a vertical reflection. Similarly, there were 12 (A, D) and 16 (D, A) transitions, both effected by a vertical reflection, so that together there were 53 transitions effected by a vertical reflection. The set of vertical transitions

$$\mathcal{O}^v = \{(A, D), (D, A), (B, C), (C, B)\},$$

is in itself an orbit of D_2 . The orbit can be generated by selecting any point $(X, Y) \in \mathcal{O}$ and evaluating $(\tau X, \tau Y)$ for all $\tau \in D_2$. Therefore, the corresponding frequency count data

$$x^v = \{12, 16, 14, 11\}$$

is indexed by (an orbit of) D_2 .

The same reasoning identifies the transitions (A, B) , (B, A) , (C, D) and (D, C) effected by horizontal reflections, and the transitions (A, C) , (C, A) , (B, D) , and (D, B) effected by double reflections. The resulting orbit and the frequency counts for the vertical transitions are given by

$$\mathcal{O}^h = \{(A, B), (B, A), (C, D), (D, C)\}, \quad x^h = \{2, 4, 10, 7\},$$

whereas the double reflection orbit and frequency counts are given by

$$\mathcal{O}^o = \{(A, C), (C, A), (B, D), (D, B)\}, \quad x^o = \{3, 6, 4, 1\}.$$

The partition of the data space into the symmetry orbits is summarized in the diagram shown in (4).

<table border="1" style="border-collapse: collapse; text-align: center;"> <thead> <tr> <th></th> <th>A</th> <th>B</th> <th>C</th> <th>D</th> </tr> </thead> <tbody> <tr> <th>A</th> <td>1</td> <td><i>h</i></td> <td><i>o</i></td> <td><i>v</i></td> </tr> <tr> <th>B</th> <td><i>h</i></td> <td>1</td> <td><i>v</i></td> <td><i>o</i></td> </tr> <tr> <th>C</th> <td><i>o</i></td> <td><i>v</i></td> <td>1</td> <td><i>h</i></td> </tr> <tr> <th>D</th> <td><i>v</i></td> <td><i>o</i></td> <td><i>h</i></td> <td>1</td> </tr> </tbody> </table>		A	B	C	D	A	1	<i>h</i>	<i>o</i>	<i>v</i>	B	<i>h</i>	1	<i>v</i>	<i>o</i>	C	<i>o</i>	<i>v</i>	1	<i>h</i>	D	<i>v</i>	<i>o</i>	<i>h</i>	1	<table border="1" style="border-collapse: collapse; text-align: center;"> <thead> <tr> <th></th> <th>A</th> <th>B</th> <th>C</th> <th>D</th> </tr> </thead> <tbody> <tr> <th>A</th> <td>0</td> <td>2</td> <td>3</td> <td>12</td> </tr> <tr> <th>B</th> <td>4</td> <td>0</td> <td>14</td> <td>4</td> </tr> <tr> <th>C</th> <td>6</td> <td>11</td> <td>0</td> <td>10</td> </tr> <tr> <th>D</th> <td>16</td> <td>1</td> <td>7</td> <td>0</td> </tr> </tbody> </table>		A	B	C	D	A	0	2	3	12	B	4	0	14	4	C	6	11	0	10	D	16	1	7	0
	A	B	C	D																																															
A	1	<i>h</i>	<i>o</i>	<i>v</i>																																															
B	<i>h</i>	1	<i>v</i>	<i>o</i>																																															
C	<i>o</i>	<i>v</i>	1	<i>h</i>																																															
D	<i>v</i>	<i>o</i>	<i>h</i>	1																																															
	A	B	C	D																																															
A	0	2	3	12																																															
B	4	0	14	4																																															
C	6	11	0	10																																															
D	16	1	7	0																																															

(4)

A more detailed summary of the transition orbits is summarized in Figure 2.

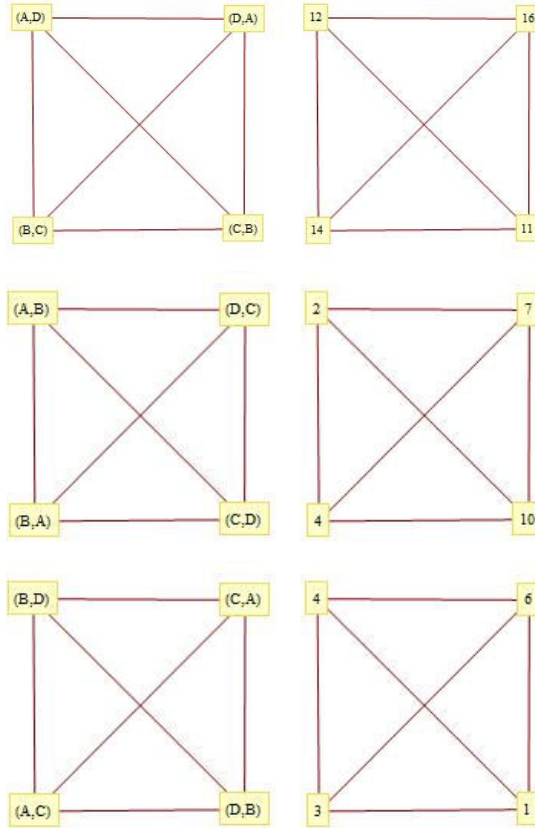


Figure 2: The orbits for the first-to-second flag preference transitions that were induced by a vertical reflection (top), a horizontal reflection (middle), and a double-reflection (bottom), and corresponding frequency counts. In each orbit, one moves between transitions according to the reflections defined by the diagrams, applied to each one of the component flags in the transition pairs.

3. THE ORBIT INVARIANTS AS FOURIER TRANSFORMS

Consider again the vertical transitions orbit and the corresponding frequency counts:

$$\mathcal{O}^v = \{(A, D), (D, A), (B, C), (C, B)\}, \quad x^v = \{12, 16, 14, 11\}.$$

Take (A, D) as its generator, arbitrarily. Because

- $(A, D) = (1A, 1D)$, then $x_1 = 12$;
- $(D, A) = (vA, vD)$, then $x_v = 16$;
- $(B, C) = (hA, hD)$, then $x_h = 14$;
- $(C, B) = (oA, oD)$, then $x_o = 11$.

We write the data in the form of the symbolic linear combination

$$x = x_1 1 + x_v v + x_o o + x_h h,$$

so that the choosing of a different orbit generator is effected by multiplying x by an element $\sigma \neq 1$ in D_2 . If $\sigma = h$, then

$$hx = x_1(h1) + x_v(hv) + x_o(ho) + x_h(hh) = x_1h + x_vo + x_ov + x_h1,$$

thus reassigning the data according to

$$x_1 \leftrightarrow x_h, \quad x_v \leftrightarrow x_o.$$

This is a D_2 -shuffle of the original data. The other shuffles are vx and ox .

We would like to summarize the orbital data in a way that does not depend on the D_2 -shuffles, that is, it does not depend on the arbitrariness of the orbit generator.

A set of invariant summaries is given by

$$\begin{aligned} x_1 + x_h + x_v + x_o, \\ x_1 + x_h - x_v - x_o, \\ x_1 - x_h + x_v - x_o, \\ x_1 - x_h - x_v + x_o, \end{aligned}$$

and will result as Fourier transforms of the data x over the group D_2 . Here D_2 acts on the data according to

$$\sigma x_\tau = x_{\sigma^{-1}\tau}, \tag{5}$$

so that, for example,

$$\tau(x_1 + x_h - x_v - x_o) = x_{\tau 1} + x_{\tau h} - x_{\tau v} - x_{\tau o} = \pm(x_1 + x_h - x_v - x_o), \tag{6}$$

for all $\tau \in \{1, h, v, o\}$, thus giving rise to a one-dimensional (invariant) subspace.

The Fourier transforms capture, in other words, precisely a set of (orbit) invariants of interest to the decomposition of the data space in (orbit) invariant subspaces. In the next section we present an overview of how the invariants will be determined.

4. THE ALGEBRAIC FRAMEWORK

Let x_τ indicate a scalar component of x , indexed by $\tau \in D_n$, and write

$$x = \sum x_\tau \tau \tag{7}$$

to denote the formal \mathbb{C} -linear combination of the dihedral elements with scalar coefficients the components of x . As it is immediately seen, the set

$$\mathbb{C}D_n = \left\{ \sum_{\tau \in G} x_\tau \tau : x_\tau \in \mathbb{C} \right\}$$

has a \mathbb{C} -vector space structure with a basis indexed by D_n , and is endowed with a multiplication

$$x y = \sum_{\tau, \sigma} x_{\tau} y_{\sigma} \tau \sigma = \sum_{\tau} \left(\sum_{\sigma} x_{\sigma} y_{\sigma^{-1} \tau} \right) \tau = \sum_{\tau} (x * y)_{\tau} \tau = x * y, \quad (8)$$

induced by the dihedral multiplication. The resulting structure is called a *dihedral algebra*. If the dihedral coefficients are points in a ring \mathcal{R} with identity, then the resulting structure is indicated by $\mathcal{R}D_n$ and called a *dihedral ring*.

The experimental data of interest for the Fourier analysis over a finite group G are then interpreted simply as points in the group algebra or group ring of G , whereas the Fourier transforms, as we shall see, are certain matrix realizations of the underlying group algebra operating on modules over that algebra.

Giving continuation to our example, associated with $D_2 = \{1, h, v, o\}$ there are four linear mappings $\xi : D_2 \mapsto \mathbb{C}$, given by

		1	h	v	o	
ξ^1		1	1	1	1	
ξ^2		1	1	-1	-1	
ξ^3		1	-1	1	-1	
ξ^4		1	-1	-1	1	

(9)

that satisfy the homomorphic property $\xi_{\tau} \xi_{\sigma} = \xi_{\tau \sigma}$ for all $\tau, \sigma \in D_2$ and lead to the linear combinations

$$\langle x, \xi \rangle = \sum_{\tau \in D_2} x_{\tau} \xi_{\tau}, \quad (10)$$

of ξ , with scalar coefficients given by the components of x , one for each

$$\xi \in \widehat{D}_2 = \{\xi^1, \xi^2, \xi^3, \xi^4\}.$$

The linear combinations $\langle x, \xi \rangle$ will be defined as the Fourier transforms of x evaluated at $\xi \in \widehat{D}_2$, so that the summaries introduced above follow directly from (10). They satisfy the (orbit) invariance property

$$\langle \tau x, \xi \rangle = \xi_{\tau} \langle x, \xi \rangle, \quad (11)$$

for all $\tau \in D_2$ and $\xi \in \widehat{D}_2$.

In the present case of D_2 , each transform gives rise to a *basis* for an invariant subspace in dimension of one, thus accounting for the dimension of the (data) vector space $\mathbb{C}D_2$. Together, these transforms are exactly two types:

1. An overall sum $\langle x, \xi^1 \rangle = x_1 + x_h + x_v + x_o$ of responses;
2. And three pairwise comparisons
 - $\langle x, \xi^2 \rangle = x_1 + x_h - x_v - x_o$,

- $\langle x, \xi^3 \rangle = x_1 - x_h + x_v - x_o$,
- $\langle x, \xi^4 \rangle = x_1 - x_h - x_v + x_o$.

If we write $H_v = \{1, v\}$, $H_h = \{1, h\}$, and $H_o = \{1, o\}$ to indicate these subgroups of D_2 then the pairwise comparisons are sums over the elements in the quotient groups

$$\begin{aligned} D_2/H_v &= \{\{1, v\}, \{h, o\}\}, \\ D_2/H_h &= \{\{1, h\}, \{v, o\}\}, \\ D_2/H_o &= \{\{1, o\}, \{v, h\}\}. \end{aligned}$$

Indicating by $(p, 1 - p)$ the probability distribution with support indexed by the elements in the quotient group, then any D_2 relabeling either leaves the distribution the same or changes it into $(1 - p, p)$. Therefore, its entropy (Ent) is an orbital invariant and can be used to statistically describe and compare the invariants when presented in the form of frequency counts. The overall sums, as they are constant, may be kept as components in the joint distribution for the three orbits.

5. INFERENCE

Here we studied the posterior distributions \mathcal{L} (relative to, say, a uniform prior) for their marginal distributions obtained from the joint total sums (53, 23, 14). The resulting Beta (*Be*) distributions

$$\mathcal{L}_v \sim Be(54, 38), \quad \mathcal{L}_h \sim Be(24, 68), \quad \mathcal{L}_o \sim Be(15, 77),$$

are shown in Figure 3.

Figure 3: Posterior densities

For the pairwise, within-orbit, invariants we may study their entropy. Within \mathcal{O}_v we have

$$\begin{aligned} \text{Ent}(12 + 16, 14 + 11) &= 0.693; \\ \text{Ent}(12 + 14, 16 + 11) &= 0.692; \\ \text{Ent}(12 + 11, 16 + 14) &= 0.684. \end{aligned}$$

Similarly, within \mathcal{O}_h , we obtained

$$\begin{aligned} \text{Ent}(2 + 7, 4 + 10) &= 0.598; \\ \text{Ent}(2 + 4, 7 + 10) &= 0.573; \\ \text{Ent}(2 + 10, 7 + 4) &= 0.692, \end{aligned}$$

whereas, within \mathcal{O}_o , we observed

$$\text{Ent}(4 + 6, 3 + 1) = 0.598;$$

$$\text{Ent}(4 + 3, 6 + 1) = 0.693;$$

$$\text{Ent}(4 + 1, 6 + 3) = 0.651.$$

We may summarize the within-orbit entropies by the minimum value, so that then we have,

$$\min \text{Ent } \mathcal{O}_v = 0.684, \quad \min \text{Ent } \mathcal{O}_h = 0.573, \quad \min \text{Ent } \mathcal{O}_o = 0.598.$$

6. COMMENTS

The human tendency of pairing asymmetrical images in a way that gives preference to stimuli that more closely approach perfect symmetry has been observed in more general contexts, e.g., [1]. The present experimental results suggest that hidden in the first-to-second rankings is the preference for matching the two flags in a way that one is a vertical reflection image of each other. This demonstration was repeated again with a second group of 13 subjects, thus producing a total of 78 rankings. In the second demonstration, the flags had the basic RGBY set of colors, and were squares in shape instead of rectangular. The results, summarized in Table (12), are strikingly similar to the those introduced in Table (3).

	<i>A</i>	<i>B</i>	<i>C</i>	<i>D</i>	total
<i>A</i>	0	4	2	14	20
<i>B</i>	2	0	15	2	19
<i>C</i>	3	17	0	2	22
<i>D</i>	14	2	1	0	17
total	19	23	18	18	78

(12)

The algebraic methods introduced in this illustration are part of the more general framework of dihedral Fourier analysis methods, described in [2]. See also [3], [4], and [5].

REFERENCES

- [1] Magnus Enquist and Rufus A. Johnstone, *Generalization and the evolution of symmetry preferences*, Proceedings: Biological Sciences **264** (1997), pp. 1345–1348.
- [2] Marlos Viana **M. Viana and H. Wynn (eds.)**, *Dihedral Fourier analysis*, Algebraic Methods in Statistics and Probability II - Contemporary Mathematics, vol. 516, American Math Soc, Providence, RI, 2010.
- [3] M. Viana and V. Lakshminarayanan, *Dihedral analysis of refraction profiles*, Proceedings 2010 Photonics Meeting (2010), Guwahati, India.
- [4] Marlos Viana, *Symmetry studies- an introduction to the analysis of structured data in applications*, Cambridge University Press, New York NY, 2008.
- [5] M. Viana and V. Lakshminarayanan, *Dihedral Fourier analysis*, Lecture Notes in Statistics, Springer, New York, NY, 2012, To appear.



ΕΥΡΕΤΗΡΙΟ ΣΥΓΓΡΑΦΕΩΝ (INDEX)

BASSIAKOS	366	ΔΑΣΚΑΛΑΚΗ	152
DAFNI	378	ΔΗΜΗΤΡΙΟΥ	104
HARALAMBOUS	397	ΔΟΝΑΤΟΣ	29
HOMBAS	373	ΔΡΙΤΣΑΚΗΣ	113
KARLIS	378	ΘΕΟΔΩΡΟΥ	125
PAPADOPOULOS	397	ΚΑΡΑΒΑΣΙΔΗΣ	134,143
POULOPOULOU	378	ΚΑΤΡΗΣ	152
PRODROMIDIS	387	ΚΑΤΣΑΚΙΩΡΗ	211
TSOLAKI	397	ΚΕΤΖΑΚΗ Δ.	160
VIANA	406	ΚΕΤΖΑΚΗ Ε.	160
ΑΔΑΜΙΔΗΣ	271	ΚΕΧΡΙΝΙΩΤΗΣ	125
ΑΛΕΒΙΖΟΣ	125, 321	ΚΙΤΙΚΙΔΟΥ	169
ΑΝΤΖΟΥΛΑΚΟΣ	303, 312	ΚΟΥΓΙΟΥΜΤΖΗΣ	254, 262, 329
ΑΝΤΩΝΙΟΥ	203	ΚΟΥΝΙΑΣ	280, 355
ΒΑΣΙΛΕΙΑΔΗΣ Β.	69	ΚΟΥΡΟΥΚΛΗΣ	178, 239
ΒΑΣΙΛΕΙΑΔΗΣ Γ.	62	ΚΟΥΤΡΑΣ	220, 338
ΒΟΤΣΗ	78	ΚΟΥΤΣΟΧΕΡΑΣ	185
ΓΕΩΡΓΙΑΔΗΣ	87	ΚΥΔΡΟΣ	143
ΓΕΩΡΓΙΟΥ	229	ΛΗΜΝΙΟΣ	78,87
ΓΚΙΟΥΛΟΣ	95	ΛΟΥΚΑΣ	271
ΔΑΛΑΜΗΤΡΑ	203	ΜΑΚΡΗ	195

ΜΑΚΡΗΣ	203, 346	ΠΟΥΠΑΚΗΣ	295
ΜΑΝΑΤΑΚΗΣ	211	ΡΑΚΙΤΖΗΣ	303, 312
ΜΑΤΗΣ	169	ΡΕΚΟΥΤΗ	321
ΜΗΛΙΕΝΟΣ	220	ΡΗΓΑΣ	69,134,143
ΜΠΑΝΤΗΣ	229	ΣΙΓΓΙΡΙΔΟΥ	329
ΜΠΟΜΠΟΤΑΣ	239	ΣΤΑΜΑΤΕΛΛΟΣ	95
ΜΠΟΡΑ-ΣΕΝΤΑ	329	ΣΠΑΝΟΥ	69
ΜΠΟΥΤΣΙΚΑΣ	312	ΤΡΙΑΝΤΑΦΥΛΛΟΥ	338
ΝΙΚΗΦΟΡΙΔΟΥ	247	ΤΣΑΚΛΙΔΗΣ	78, 95
ΠΑΓΓΕ	247	ΤΣΙΜΗΚΑΣ	229
ΠΑΠΑΔΗΜΗΤΡΙΟΥ	78	ΦΑΡΜΑΚΗΣ	160, 346
ΠΑΠΑΪΩΑΝΝΟΥ	185	ΦΙΛΙΠΠΟΥ	45
ΠΑΠΑΝΑ	254	ΧΑΛΙΚΙΑΣ	355
ΠΑΠΑΠΕΤΡΟΥ	262	ΧΑΤΖΗΠΑΝΤΕΛΗΣ	247
ΠΑΠΠΑΣ	271	ΧΡΙΣΤΟΦΙΔΗΣ	53
ΠΕΡΙΚΛΕΟΥΣ	280	ΨΥΛΛΑΚΗΣ	195
ΠΕΤΡΟΠΟΥΛΟΣ	288		