



Ελληνικό Στατιστικό Ινστιτούτο



Πανεπιστήμιο Πειραιά
Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης
ΠΜΣ στην Αναλογιστική Επιστήμη και Διοικητική Κινδύνου
ΠΜΣ στην Εφαρμοσμένη Στατιστική

26^ο Πανελλήνιο Συνέδριο Στατιστικής

8-11
ΜΑΪΟΥ 2013

Στατιστική στον Αναλογισμό,
τα Χρηματοοικονομικά
και τη Διοικητική Κινδύνου

<http://esi26.unipi.gr>



ΠΕΡΙΕΧΟΜΕΝΑ



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

Π Ρ Α Κ Τ Ι Κ Α

26^{ου} Πανελληνίου
Συνεδρίου Στατιστικής

PROCEEDINGS

of the 26th Panhellenic
Statistics Conference

*Στατιστική στον Αναλογισμό, τα
Χρηματοοικονομικά και τη Διοικητική
Κινδύνου*

*Statistics in Actuarial, Finance and Risk
Management*

Πειραιάς, 8-11 Μαΐου 2013



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

Π Ρ Α Κ Τ Ι Κ Α

26^{ου} Πανελληνίου
Συνεδρίου Στατιστικής

***Στατιστική στον Αναλογισμό, τα
Χρηματοοικονομικά και τη Διοικητική
Κινδύνου***

Οργάνωση

ΕΛΛΗΝΙΚΟ ΣΤΑΤΙΣΤΙΚΟ ΙΝΣΤΙΤΟΥΤΟ

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΑ

Τμήμα Στατιστικής κα Ασφαλιστικής Επιστήμης
ΠΜΣ στην Αναλογιστική Επιστήμη και Διοικητική Κινδύνου
ΠΜΣ στην Εφαρμοσμένη Στατιστική

Πειραιάς, 8-11 Μαΐου 2013



ΕΛΛΗΝΙΚΟ ΣΤΑΤΙΣΤΙΚΟ ΙΝΣΤΙΤΟΥΤΟ

Σολωμού 5 (Πλατεία Εξαρχείων)

Τηλ.& Fax 210 33.03.909

Email: esi-stat@hol.gr

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

ISBN:978-618-80672-3-3

ISSN:1792-2461

ΠΕΡΙΕΧΟΜΕΝΑ

ΠΡΟΛΟΓΟΣ	αρ.σελ. 7
ΧΟΡΗΓΟΙ.....	11
ΠΡΟΓΡΑΜΜΑ ΣΥΝΕΔΡΙΟΥ.....	12
ΕΠΙΤΡΟΠΕΣ ΣΥΝΕΔΡΙΟΥ.....	23

Εργασίες στα Ελληνικά	
ΒΑΣΙΛΕΙΑΔΗΣ Β., ΡΗΓΑΣ Α.: Μοντελοποίηση ενός στοχαστικού συστήματος με μία είσοδο και δύο εξόδους χρησιμοποιώντας τη συνάρτηση πιθανοφάνειας.....	25
ΔΗΜΗΤΡΑΚΟΣ Θ., ΚΥΡΙΑΚΙΔΗΣ Ε.: Ένα πρόβλημα βέλτιστης δρομολόγησης ενός οχήματος που διανέμει ένα προϊόν σε διατεταγμένους πελάτες με στοχαστικές διακριτές ζητήσεις και ποινές για τη μερική ικανοποίηση των ζητήσεων.....	35
ΔΟΝΑΤΟΥ Α., ΛΕΒΕΝΤΙΔΗΣ Ι.: Ποιοτικός έλεγχος και αποδοτικότητα των μετοχών των ελληνικών εταιρειών μεταποιητικού κλάδου.....	44
ΔΡΙΤΣΑΚΗΣ Ν.: Αιτιακές σχέσεις μεταξύ του ελλείμματος ισοζυγίου τρέχουσων συναλλαγών και των δημοσιονομικών ελλειμμάτων: Μία εμπειρική έρευνα για την Ελλάδα.....	55
ΕΥΘΥΜΙΟΥ Ι., ΓΕΩΡΓΙΟΥ ΣΤ.: Σχεδιασμοί λατινικών υπερκύβων	65
ΘΕΟΔΟΣΙΑΔΟΥ Ο., ΠΟΛΥΜΕΝΗΣ Β., ΤΣΑΚΛΙΔΗΣ Γ.: Διερεύνηση ενός στοχαστικού μοντέλου για το διαχωρισμό του θορύβου από την άφιξη πληροφορίας στην εξέλιξη της τιμής μιας μετοχής.....	76
ΚΑΛΑΜΑΡΑΣ Δ., ΚΑΛΑΜΑΤΙΑΝΟΥ Α.: Ένα μοντέλο διατάξιμης λογιστικής παλινδρόμησης για τη μελέτη παραγόντων συνδεόμενων με τη διάρκεια σπουδών: Μία εφαρμογή σε δεδομένα του Πάντειου Πανεπιστημίου	87
ΚΑΡΑΚΑΣΙΔΟΥ Σ.: Μη γραμμική ανάλυση χρονοσειρών τιμών αγροτικών προϊόντων: Εφαρμογή στην κεντρική αγορά Θεσσαλονίκης Α.Ε.	98
ΚΟΥΤΡΑΣ Μ.Β., BALAKRISHNAN N., ΜΗΛΙΕΝΟΣ Φ.Σ.: Ένα γενικό μοντέλο για τεστ εκκίνησης	108
ΚΟΥΤΡΟΥΒΕΛΗΣ Ι. Α.: Προσαρμογή προτύπων Poisson - γάμμα με τη μέθοδο μικτών ροπών	118
ΜΠΑΡΤΖΗΣ Γ., ΜΠΕΡΤΟΛΗΣ Ι., ΛΑΓΟΥ Β., ΣΤΑΜΑΤΟΠΟΥΛΟΥ Ε., ΣΕΛΗΝΙΔΟΥ Α., ΤΡΙΚΑΛΙΩΤΟΥ Α.: Ανάλυση οδοντιατρικών δεδομένων επιδημιολογικής έρευνας σε σχολεία του δήμου Πειραιά.....	128

ΜΠΟΥΤΣΙΚΑΣ Μ.Β.: Προσέγγιση μέσω της κατανομής γάμμα αντί της κανονικής κατανομής στο κεντρικό οριακό θεώρημα.....	138
ΜΩΥΣΙΑΔΗΣ Π.: Μέτρα κεντρικότητας σε δίκτυα.....	148
ΠΑΠΑΣΠΥΡΟΠΟΥΛΟΣ Κ.Γ., ΚΟΥΓΙΟΥΜΤΖΗΣ Δ.: Αιτιότητα κατά Granger σε απαριθμητές χρονοσειρές και εφαρμογή σε ιστορικά δεδομένα προστασίας της βιοποικιλότητας.....	158
ΠΕΡΙΚΛΕΟΥΣ Κ., ΚΟΥΚΟΥΒΙΝΟΣ Χ., ΚΟΥΝΙΑΣ Σ.: Βέλτιστοι πειραματικοί σχεδιασμοί με εξαρτημένες παρατηρήσεις	170
ΠΕΡΙΚΛΕΟΥΣ Κ., ΧΑΤΖΟΠΟΥΛΟΣ Σ., ΚΟΥΛΥΒΑ-ΜΑΧΑΙΡΑ Φ., ΚΟΥΝΙΑΣ Σ.: 3 ^k Βέλτιστοι σχεδιασμοί για την εκτίμηση γραμμικών και τετραγωνικών αντιθέσεων, όταν $N \equiv 0 \pmod{3}$	180
ΠΕΤΡΟΣ Σ., ΠΡΟΔΡΟΜΙΔΗΣ Π.: Παράγοντες που επηρεάζουν τη λιανική τιμή αμόλυβδης βενζίνης 95 οκτανίων: Απρ.2011- Δεκ.2012.....	190
ΡΑΚΙΤΖΗΣ Α.Χ., ΜΠΟΥΤΣΙΚΑΣ Μ.Β., ΑΝΤΖΟΥΛΑΚΟΣ Δ.Λ.: Κατανομές χρόνων διακοπής κα αθροισμάτων διακοπής σε πλειότιμες ανταλλάξιμες δοκιμές.....	201
ΣΑΧΛΑΣ Α., ΜΠΕΡΣΙΜΗΣ Σ., ΠΑΠΑΪΩΑΝΝΟΥ Τ.: Κανόνες διακοπής για πολυδιάστατες κλινικές δοκιμές φάσης II με δύο μεταβλητές απόκρισης.....	211
ΤΣΙΜΠΟΣ Κ.: Συμπληρωμένη γονιμότητα και πιθανότητες διεύρυνσης οικογένειας του πληθυσμού της Ελλάδας κατά επίπεδο εκπαίδευσης, αστικότητα και νομό.....	222
ΦΑΡΜΑΚΗΣ Ν.: Συμμετρικές διακριτές κατανομές και συντελεστής μεταβλητότητας.....	231
ΧΑΛΚΟΣ Γ., ΚΙΤΣΟΥ Δ.: Προσδιορισμός άριστου επιπέδου ρύπανσης και συνθήκες εξαγωγής περιοχής ωφέλειας.....	242
Εργασίες στα Αγγλικά	
CHARITIDOU E., FOUSKAKIS D., NTZOUFRAS I.: On Bayesian transformation selection: Problem formulation and preliminary results.....	253
DONATOS G.S., MICHAILIDIS G.C.: Small sample properties of least squares and ridge predictors under normal and nonnormal autocorrelated disturbances.....	261
DRAKOS K., KOUTRAS V.: A Markov chain approach for modelling USA bank's capitalization.....	271
KITSOS C.: Statistical invariance due to affine transformations.....	281
MAKRIDES A., KARAGRIGORIOU A.: Statistical inference for multi-state systems	291

OIKONOMOMPOULOU G., KARLIS D.: Statistical modeling of social networks: An application to co-authorship data.....	302
ORFANOIANNAKI K., KARLIS D: Hidden Markov models in modelling time series of earthquakes.....	312
PETRALIAS A., PRODROMIDIS P.: Price discovery under crisis: Uncovering the determinant factors of prices using efficient bayesian model selection methods.....	322
POLIMENIS V., PAPANTONIS I.: On the joint estimation of information betas.....	336
POULOPOULOU S., KARLIS D., DAFNI U.: A new two-stage, sample size re-estimation design for classical phase II clinical trials.....	346
SIGGIRIDOU E., KUGIUMTZIS D.: Granger causality in multivariate systems using restricted vector autoregressive models.....	356

ΠΡΟΛΟΓΟΣ

Το 26^ο Πανελλήνιο Συνέδριο Στατιστικής διοργανώθηκε από το Ελληνικό Στατιστικό Ινστιτούτο (Ε.Σ.Ι.) στον Πειραιά την περίοδο 8-11 Μαΐου 2013 σε συνεργασία με το Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς. Το θέμα του Συνεδρίου ήταν *Στατιστική στον Αναλογισμό, τα Χρηματοοικονομικά και τη Διοικητική Κινδύνου*.

Έλαβαν μέρος 255 Σύεδροι, από τους οποίους 188 ήταν φοιτητές και 23 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 ο κ. Νικήτας Πιττής, Καθηγητής του Τμήματος Χρηματοοικονομικής και Τραπεζικής Διοικητικής του Πανεπιστημίου Πειραιώς και ο κ. Άγγελος Ντάσσιος, Department of Statistics, London School of Economics.

Το Επιστημονικό Πρόγραμμα ήταν ιδιαίτερα πλούσιο και περιελάμβανε εβδομήντα δύο (72) ανακοινώσεις και εννέα (9) posters (αναρτήσεις) σε θέματα Αναλογισμού-Διοικητικής Κινδύνου, Δημογραφίας, Πειραματικών Σχεδιασμών, Πιθανοτήτων, Θεωρητικής και Εφαρμοσμένης Στατιστικής, Στοχαστικών Μοντέλων, Πρόβλεψης- Χρονοσειρών, Βιοστατιστικής, κτλ.

Η εναρκτήρια τελετή του Συνεδρίου πραγματοποιήθηκε στο μεγάλο αμφιθέατρο του Πανεπιστημίου Πειραιώς. Την έναρξη κήρυξε Αντιπρύτανης Οικονομικού Προγραμματισμού του Πανεπιστημίου Πειραιώς Καθηγητής Νικόλαος Γεωργόπουλος και ο Αναπληρωτής Πρόεδρος του Τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς Αναπληρωτής Καθηγητής Δημήτριος Αντζουλάκος. Επίσης, στην εναρκτήρια τελετή απηύθυνε χαιρετισμούς ο Πρόεδρος του Ε.Σ.Ι. Καθηγητής Χαράλαμπος Χαραλαμπίδης. Η τελετή έναρξης ολοκληρώθηκε με την εναρκτήρια ομιλία του προσκεκλημένου ομιλητή, του κ. Νικήτα Πιττή με τίτλο «Στατιστική μοντελοποίησης κερδοσκοπικών τιμών: Μία ιστορική προσέγγιση με μεθοδολογικές προεκτάσεις». Το Σάββατο σε ολομέλεια δόθηκε και η ομιλία του κ. Άγγελου Ντάσσιου με τίτλο: «A dynamic contagion process with applications to finance & insurance».

Κατά την εναρκτήριο τελετή απονεμήθηκε και το Ελένιο Βραβείο Διδακτορικής Διατριβής στη Στατιστική για τη διετία 2010-2012 στον κ. Ιωάννη Παπασταθόπουλο, Department of Mathematics and Statistics, Lancaster University, ο οποίος την επόμενη μέρα παρουσίασε τη διατριβή του με τίτλο «Statistical Models for Pharmaceutical Extremes». Ο κ. Παπασταθόπουλος εκπόνησε τη διατριβή του στο Lancaster University με

επιβλέποντα καθηγητή τον Α. Tawn. Για το βραβείο υποβλήθηκαν και οι ακόλουθες πέντε (5) διατριβές:

Ονοματεπώνυμο:	Vasiliki Dimitrakopoulou
Τίτλος Διατριβής:	Bayesian variable selection in cluster analysis
Επιβλέπων Καθηγητής:	Jim Griffin
Πανεπιστήμιο:	University of Kent

Ονοματεπώνυμο:	Emmanouil A. Varouchakis
Τίτλος Διατριβής:	Geostatistical analysis and space-time models of aquifer levels: application to Mires hydrological basin in the Prefecture of Crete
Επιβλέπων Καθηγητής:	Dionisios T. Hristopoulos
Πανεπιστήμιο:	Technical University of Crete

Ονοματεπώνυμο:	Ελευθέριος Μαυρουδής
Τίτλος Διατριβής:	Έλεγχοι διεργασιών υψηλής ποιότητας και εφαρμογές στη δειγματοληψία αποδοχής
Επιβλέπων Καθηγητής:	Νικόλαος Φαρμάκης
Πανεπιστήμιο:	Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης

Ονοματεπώνυμο:	Anestis Touloumis
Τίτλος Διατριβής:	Generalized estimating equations for multinomial responses
Επιβλέπων Καθηγητής:	Alan Agresti
Πανεπιστήμιο:	University of Florida

Ονοματεπώνυμο:	Xanthi Pedeli
Τίτλος Διατριβής:	Modelling multivariate time series for count data
Επιβλέπων Καθηγητής:	Dimitrios Karlis
Πανεπιστήμιο:	Athens University of Economics and Business

Για έβδομη χρονιά απονεμήθηκε το Βραβείο Καλύτερη Εργασία Νέου Στατιστικού. Για το σκοπό αυτό υποβλήθηκαν τρεις (3) 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 2013 απονεμήθηκε στην **κ.Αικατερίνη Ορφανογιαννάκη**, υποψήφια διδάκτορα του *Οικονομικού Πανεπιστημίου Αθηνών* για την από κοινού με τον Αναπληρωτή Καθηγητή κ. Δημήτριο Καρλή εργασία της με τίτλο: «Hidden Markov models in modelling time series of earthquakes». Έπαινος απονεμήθηκε και στους άλλους δύο υποψήφιους διδάκτορες κ. **Γεωργία Οικονομοπούλου**, *Οικονομικό Πανεπιστήμιο Αθηνών*, για την από κοινού με τον Αναπληρωτή Καθηγητή κ. Δημήτριο Καρλή εργασία της με τίτλο «Statistical modeling of social networks: An application to co-authorship data» και κ. **Ιωάννη Παπαντώνη**, *Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης*, για την από κοινού με τον Επίκουρο Καθηγητή κ. Βασίλειο Πολυμένη εργασία του με τίτλο «On the joint estimation of information betas». Την Τριμελή Επιτροπή αποτελούσαν οι συνάδελφοι *Στρατής Κουνιάς*, *Ανδρέας Κυριακούσης* και *Χρόνης Μωυσιάδης*.

Οι κοινωνικές εκδηλώσεις ήταν πλούσιες: Cocktail προσφέρθηκε την πρώτη μέρα του Συνεδρίου στο χώρο του Πανεπιστημίου Πειραιώς. Εκδρομή για τους συνέδρους και τα συνοδεύοντα μέλη στην Αίγινα με επίσκεψη στον αρχαιολογικό χώρο του ναού της Αφαίας και στο μοναστήρι του Αγίου Νεκταρίου. Το βράδυ της Παρασκευής διοργανώθηκε το επίσημο δείπνο του συνεδρίου στο κέντρο «1911» στον Πειραιά, 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. Οι εργασίες πρέπει να έχουν ικανό στατιστικό περιεχόμενο, να αναδεικνύουν το πρόβλημα που μελετούν, να μην περιορίζονται μόνο σε Περιγραφική Στατιστική κτλ. 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 τριάντα εννέα (39) εργασίες, από τις οποίες δύο (2) εργασίες ανακλήθηκαν από τους συγγραφείς μετά την πρώτη

αξιολόγηση και τέσσερις (4) απορρίφθηκαν, 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: Χ. Αγιακλόγλου, Φ. Αλεβίζος, Μ. Βαμβακάρη, Ι. Βόντα, Χ. Ευαγγελάρας, Γ. Ηλιόπουλος, Δ. Ιωαννίδης, Α. Καλαματιανού, Α. Καραγρηγορίου, Δ. Καρλής, Σ. Κουνιάς, Σ. Κουρούκλης, Σ. Μαλεφάκη, Π. Μαραβελάκης, Λ. Μελικοτσίδου, Θ. Μοσχονά, Α. Μπατσίδης, Σ. Μπερσίμης, Χ. Μωυσιάδης, Σ. Ντεγιαννάκης, Ι. Ντζούφρας, Δ. Παναγιωτάκος, Α. Πανοπούλου, Σ. Παπαδόπουλος, Χ. Παυλόπουλος, Α. Πετραλιάς, Α. Σαχλάς, Γ. Τζαβελάς, Γ. Τσακλίδης, Ν. Τσάντας, Κ. Φερεντίνος, Δ. Φουσκάκης, Τ. Χριστοφίδης. 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 Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς, το ΠΜΣ στην Αναλογιστική Επιστήμη και Διοικητική Κινδύνου, το ΠΜΣ στην Εφαρμοσμένη Στατιστική, 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 Γραμματέα 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 26^{ου} Συνεδρίου

Γεώργιος Ηλιόπουλος
Δημήτριος Καρλής

Ιωάννης Κουτρουβέλης
Τάκης Παπαϊωάννου

ΧΟΡΗΓΟΙ ΣΥΝΕΔΡΙΟΥ



ΠΡΟΓΡΑΜΜΑ ΣΥΝΕΔΡΙΟΥ

Τετάρτη 8 Μαΐου 2013

Αίθουσα Συνεδρίων	
16:30	Έναρξη του Συνεδρίου – Χαιρετισμοί

Αίθουσα Συνεδρίων	<u>ΟΜΙΛΙΑ ΣΕ ΟΛΟΜΕΛΕΙΑ</u> Προεδρεύων: Τ. Παπαϊωάννου
17:00	Ν.Πιττής Στατιστική μοντελοποίηση κερδοσκοπικών τιμών: Μία ιστορική προσέγγιση με μεθοδολογικές προεκτάσεις

Αίθουσα Συνεδρίων	<u>ΑΝΑΛΟΓΙΣΜΟΣ - ΔΙΟΙΚΗΤΙΚΗ ΚΙΝΔΥΝΟΥ</u> Προεδρεύων: Ν.Πιττής
18:00	D.G. Konstantinides, C.E. Kountzakis Risk measure theory under heavy-tails and conic assumptions
18:25	A.G. Bardoutsos, D.G. Konstantinides Structure of subclasses in the frame of heavy-tailed distributions
18:50	Γ. Ψαρράκος Μελέτη της ολοκληρώσιμης κατανομής του ελλείμματος στο ανανεωτικό μοντέλο της θεωρίας κινδύνων
19:15	C.E. Kountzakis Deviation measures on Banach spaces and applications

Αίθ. 106	<u>ΔΗΜΟΓΡΑΦΙΑ</u> Προεδρεύων: Δ. Καρλής
18:00	Κ. Τσίμπος Συμπληρωμένη γονιμότητα και πιθανότητες διεύρυνσης οικογένειας του πληθυσμού της Ελλάδας κατά επίπεδο εκπαίδευσης, αστικότητα και νομό
18:25	Γ. Βερροπούλου Δημογραφικά, κοινωνικοοικονομικά χαρακτηριστικά και παράγοντες υγείας που σχετίζονται με θνησιμότητα στους ηλικιωμένους: Ανάλυση panel δεδομένων της έρευνας SHARE
18:50	Γ. Ματαλλιωτάκης, Χ. Σκιαδάς Στοχαστική μοντελοποίηση δεδομένων επιβίωσης και πρόβλεψη θνησιμότητας
19:15	C.H. Skiadas, C. Skiadas Inversion problems in the first exit time density functions

Αίθ. 107	<u>ΠΕΙΡΑΜΑΤΙΚΟΙ ΣΧΕΔΙΑΣΜΟΙ</u> Προεδρεύων: Χ. Μωυσιάδης
18:00	Κ. Περικλέους, Σ. Χατζόπουλος, Φ. Κολυβά-Μαχαίρα, Σ. Κουνιάς 3^k κλασματικοί παραγοντικοί σχεδιασμοί, βέλτιστη εκτίμηση γραμμικών και τετραγωνικών αντιθέσεων όταν $N = 0 \bmod 3$
18:25	Κ. Περικλέους, Χ. Κουκουβίνος, Σ. Κουνιάς Βέλτιστοι πειραματικοί σχεδιασμοί γραμμής με εξαρτημένες παρατηρήσεις
18:50	Ι. Ευθυμίου, Σ. Γεωργίου Σχεδιασμοί λατινικών υπερκύβων
19:15	Χ. Ευαγγελάρας Ορθογώνιοι σχηματισμοί τριών επιπέδων και εφαρμογές

20:00 WELCOME COCKTAIL

Πέμπτη 9 Μαΐου 2013

Αίθουσα Συνεδρίων	<u>ΕΛΕΝΕΙΟ ΒΡΑΒΕΙΟ ΔΙΔΑΚΤΟΡΙΚΗΣ ΔΙΑΤΡΙΒΗΣ</u> Προεδρεύων: Α. Καραγρηγορίου
9:00	Ι. Παπασταθόπουλος Statistical models for pharmaceutical extremes

Αίθουσα Συνεδρίων	<u>ΒΡΑΒΕΙΟ ΚΑΛΥΤΕΡΟΥ ΝΕΟΥ ΣΤΑΤΙΣΤΙΚΟΥ</u> Προεδρεύων: Σ. Κουνιάς
10:00	<u>K. Orfanogiannaki</u>, D. Karlis Hidden Markov models in modelling time series of earthquakes
10:25	V. Polimenis, <u>I. Papantonis</u> On the joint estimation of the information betas
10:50	<u>G. Oikonomopoulou</u>, D. Karlis Statistical modeling of social networks: An application to co-authorship data

Αίθ. 106	<u>ΜΑΘΗΜΑΤΙΚΗ ΣΤΑΤΙΣΤΙΚΗ</u> Προεδρεύων: Θ. Κάκουλλος
10:00	Π. Μπομποτάς, Γ. Ηλιόπουλος, Σ. Κουρούκλης Εκτιμητές διπλής προσαρμογής για τον λόγο παραμέτρων κλίμακος
10:25	A. Batsidis, N. Martin, L. Pardo, K. Zografos Εντοπισμός αλλαγής σημείου με φ -αποκλίσεις
10:50	Ι.Α. Κουτρουβέλης Προσαρμογή προτύπων Poisson-γάμμα με μέθοδο μικτών ροπών
11:15	C.P. Kitsos Statistical invariance due to affine transformations

Αίθ.107	<u>ΣΤΑΤΙΣΤΙΚΗ ΣΤΑ ΧΡΗΜΑΤΟΟΙΚΟΝΟΜΙΚΑ</u> Προεδρεύων: Γ. Δονάτος
10:00	Ο. Θεοδοσιάδου, Β. Πολυμένης, Γ. Τσακλίδης Διερεύνηση ενός στοχαστικού μοντέλου για το διαχωρισμό του θορύβου από την άφιξη πληροφορίας στη μεταβλητότητα της τιμής μιας μετοχής
10:25	S. Papadopoulos, G. Papadopoulos Testing capital needs, profit and credit of European banks under stress
10:50	S. Papadopoulos, G. Papadopoulos Causality and modeling among Greek CDS, bonds and stock
11:15	V. Polimenis, I. Papantonis Uncovering idiosyncratic skews

11:40 ΔΙΑΛΕΙΜΜΑ – ΚΑΦΕΣ

Αίθ.105	<u>ΧΡΟΝΟΣΕΙΡΕΣ</u> Προεδρεύων: Χ. Κίτσος
12:10	Σ. Καρακασίδου Μη γραμμική ανάλυση χρονοσειρών τιμών αγροτικών προϊόντων: Εφαρμογή στην Κεντρική Αγορά Θεσσαλονίκης Α.Ε.
12:35	Κ.Γ. Παπασπυρόπουλος, Δ. Κουγιουμτζής Αιτιότητα κατά Granger σε απεριθμητές χρονοσειρές και εφαρμογή σε ιστορικά δεδομένα προστασίας της βιοποικιλότητας
13:00	E. Siggiridou, D. Kugiumtzis Granger causality in multivariate systems using restricted vector autoregressive models
13:25	X. Pedeli On a multivariate integer-valued autoregressive model with Poisson innovations

Αίθ.106	<u>ΠΙΘΑΝΟΤΗΤΕΣ</u> Προεδρεύων: Δ. Αντζουλάκος
12:10	A. Κυριακούσης, Μ. Βαμβακάρη On asymptotic behaviour of Bailey-Daum distribution
12:35	A.X. Ρακιτζής, Μ.Β. Μπούτσικας, Δ.Α. Αντζουλάκος Κατανομές χρόνων διακοπής και αθροισμάτων διακοπής σε πλειότιμες ανταλλάξιμες δοκιμές
13:00	Γ. Τζαβελάς, Π. Οικονόμου Ιδιότητες χαρακτηρισμού για την λογαριθμοκανονική κατανομή με την βοήθεια σταθμισμένων κατανομών
13:25	V. Hombas Generalizing Galileo's passé-dix game
13:50	Θ. Κάκουλλος Η γένεση της πολυδιάστατης Cauchy

Αίθ.107	<u>ΒΙΟΣΤΑΤΙΣΤΙΚΗ</u> Προεδρεύων: Δ. Παναγιωτάκος
12:10	S. Pouloupoulou, U. Dafni A new two-stage, sample size re-estimation design for classical phase II clinical trials
12:35	A. Σαχλός, Σ. Μπερσίμης, Τ. Παπαϊωάννου Κανόνες διακοπής για πολυσταδιακές συγκριτικές κλινικές δοκιμές φάσης II με χρήση δύο μεταβλητών απόκρισης
13:00	Β.Γ. Βασιλειάδης, Α.Γ. Ρήγας Μοντελοποίηση ενός στοχαστικού συστήματος με μία είσοδο και δύο εξόδους χρησιμοποιώντας την συνάρτηση πιθανοφάνειας
13:25	Γ. Τσιμήκας, Α. Μπαντής Γενικευμένα γραμμικά μοντέλα με λογοκριμένη συμμεταβλητή σε δεδομένα επαναλαμβανόμενων μετρήσεων
13:50	Σ. Μαλεφάκη, Π. Οικονόμου, Χ. Καρώνη Παλινδρόμηση κατωφλιού με τυχαίες επιδράσεις για επαναλαμβανόμενα γεγονότα

14:15 ΜΕΣΗΜΒΡΙΝΗ ΔΙΑΚΟΠΗ

Αίθ.105	<u>ΜΠΕΥΖΙΑΝΗ ΣΤΑΤΙΣΤΙΚΗ</u> Προεδρεύων: Ι. Κουτρουβέλης
16:15	Σ. Μπουλοκώστα, Φ. Κολυβά Μαχαίρα, Χ. Αναγνωστοπούλου, Π. Μαχαίρας Μπεϋζιανή εκτίμηση παραμέτρων κατανομών ακραίων τιμών
16:40	A. Petralias, P. Prodromidis Price discovery under crisis: Uncovering the determinant factors of prices using efficient Bayesian model selection
17:05	E. Charitidou, D. Fouskakis, I. Ntzoufras On Bayesian transformation selection: Problem formulation and preliminary results
17:30	D. Fouskakis, I. Ntzoufras, D. Draper Power-expected-posterior priors variable selection in Gaussian linear models
17:55	K. Antonio, A. Bardoutsos, W. Ouburg Bayesian stochastic mortality models for two populations: A technical note on MCMC sampling

Αίθ.106	<u>ΣΤΑΤΙΣΤΙΚΗ</u> Προεδρεύων: Α. Κυριακούσης
16:15	Μ. Παπαπέτρου, Δ. Κουγιουμτζής Εκτίμηση τάξης μαρκοβιανής αλυσίδας και συσχέτισης μακράς εμβέλειας σε συμβολοσειρές DNA
16:40	Papastamoulis, M-L. Martin-Magniette, C-M. Rabusseau High dimensional mixtures of Poisson generalized linear models with application to RNA-seq data clustering
17:05	D. Karlis Multivariate count data with censoring
17:30	I. Vonta, A. Karagrigoriou Statistical inference for grouped data
17:55	Ν. Φαρμάκης Συμμετρικές διακριτές κατανομές και συντελεστής μεταβλητότητας

Αίθ. 107	<u>ΣΤΑΤΙΣΤΙΚΗ ΘΕΩΡΙΑ ΑΞΙΟΠΙΣΤΙΑΣ - ΕΠΙΧΕΙΡΗΣΙΑΚΗ</u> <u>ΕΡΕΥΝΑ</u> Προεδρεύων: Γ. Τσακλίδης
16:15	A. Makridis, A. Karagrigoriou Statistical inference for multi-state systems
16:40	Ι. Τριανταφύλλου Συστήματα συνεχόμενα k -από-τα- n : F και γενικεύσεις τους: Μία ανασκόπηση των πρόσφατων ερευνητικών αποτελεσμάτων
17:05	Θ. Δημητράκος, Ε. Κυριακίδης Ένα πρόβλημα βέλτιστης δρομολόγησης ενός οχήματος που διανέμει ένα προϊόν σε διατεταγμένους πελάτες με στοχαστικές ζητήσεις και ποινές για μερική ικανοποίηση των ζητήσεων
17:30	Ε. Κυριακίδης, Κ. Καραματσούκης Ένα στοχαστικό πρόβλημα δρομολόγησης ενός οχήματος για τη διανομή δύο παρόμοιων προϊόντων σε μία προκαθορισμένη ακολουθία πελατών
17:55	Ο. Καναβέτας, Α. Μπουρνέτας Προσαρμοστικές πολιτικές δειγματοληψίας υπό ελλιπή πληροφόρηση και πρόσθετους περιορισμούς

21:15 ΕΠΙΣΗΜΟ ΔΕΙΠΝΟ

Παρασκευή 10 Μαΐου 2013

10:30 ΕΚΔΡΟΜΗ

Σάββατο 11 Μαΐου 2013

Αίθ.106	<u>ΟΜΙΛΙΑ ΣΕ ΟΛΟΜΕΛΕΙΑ</u> Προεδρεύων: Δ. Κωνσταντινίδης
9:00	A. Dassios A dynamic contagion process with applications to Finance & Insurance

Αίθ.105	<u>ΣΤΑΤΙΣΤΙΚΗ ΣΤΗΝ ΟΙΚΟΝΟΜΙΑ</u> Προεδρεύων: Κ. Τσίμπος
10:00	Σ. Πέτρος, Π. Προδρομίδης Ανάλυση των παραγόντων που επηρεάζουν την τιμή της αμόλυβδης βενζίνης
10:25	N. Δριτσάκης Αιτιακές σχέσεις μεταξύ του ελλείμματος ισοζυγίου τρεχουσών συναλλαγών και των δημοσιονομικών ελλειμμάτων: Μία εμπειρική έρευνα για την Ελλάδα
10:50	Γ. Χάλκος, Δ. Κίτσου Προσδιορισμός άριστου επιπέδου ρύπανσης και οι συνθήκες εξαγωγής της περιοχής ωφέλειας
11:15	Γ. Δονάτος, Γ. Μιχαηλίδης Μελέτη Monte Carlo των προβλεπτριών ελαχίστων τετραγώνων και ridge για κανονικά και μη κανονικά αυτοσυσχετισμένα σφάλματα στην περίπτωση μικρών δειγμάτων

Αίθ.106	<u>ΘΕΩΡΙΑ ΠΙΘΑΝΟΤΗΤΩΝ</u> Προεδρεύων: A. Dassios
10:00	C.E. Kountzakis, S.D. Tsaros Generalized likelihood functions via random measures
10:25	Δ. Λυμπερόπουλος, Σ. Τζανίνης, Ν. Μαχαιράς Το πρόβλημα της αναγωγής μεμειγμένων ανανεωτικών διαδικασιών σε μεμειγμένες Poisson και συνέπειες μίας λύσης του
10:50	Δ. Λυμπερόπουλος, Ν. Μαχαιράς Χαρακτηρισμοί και μία κατασκευή των μεμειγμένων ανανεωτικών διαδικασιών
11:15	Μ. Μπούτσικας Προσέγγιση μέσω της κατανομής γάμμα αντί της κανονικής κατανομής στο Κεντρικό Οριακό Θεώρημα

Αίθ.107	<u>ΒΙΟΣΤΑΤΙΣΤΙΚΗ</u> Προεδρεύουσα: Χ. Καρώνη
10:00	Κ. Δασκαλοπούλου, Α. Κωνσταντόπουλος, Κ. Πουλάκης, Μ. Χουζούρης Πρόγνωση εξέλιξης του οξέως εμφράγματος του μυοκαρδίου μέσω των δεικτών CRP, TnI, AST
10:25	Γ. Μπάρτζης, Ι. Μπερτόλης, Β. Λαγού, Ε. Σταματοπούλου, Α. Σεληνίδου, Α. Τρικαλιώτου Ανάλυση οδοντιατρικών δεδομένων επιδημιολογικής έρευνας σε σχολεία του Δήμου Πειραιά
10:50	Σ. Μανωλιά Οι ιδιαίτερες δυσκολίες ασθενών με σπάνιες παθήσεις στην Ευρώπη και ο ρόλος της έγκαιρης διάγνωσης: Μελέτη περίπτωσης του συνδρόμου Prader-Willi (PWS)
11:15	Χ. Παρπούλα, Χ. Κουκουβίνος Επιλογή σχεδιασμού και τεχνικές εξόρυξης πληροφορίας στην στατιστική ανάλυση και μοντελοποίηση βάσεων δεδομένων

11:40 ΔΙΑΛΕΙΜΜΑ – ΚΑΦΕΣ

12:15-14:15 POSTERS (Αναρτημένες Εργασίες)*

1	Ε. Γεωργουσopoύλου, Δ. Παναγιωτάκος Σύγκριση στατιστικών μεθόδων (ημιπαραμετρικά μοντέλα αναλογικών κινδύνων, παραμετρικό εκθετικό μοντέλο και παραμετρικό μοντέλο Weibull) σε μοντέλα εκτίμησης κινδύνου εμφάνισης καρδιαγγειακού συμβάντος
2	Κ. Δράκος, Β. Κούτρας Μία μαρκοβιανή προσέγγιση για τη μοντελοποίηση της κεφαλαιοποίησης των αμερικανικών τραπεζών
3	Ι. Ευθυμίου, Σ. Γεωργίου, Σ. Στυλιανού Ορθογώνιοι σχεδιασμοί και σχεδιασμοί με μικρές συσχετίσεις κατάλληλοι για πειράματα με υπολογιστές
4	Δ.Α. Καλαμάρας, Α. Καλαματιανού Ένα μοντέλο διατάξιμης λογιστικής παλινδρόμησης για τη μελέτη παραγόντων συνδεόμενων με τη χρονική διάρκεια πανεπιστημιακών σπουδών
5	Δ. Λιόνης, Δ. Παναγιωτάκος

* Τα posters θα παραμείνουν αναρτημένα στον ημιώροφο του Κεντρικού Κτηρίου του Πανεπιστημίου Πειραιώς καθ' όλη τη διάρκεια του συνεδρίου.

	Σύγκριση λογιστικής παλινδρόμησης, διαχωριστικής ανάλυσης και δένδρων ταξινόμησης: Εφαρμογή σε δεδομένα μελέτης ασθενών-μαρτύρων και σε προσομοιωμένα δεδομένα
6	Γ. Ματαλλιωτάκης, Κ. Ζοπουνίδης Δημοσκόπηση για την οικονομία και την επιχειρηματικότητα
7	Θ. Μοσχονά, Μ. Χαλκιάς, Γ. Ντούρος Μελέτη των δεδομένων ΕΛΣΤΑΤ για την έρευνα εισοδήματος και τις συνθήκες διαβίωσης των ελληνικών νοικοκυριών
8	Κ.Γ. Παπασπυρόπουλος, Χ.Κ. Σώκος, Η. Γιαλαμπουκίδης, Π.Κ. Μπίρτσας Σύγκριση μοντέλων διαφορικών εξισώσεων για τον έλεγχο του συστήματος κάρπωσης θηραμάτων στην Ελλάδα
9	Ι. Τριανταφύλλου Διαγράμματα ελέγχου φάσης II με χρήση βαθμολογικών συναρτήσεων τύπου Wilcoxon: Γενικεύσεις και εφαρμογές

14:15 ΜΕΣΗΜΒΡΙΝΗ ΔΙΑΚΟΠΗ

Αίθ.106	ΠΟΙΟΤΙΚΟΣ ΕΛΕΓΧΟΣ Προεδρεύουν: Χ. Χαραλαμπίδης
16:30	Μ.Β. Κούτρας, Ν. Balakrishnan, Φ.Σ. Μηλιένος Ένα γενικό μοντέλο για τεστ εκκίνησης με δύο δυνατά αποτελέσματα
16:55	P.Castagliola, M.B.C. Khoo, P.E. Maravelakis Το EWMA διάγραμμα ελέγχου με κανόνες ροών
17:20	Σ. Μπερσίμης, P. Castagliola Μέθοδοι ελέγχου πολυμεταβλητών διεργασιών υψηλής απόδοσης
17:45	Α. Δονάτου, Ι. Λεβεντιδής Ποιοτικός έλεγχος και αποδοτικότητα των μετοχών των ελληνικών εταιρειών μεταποιητικού κλάδου

Αίθ.107	ΣΤΑΤΙΣΤΙΚΗ ΚΑΙ ΔΙΑΔΙΚΤΥΟ Προεδρεύουσα: Μ. Βαμβακάρη

16:30	Ι. Παπατσούμα, Ν. Φαρμάκης Προσωπικά δεδομένα και διαδίκτυο
16:55	Γ.Χ. Μακρής, Ε. Δαλαμήτρα Κρυπτογραφία και στεγανογραφία
17:20	Π. Μωυσιάδης Μέτρα κεντρικότητας σε δίκτυα

18:30 ΛΗΞΗ ΣΥΝΕΔΡΙΟΥ

ΚΟΙΝΩΝΙΚΕΣ ΕΚΔΗΛΩΣΕΙΣ & Πρόγραμμα Μετακινήσεων

Τετάρτη 8 Μαΐου 2013

Welcome Cocktail: Τμήμα Στατιστικής & Ασφαλιστικής Επιστήμης,
Πανεπιστήμιο Πειραιώς

Ωρα: 20:00

Πέμπτη 9 Μαΐου 2013

Επίσημο Δείπνο Συνεδρίου

Προορισμός: κέντρο «1911»

Ωρα: 21:00

ΕΚΔΡΟΜΗ

Παρασκευή 10 Μαΐου 2013

Πρόγραμμα: Αναχώρηση στις 10:30.

Περιλαμβάνεται στο δικαίωμα συμμετοχής. Η εκδρομή δεν περιλαμβάνει φαγητό.

Επιστημονική Επιτροπή

Μ. Γκλεζάκος, Πανεπιστήμιο
Πειραιώς

Γ. Δονάτος, Πανεπιστήμιο Αθηνών

Κ. Ζωγράφος, Πανεπιστήμιο
Ιωαννίνων

Δ. Καρλής, Οικονομικό
Πανεπιστήμιο Αθηνών

Χ. Κίτσος, Τ.Ε.Ι. Αθηνών

Μ. Κούτρας, Πανεπιστήμιο
Πειραιώς

Ι. Κουτρουβέλης, Πανεπιστήμιο
Πατρών

Δ. Κωνσταντινίδης, Πανεπιστήμιο
Αιγαίου

Χ. Μουσιάδης, Αριστοτέλειο
Πανεπιστήμιο Θεσσαλονίκης

Τ. Παπαϊωάννου, Πανεπιστήμια
Πειραιώς και Ιωαννίνων
(Πρόεδρος)

Α. Ρήγας, Δημοκρίτειο
Πανεπιστήμιο Θράκης

Ν. Φράγκος, Οικονομικό
Πανεπιστήμιο Αθηνών

Ε. Χατζηκωνσταντινίδης,
Πανεπιστήμιο Πειραιώς

Χ. Χαραλαμπίδης, Πανεπιστήμιο
Αθηνών

Τ. Χριστοφίδης, Πανεπιστήμιο
Κύπρου

Οργανωτική Επιτροπή

Μέλη ΔΕΠ του Τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης

Χ. Ευαγγελάρας

Γ. Ηλιόπουλος (Πρόεδρος)

Σ. Μπερσίμης

Γ. Πιτσέλης

Γ. Ψαρράκος

Προπτυχιακοί και μεταπτυχιακοί φοιτητές του Τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς

Ι. Μπαντούνας

Γ. Μπάρτζης

Γ. Σαμαρά

Μ. Ταφιάδη

Ι. Μπερτόλης

Ε. Σταματοπούλου

Υπεύθυνος Προγράμματος

Γ. Ηλιόπουλος, Πανεπιστήμιο Πειραιώς

εργασίες

στα ελληνικά

ΜΟΝΤΕΛΟΠΟΙΗΣΗ ΕΝΟΣ ΣΤΟΧΑΣΤΙΚΟΥ ΣΥΣΤΗΜΑΤΟΣ ΜΕ ΜΙΑ ΕΙΣΟΔΟ ΚΑΙ ΔΥΟ ΕΞΟΔΟΥΣ ΧΡΗΣΙΜΟΠΟΙΩΝΤΑΣ ΤΗ ΣΥΝΑΡΤΗΣΗ ΠΙΘΑΝΟΦΑΝΕΙΑΣ

Β.Γ. Βασιλειάδης, Α.Γ. Ρήγας

Δημοκρίτειο Πανεπιστήμιο Θράκης

{bvasil, rigas}@ee.duth.gr

ΠΕΡΙΛΗΨΗ

Στην εργασία αυτή επιχειρείται η μοντελοποίηση ενός στοχαστικού συστήματος στάσιμων σημειακών ανεξίτηλων με τη βοήθεια των γενικευμένων γραμμικών μοντέλων (ΓΓΜ). Η εκτίμηση των παραμέτρων του ΓΓΜ γίνεται στο πεδίο του χρόνου χρησιμοποιώντας τη συνάρτηση πιθανοφάνειας. Το στοχαστικό μοντέλο το οποίο προτείνεται για την περιγραφή του συστήματος περιέχει το κατώφλι, τη συνάρτηση άθροισης και τη συνάρτηση ανάκαμψης, η εκτίμηση των οποίων είναι σημαντική καθώς περιγράφουν την επίδραση της εισόδου στο σύστημα και την ύπαρξη ή μη αυτοπυροδότησης. 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 σε δύο περιπτώσεις. Όταν δεν υπάρχουν αλλαγές στο μήκος του γονικού μυός και όταν υπάρχουν.

Keywords: γενικευμένο γραμμικό μοντέλο, στάσιμη σημειακή ανέλιξη, συνάρτηση πιθανοφάνειας, μυϊκή ατράκτος

1. ΕΙΣΑΓΩΓΗ

Η εργασία αυτή έχει ως στόχο να παρουσιάσει μια μέθοδο μοντελοποίησης ενός πολύπλοκου στοχαστικού συστήματος στάσιμων σημειακών ανεξίτηλων με τη βοήθεια των γενικευμένων γραμμικών μοντέλων (ΓΓΜ). Η εκτίμηση των παραμέτρων του ΓΓΜ γίνεται στο πεδίο του χρόνου χρησιμοποιώντας τη συνάρτηση πιθανοφάνειας. Κίνητρο για την εργασία αυτή είναι η μελέτη της συμπεριφοράς της μυϊκής ατράκτου. Σε προηγούμενες εργασίες (Karavasilis et al., 2005, Kotti and Rigas, 2005) έχει μελετηθεί η επίδραση του γάμμα κινητονεύρου στην απόκριση του πρωτεύοντος αισθητήριου άξονα (Boyd, 1980). Στην εργασία αυτή μελετώνται τόσο η απόκριση του πρωτεύοντος όσο και του δευτερεύοντος άξονα (έξοδοι του συστήματος της μυϊκής ατράκτου) υπό την επίδραση ενός γάμμα κινητονεύρου (είσοδος του συστήματος) σε δύο περιπτώσεις.

Όταν δεν υπάρχουν αλλαγές στο μήκος του γονικού μύος και όταν υπάρχουν.

Στην Ενότητα 2 παρουσιάζουμε το λογιστικό μοντέλο, τη συνάρτηση πιθανοφάνειας και τη μέθοδο μεγιστοποίησης αυτής μέσω των εξισώσεων σκορ. Στην Ενότητα 3 επικεντρωνόμαστε στην μοντελοποίηση της λειτουργίας της μυϊκής ατράκτου. Παρουσιάζουμε λεπτομερώς τον τρόπο κατασκευής των πινάκων που θα χρησιμοποιηθούν για την υλοποίηση της μεθόδου και τέλος στην Ενότητα 4 παρουσιάζουμε τα σημαντικότερα αποτελέσματα που προέκυψαν από την εφαρμογή των προτεινόμενων μεθόδων στα δεδομένα μας.

2. ΤΟ ΛΟΓΙΣΤΙΚΟ ΜΟΝΤΕΛΟ

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 ένα διάνυσμα συμμεταβλητών και ψευδομεταβλητών που αντιστοιχούν σε παράγοντες κινδύνου. Τότε το μοντέλο λογιστικής παλινδρόμησης δίνεται ως εξής

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

όπου $\boldsymbol{\beta}$ το διάνυσμα των παραμέτρων (McCullagh and Nelder, 1989, Dobson and Barnett, 2008). Οι εκτιμητές μεγίστης πιθανοφάνειας των παραμέτρων β_j , $j = 1, \dots, k$, και κατά συνέπεια των πιθανοτήτων π_i , επιτυγχάνονται μεγιστοποιώντας το λογάριθμο της συνάρτησης πιθανοφάνειας $l(\boldsymbol{\beta}|\mathbf{y}) = \log L(\boldsymbol{\beta}|\mathbf{y})$,

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

Όταν χρησιμοποιείται η μέθοδος scoring, οι εκτιμητές των παραμέτρων β_j είναι οι λύσεις των εξισώσεων σκορ

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

Οι εκτιμητές μεγίστης πιθανοφάνειας βρίσκονται μόλις επιτευχθεί σύγκλιση μετά από διαδοχικές επαναλήψεις. Στην m -οστή επανάληψη, το διάνυσμα των εκτιμητών θα δίνεται από την εξίσωση

$$\boldsymbol{\beta}^{(m)} = \boldsymbol{\beta}^{(m-1)} + \mathbf{I}(\boldsymbol{\beta}^{(m-1)})^{-1} \mathbf{U}(\boldsymbol{\beta}^{(m-1)}), \quad (4)$$

όπου $\mathbf{I}(\boldsymbol{\beta})^{-1}$ είναι ο αντίστροφος του πίνακα πληροφορίας Fisher υπολογισμένου στις τιμές του $\boldsymbol{\beta}$ της προηγούμενης επανάληψης. Για τον υπολογισμό του επιπέδου σημαντικότητας καθώς και των διαστημάτων εμπιστοσύνης χρησιμοποιούνται στατιστικές συναρτήσεις τύπου Wald.

3. ΜΟΝΤΕΛΟΠΟΙΗΣΗ ΤΗΣ ΛΕΙΤΟΥΡΓΙΑΣ ΤΗΣ ΜΥΙΚΗΣ ΑΤΡΑΚΤΟΥ

Στην ενότητα αυτή, κατασκευάζεται ένα λογιστικό μοντέλο με σκοπό να περιγραφεί η συμπεριφορά της μυϊκής ατράκτου καταγράφοντας τις αποκρίσεις του πρωτεύοντος

(I_a) και του δευτερεύοντος άξονα (II) υπό την επίδραση ενός γάμμα κινητονεύρου (γ) σε δύο περιπτώσεις: Όταν δεν υπάρχουν αλλαγές στο μήκος του γονικού μυός ($L - \text{off}$) και όταν υπάρχουν ($L - \text{on}$). Οι δυο άξονες (I_a , II) πυροδοτούν όταν το δυναμικό της μεμβράνης ξεπεράσει μία κρίσιμη τιμή που ονομάζεται κατώφλι (threshold). Το δυναμικό της μεμβράνης επηρεάζεται τόσο από εξωτερικές όσο και από εσωτερικές διεργασίες. Το μοντέλο που χρησιμοποιείται για την περιγραφή της λειτουργίας της μυικής ατράκτου εμπλέκει τρεις συναρτήσεις: το κατώφλι, τη συνάρτηση ανάκαμψης (recovery) και τη συνάρτηση άθροισης (summation) (Karavasilis et al., 2005).

Έστω Y_t η διεργασία πυροδότησης ενός από τους δύο άξονες. Αν h είναι το διάστημα δειγματοληψίας, τότε οι παρατηρήσεις της εξόδου γράφονται ως:

$$y_t = \begin{cases} 1, & \text{όταν ένας παλμός εξόδου συμβαίνει στο } (t, t + h] \\ 0, & \text{αλλιώς,} \end{cases} \quad (5)$$

όπου $t = h, \dots, nh$ και $N = nh$ είναι η διάρκεια της παρατήρησης. Στην περίπτωση μας έχει επιλεγεί $h = 1 \text{ msec}$. Έστω X_t η διεργασία της εισόδου του συστήματος που προέρχεται από ένα γάμμα κινητονεύρο και αποτελείται από παρατηρήσεις x_t (διακριτές δυαδικές χρονοσειρές), όπου

$$x_t = \begin{cases} 1, & \text{όταν ένας παλμός εισόδου συμβαίνει στο } (t, t + h] \\ 0, & \text{αλλιώς.} \end{cases} \quad (6)$$

Με θ_t συμβολίζουμε το επίπεδο του κατωφλίου στη χρονική στιγμή t , που ορίζεται ως $\theta_t = \theta_t^* + \epsilon_t$, όπου ϵ_t είναι μια διεργασία θορύβου. Το ϵ_t ακολουθεί προσεγγιστικά κανονική κατανομή (Brillinger, 1992). Το θ_t^* είναι μια συνάρτηση του t , που παριστάνει τη μορφή του κατωφλίου. Στην εργασία αυτή θεωρούμε ότι $\theta_t^* = \beta_0$ δηλαδή ότι είναι μια άγνωστη σταθερά.

Η συνάρτηση που αντιπροσωπεύει τις εξωτερικές διεργασίες που επηρεάζουν το δυναμικό της μεμβράνης είναι η συνάρτηση άθροισης (SF). Η συνάρτηση αυτή δείχνει την επιρροή της εισόδου στο σύστημα και περιγράφεται από το σύνολο των παραμέτρων $\{\beta_{1,u}\}$. Έτσι, το δυναμικό της μεμβράνης τη χρονική στιγμή t , εξαιτίας των εξωτερικών ερεθισμάτων ορίζεται ως

$$SF_t = \sum_{u=1}^{u_M} \beta_{1,u} x_{t-u}, \quad (7)$$

όπου x_{t-u} είναι η παρατήρηση της γάμμα εισόδου τη χρονική στιγμή $t - u$.

Οι εσωτερικές διεργασίες είναι υπεύθυνες για πιθανές αποκρίσεις του συστήματος λόγω αυτοπυροδότησης. Μια συνάρτηση που περιγράφει τις εσωτερικές διεργασίες είναι η συνάρτηση ανάκαμψης. Στην παρούσα εργασία, χρησιμοποιούμε μια πολυωνυμική συνάρτηση βαθμού s , που ορίζεται ως

$$V_t = \begin{cases} \sum_{i=1}^s \beta_{2,i} (\tau_t - \zeta - 1)^i, & \text{αν } \tau_t \geq \zeta + 1 \\ 0, & \text{αν } \tau_t < \zeta + 1, \end{cases} \quad (8)$$

όπου $\{\beta_{2,i}\}$ είναι οι συντελεστές της συνάρτησης και ζ είναι το ελάχιστο διάστημα μεταξύ δύο διαδοχικών πυροδοτήσεων της εξόδου. Επιπροσθέτως, το τ_t παριστάνει το χρόνο που έχει περάσει από το τελευταίο γεγονός εξόδου.

Το μοντέλο λογιστικής παλινδρόμησης που περιγράφει τη συμπεριφορά της μυικής ατράκτου μπορεί να εκφραστεί στη μορφή

$$\log\left(\frac{\pi_t}{1 - \pi_t}\right) = SF_t + V_t - \beta_0, \quad (9)$$

όπου π_t είναι η πιθανότητα ενός γεγονότος εξόδου. Οι άγνωστες παράμετροι του μοντέλου είναι οι συντελεστές της συνάρτησης άθροισης $\{\beta_{1,u}, u = 1, \dots, u_M\}$, οι συντελεστές της συνάρτησης ανάκαμψης $\{\beta_{2,i}, i = 1, \dots, s\}$ και το κατώφλι, β_0 . Το πρόσημο μπροστά από τη σταθερά κατωφλίου μας δείχνει ότι το άθροισμα τόσο των εσωτερικών όσο και των εξωτερικών διεργασιών πρέπει να ξεπεράσει το επίπεδο του κατωφλίου ώστε να προκύψει πυροδότηση.

Στη μορφή πινάκων το διάνυσμα των παραμέτρων γράφεται ως

$$\boldsymbol{\beta} = (\beta_0 \ \beta_{1,1} \ \dots \ \beta_{1,u_M} \ \beta_{2,1} \ \dots \ \beta_{2,s})^T. \quad (10)$$

Η t -οστή γραμμή του πίνακα συμμεταβλητών $\mathbf{X}$ γράφεται ως

$$\mathbf{x}_t^T = (-1 \ x_{1,t-1} \ \dots \ x_{1,t-u_M} \ (\tau_t - \zeta - 1) \ \dots \ (\tau_t - \zeta - 1)^s) \quad (11)$$

αν $\tau_t \geq \zeta + 1$, ή

$$\mathbf{x}_t^T = (-1 \ x_{1,t-1} \ \dots \ x_{1,t-u_M} \ 0 \ \dots \ 0), \quad (12)$$

αν $\tau_t < \zeta + 1$.

Στον παρακάτω Πίνακα 1 παρουσιάζονται οι χρόνοι καταγραφής καθώς και ο αριθμός των παλμών για τις σειρές δεδομένων που χρησιμοποιήθηκαν.

	L-off	L-on
T	60000ms	60000ms
Αριθμός γ Γεγονότων	3844	3932
Αριθμός I_a Γεγονότων	1410	4352
Αριθμός II Γεγονότων	2128	2187

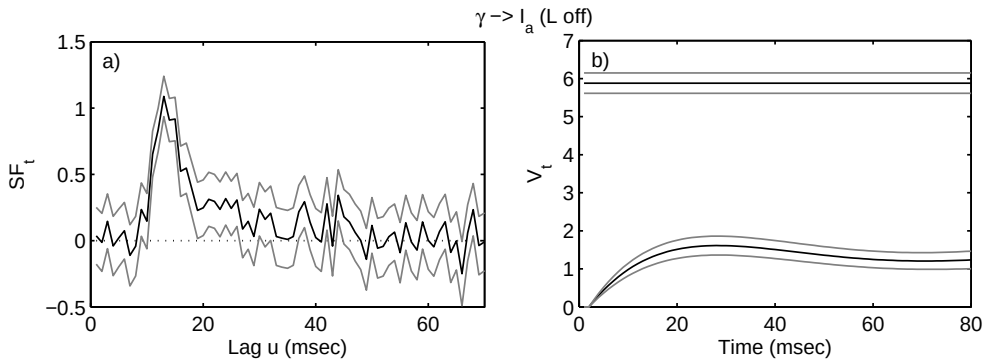
Θα πρέπει να σημειωθεί ότι σε καμία περίπτωση δεν προέκυψε το φαινόμενο του διαχωρισμού που έχει παρατηρηθεί σε προηγούμενη εργασία μελετώντας την επίδραση του άλφα κινητονεύρου (Βασιλειάδης και Ρήγας, 2010).

4. ΑΠΟΤΕΛΕΣΜΑΤΑ

Αρχικά, παρουσιάζονται τα αποτελέσματα για την περίπτωση της I_a εξόδου όταν δεν επηρεάζεται από το μήκος του μυός ($\gamma \rightarrow I_a(L - \text{off})$). Στην περίπτωση αυτή εκτιμήθηκαν 78 παράμετροι (70 για τη συνάρτηση άθροισης, 7 για τη συνάρτηση ανάκαμψης

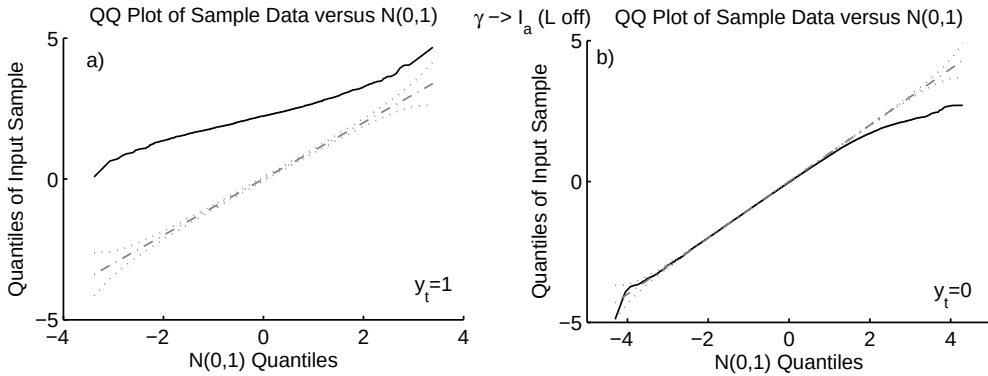
και 1 για το κατώφλι). Ο εκτιμητής του λογαρίθμου της πιθανοφάνειας παίρνει την τιμή -6307.1. Στο Σχήμα 1a παρουσιάζονται οι εκτιμητές των παραμέτρων της συνάρτησης άθροισης για την περίπτωση αυτή καθώς και το 95% διάστημα εμπιστοσύνης τύπου Wald. Παρατηρούμε ότι η μηδενική τιμή περιέχεται στο διάστημα εμπιστοσύνης για τα πρώτα 10 msec και οι συντελεστές της συνάρτησης άθροισης είναι θετικοί μεταξύ 10 και 30 msec και έξω από το διάστημα εμπιστοσύνης. Αυτό μας δείχνει ότι η συμπεριφορά του συστήματος αρχικά δε μεταβάλλεται, ενώ η απόκριση του I_a επιταχύνεται στο διάστημα από 10 έως 30 msec. Οι συντελεστές μετά από τα 40 msec δείχνουν ότι το σύστημα παραμένει σταθερό και δε μεταβάλλεται. Στο Σχήμα 1b παρουσιάζονται η συνάρτηση ανάκαμψης μαζί με τη συνάρτηση κατωφλίου και τα 95% διαστήματα εμπιστοσύνης τους. Παρατηρούμε ότι η συνάρτηση ανάκαμψης αυξάνεται μέχρι το 200 msec και μετά σταθεροποιείται κάτω από το κατώφλι. Αυτό μας δείχνει ότι, στην περίπτωση αυτή, ο I_a άξονας δεν αυτοπυροδοτεί. Για να μελετηθεί η εγκυρότητα του προσαρμοσμένου μοντέλου, χρησιμοποιήθηκαν τα τυχαιοποιημένα υπόλοιπα Dunn and Smyth, 1996 και χρησιμοποιήθηκε ο έλεγχος της εγκυρότητας με χρήση του γραφήματος Q-Q. Για λόγους οικονομίας δεν παρουσιάζουμε τον έλεγχο αυτό στην παρούσα εργασία, αφού περισσότερες πληροφορίες μπορούν να βρεθούν στις εργασίες Βασιλειάδης και συν.(2008) και Βασιλειάδης και Ρήγας (2010). Στο Σχήμα 2a παρουσιάζεται το γράφημα Q-Q των 1410 τυχαιοποιημένων υπολοίπων που αντιστοιχούν στις αποκρίσεις με $y_t = 1$. Όλα τα σημεία βρίσκονται έξω από τις 5% απορριπτικές περιοχές που υπολογίστηκαν με 1000 επαναλήψεις Monte Carlo. Η θέση των σημείων μάς υποδεικνύει ότι οι αποκρίσεις με $y_t = 1$ έχουν υποεκτιμηθεί. Στο Σχήμα 2b παρουσιάζεται το γράφημα Q-Q των 60000-1410=58590 τυχαιοποιημένων υπολοίπων που αντιστοιχούν στις αποκρίσεις με $y_t = 0$. Όλα τα σημεία βρίσκονται έξω από τις 5% απορριπτικές περιοχές που υπολογίστηκαν με 1000 επαναλήψεις Monte Carlo. Η θέση των σημείων μάς υποδεικνύει ότι οι αποκρίσεις με $y_t = 0$ έχουν υπερεκτιμηθεί. Τα αποτελέσματα

Σχήμα 1: α) Η συνάρτηση άθροισης και β) η συνάρτηση ανάκαμψης, για την έξοδο I_a , χωρίς την επίδραση του μήκους του μυός. Με γκρι χρώμα εμφανίζονται τα 95% διαστήματα εμπιστοσύνης.



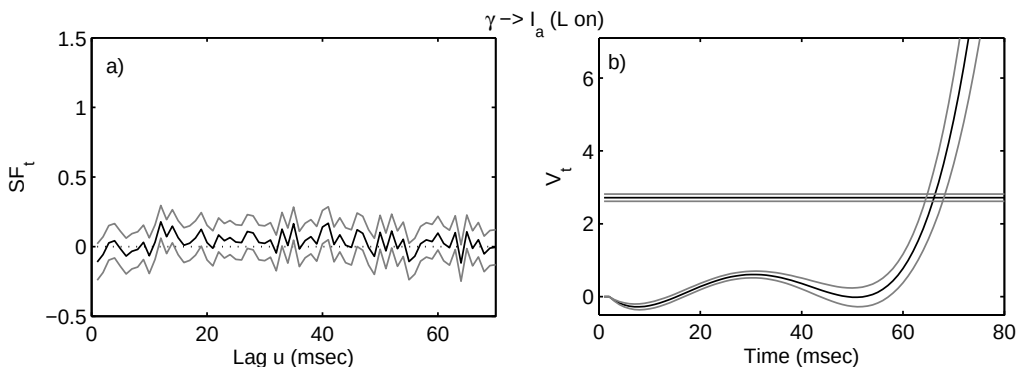
αντιπαραβάλλονται με αυτά της περίπτωσης της I_a εξόδου όταν επηρεάζεται από το μήκος του μυός ($\gamma \rightarrow I_a(L - on)$). Στην περίπτωση αυτή εκτιμήθηκαν 75 παράμετροι (70 για τη συνάρτηση άθροισης, 4 για τη συνάρτηση ανάκαμψης και 1 για το κατώφλι).

Σχήμα 2: Η γραφική παράσταση Q-Q των τυχαιοποιημένων υπολοίπων του προσαρμοσμένου λογιστικού μοντέλου που αντιστοιχούν α) στις παρατηρήσεις με $y_t = 1$ και β) στις παρατηρήσεις με $y_t = 0$, για την περίπτωση της εξόδου I_a χωρίς την επίδραση του μήκους του μύδς. Οι 5% απορριπτικές περιοχές μετά από 1000 επαναλήψεις Monte Carlo εμφανίζονται με γκρι χρώμα.



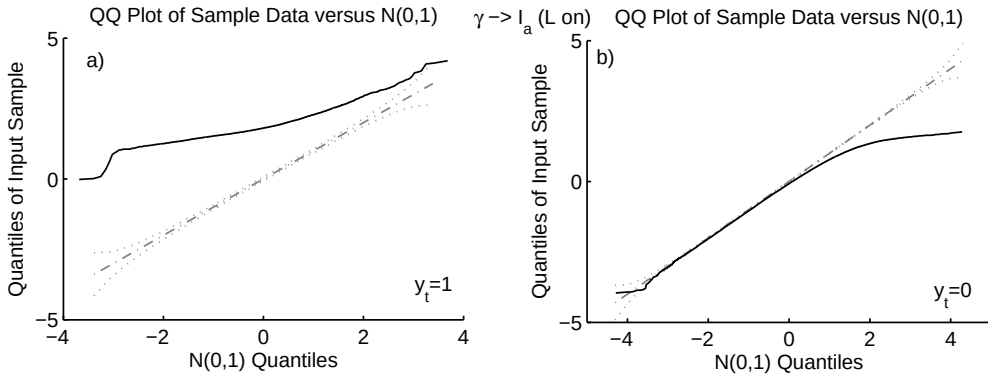
Ο εκτιμητής του λογαρίθμου της πιθανοφάνειας παίρνει την τιμή -1.5399 . Στο Σχήμα 3α παρουσιάζονται οι εκτιμητές των παραμέτρων της συνάρτησης άθροισης για την περίπτωση αυτή καθώς και το 95% διάστημα εμπιστοσύνης τύπου Wald. Παρατηρούμε ότι η μηδενική τιμή περιέχεται στο διάστημα εμπιστοσύνης σε όλο το διάστημα από 0 έως 80 msec. Αυτό μας δείχνει ότι η συμπεριφορά του συστήματος δε μεταβάλλεται εξαιτίας της επίδρασης του γάμμα κινητονεύρου. Στο Σχήμα 3β παρουσιάζονται η συνάρτηση ανάκαμψης μαζί με τη συνάρτηση κατωφλίου και τα 95% διαστήματα εμπιστοσύνης τους. Παρατηρούμε ότι η συνάρτηση ανάκαμψης αυξομειώνεται μέχρι το 60 msec, αλλά μετά αυξάνει απότομα, τέμνει το κατώφλι περίπου στο 650 msec και μετά παραμένει σε μεγάλες τιμές. Αυτό μας δείχνει ότι το σύστημα στην περίπτωση αυτή αναμένεται να αυτοπυροδοτήσει λόγω εσωτερικών διεργασιών περίπου στο 650 msec.

Σχήμα 3: α) Η συνάρτηση άθροισης και β) η συνάρτηση ανάκαμψης, για την έξοδο I_a , με την επίδραση του μήκους του μύδς. Με γκρι χρώμα εμφανίζονται τα 95% διαστήματα εμπιστοσύνης.



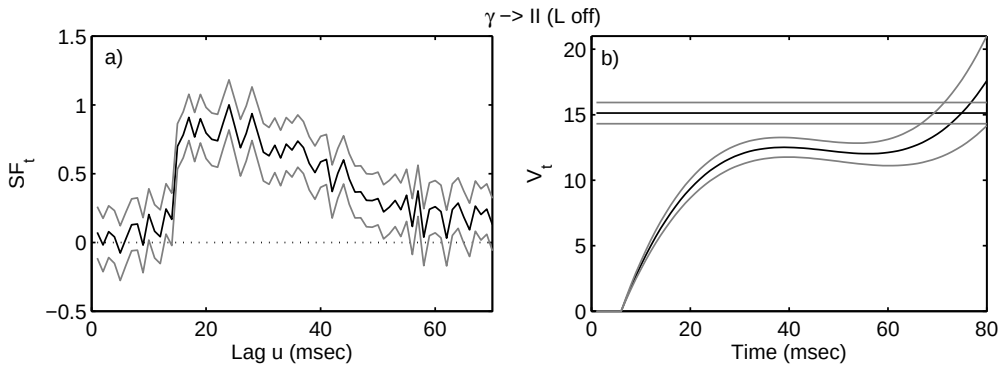
Στο Σχήμα 4α παρουσιάζεται το γράφημα Q-Q των 4352 τυχαιοποιημένων υπολοί-

Σχήμα 4: Η γραφική παράσταση Q-Q των τυχαιοποιημένων υπολοίπων του προσαρμοσμένου λογιστικού μοντέλου που αντιστοιχούν α) στις παρατηρήσεις με $y_t = 1$ και β) στις παρατηρήσεις με $y_t = 0$, για την περίπτωση της εξόδου I_a με την επίδραση του μήκους του μυός. Οι 5% απορριπτικές περιοχές μετά από 1000 επαναλήψεις Monte Carlo εμφανίζονται με γκρι χρώμα.



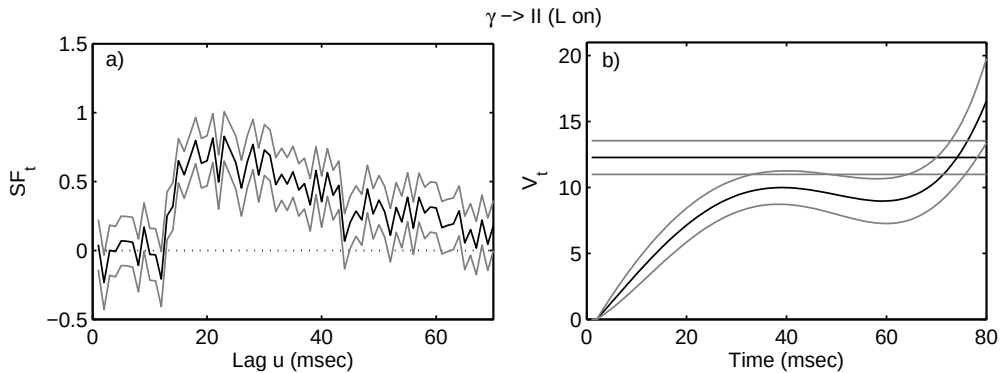
πων που αντιστοιχούν στις αποκρίσεις με $y_t = 1$. Όλα τα σημεία βρίσκονται έξω από τις 5% απορριπτικές περιοχές που υπολογίστηκαν με 1000 επαναλήψεις Monte Carlo. Η θέση των σημείων μάς υποδεικνύει ότι οι αποκρίσεις με $y_t = 1$ έχουν υπο-εκτιμηθεί. Στο Σχήμα 4b παρουσιάζεται το γράφημα Q-Q των 60000-4352=55648 τυχαιοποιημένων υπολοίπων που αντιστοιχούν στις αποκρίσεις με $y_t = 0$. Όλα τα σημεία βρίσκονται έξω από τις 5% απορριπτικές περιοχές που υπολογίστηκαν με 1000 επαναλήψεις Monte Carlo. Η θέση των σημείων μάς υποδεικνύει ότι οι αποκρίσεις με $y_t = 0$ έχουν υπερεκτιμηθεί. Σημειώνεται ότι και για τα υπόλοιπα παραδείγματα τα γραφήματα Q-Q παρουσιάζουν την ίδια μορφή και για λόγους οικονομίας παραλείπονται.

Σχήμα 5: α) Η συνάρτηση άθροισης και β) η συνάρτηση ανάκαμψης, για την έξοδο II , χωρίς την επίδραση του μήκους του μυός. Με γκρι χρώμα εμφανίζονται τα 95% διαστήματα εμπιστοσύνης.



Στην περίπτωση της II εξόδου όταν δεν επηρεάζεται από το μήκος του μυός ($\gamma \rightarrow II(L - \text{off})$) εκτιμήθηκαν 74 παράμετροι (70 για τη συνάρτηση άθροισης, 3 για τη συνάρτηση ανάκαμψης και 1 για το κατώφλι). Ο εκτιμητής του λογαρίθμου της πιθανότητας

Σχήμα 6: α) Η συνάρτηση άθροισης και β η συνάρτηση ανάκαμψης, για την έξοδο II, με την επίδραση του μήκους του μυός. Με γκρι χρώμα εμφανίζονται τα 95% διαστήματα εμπιστοσύνης.



νοφάνειας παίρνει την τιμή -6273.9 . Στο Σχήμα 5α παρουσιάζονται οι εκτιμητές των παραμέτρων της συνάρτησης άθροισης για την περίπτωση αυτή καθώς και το 95% διάστημα εμπιστοσύνης τύπου Wald. Παρατηρούμε ότι οι συντελεστές της συνάρτησης άθροισης είναι θετικοί μεταξύ 20 και 60 msec και έξω από το διάστημα εμπιστοσύνης. Αυτό μας δείχνει ότι η συμπεριφορά του συστήματος αρχικά δε μεταβάλλεται, ενώ η απόκριση του II επιταχύνεται στο διάστημα από 20 έως 60 msec εξαιτίας της επίδρασης του γάμμα κινητονεύρου. Στο Σχήμα 5β παρουσιάζονται η συνάρτηση ανάκαμψης μαζί με τη συνάρτηση κατώφλιου και τα 95% διαστήματα εμπιστοσύνης τους. Παρατηρούμε ότι η συνάρτηση ανάκαμψης συνεχώς αυξάνεται και τέμνει το κατώφλι περίπου στο 650 msec και μετά παραμένει σε μεγάλες τιμές. Αυτό μας δείχνει ότι το σύστημα στην περίπτωση αυτή αναμένεται να αυτοπυροδοτήσει λόγω εσωτερικών διεργασιών περίπου στο 650 msec.

Στην περίπτωση της II εξόδου όταν επηρεάζεται από το μήκος του μυός ($\gamma \rightarrow II(L - on)$) εκτιμήθηκαν 75 παράμετροι (70 για τη συνάρτηση άθροισης, 4 για τη συνάρτηση ανάκαμψης και 1 για το κατώφλι). Ο εκτιμητής του λογαρίθμου της πιθανοφάνειας παίρνει την τιμή -6673.7 . Στο Σχήμα 6α παρουσιάζονται οι εκτιμητές των παραμέτρων της συνάρτησης άθροισης για την περίπτωση αυτή καθώς και το 95% διάστημα εμπιστοσύνης τύπου Wald. Παρατηρούμε ότι συμπεριφορά της συνάρτησης άθροισης μοιάζει με αυτή του Σχήματος 4α άλλα η επιτάχυνση του συστήματος λόγω της επίδρασης του γάμμα κινητονεύρου αρχίζει μερικά (~ 5) msec νωρίτερα και υπάρχει ένδειξη ότι διαρκεί λιγότερο. Το Σχήμα 6β πάλι μας δείχνει ότι το σύστημα στην περίπτωση αυτή αναμένεται να αυτοπυροδοτήσει λόγω εσωτερικών διεργασιών περίπου στο 650 mses.

5. ΣΥΜΠΕΡΑΣΜΑΤΑ

Στην εργασία αυτή παρουσιάστηκε μια μέθοδος μοντελοποίησης του συστήματος τη μυϊκής ατράκτου με τη βοήθεια ενός λογιστικού μοντέλου, οι παράμετροι του οποίου εκτιμήθηκαν με τη χρήση της συνάρτησης πιθανοφάνειας. Δόθηκε κυρίως βαρύτητα

στον τρόπο που μπορεί κάποιος να συμπεριλάβει στο μοντέλο τη συνάρτηση άθροισης και τη συνάρτηση ανάκαμψης για να μελετήσει τις εξωτερικές όσο και τις εσωτερικές διεργασίες του συστήματος.

Η εφαρμογή του μοντέλου στην περίπτωση της επίδρασης ενός γάμμα κινητονεύρου στην απόκριση του πρωτεύοντος και του δευτερεύοντος αισθητήριου άξονα μας δείχνει ξεκάθαρα ότι στον πρωτεύοντα άξονα η επίδραση του μήκους του μυός παίζει σημαντικό ρόλο αφού εξαλείφει την επίδραση του γάμμα κινητονεύρου και επιτρέπει την αυτοπυροδότηση του συστήματος. Στον δευτερεύοντα άξονα υπάρχει αυτοπυροδότηση του συστήματος ανεξαρτήτως από το αν στο σύστημα επιδρά στο μήκος του μυός.

Η μελέτη της εγκυρότητας των μοντέλων με τη χρήση των τυχαιοποιημένων υπολοίπων και του ελέγχου καλής προσαρμογής με τη βοήθεια του γραφηματος Q-Q μας έδειξε ότι υπάρχουν περιθώρια βελτίωσης του μοντέλου ώστε να αποφευχθεί η υπο- ή η υπερεκτίμηση των αποκρίσεων των εξόδων. Θα είχε ενδιαφέρον να διερευνηθεί στο μέλλον η χρήση διαφορετικών μορφών συναρτήσεων ανάκαμψης (όχι πολυωνυμικών) και η χρήση μεταβλητής συνάρτησης κατωφλίου.

ABSTRACT

In this work, a way to model a stochastic system of stationary point processes by using a generalized linear model (GLM) is presented. The estimation of the parameters of the GLM is achieved in the time domain through the likelihood function. The proposed stochastic model includes the threshold, the summation and the recovery functions, whose estimation is important since they describe the effect of the input to the system and the existence or not of spontaneous response. An illustrative example from the field of Neurophysiology is presented in which the behaviour of the muscle spindle is examined under the influence of a gamma motoneurone in two cases; when there are changes in the “parent” muscle’s length and when there are not.

ΑΝΑΦΟΡΕΣ

- Boyd, I.A. (1980). The isolated mammalian muscle spindle. *Trends in Neurosciences* **3**, 258-265.
- Brillinger, D.R. (1992). Nerve cell spike train data analysis: a progression of technique. *JASA* **87**, 260-271.
- Dobson, A.J. and Barnett, A.G. (2008). *An Introduction to Generalized Linear Models, Third Edition*, Boca Raton: Chapman and Hall.
- Dunn, P.K. and Smyth, G.K. (1996). Randomized quantile residuals. *J. Comput. Graph. Stat.* **5(3)**, 236-244.
- McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models, Second Edition*, London: Chapman and Hall.
- 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.

- Kotti, V.K. and Rigas, A.G. (2005). Logistic regression methods and their implementation. In: Edler, L. and Kitsos, C.P. (Eds.), *Recent Advances in Quantitative Methods in Cancer and Human Health Risk Assessment*, New York: Wiley, pp. 355-369.
- Βασυλειάδης, 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.
- Βασυλειάδης, Β.Γ. και Ρήγας, Α.Γ. (2010). Επίλυση του φαινομένου του διαχωρισμού σε λογιστικά μοντέλα με τη χρήση της ποινικοποιημένης συνάρτησης πιθανοφάνειας. *Πρακτικά 23ου Πανελληνίου Συνεδρίου Στατιστικής*, 52-59.

ΕΝΑ ΠΡΟΒΛΗΜΑ ΒΕΛΤΙΣΤΗΣ ΔΡΟΜΟΛΟΓΗΣΗΣ ΕΝΟΣ ΟΧΗΜΑΤΟΣ ΠΟΥ ΔΙΑΝΕΜΕΙ ΕΝΑ ΠΡΟΪΟΝ ΣΕ ΔΙΑΤΕΤΑΓΜΕΝΟΥΣ ΠΕΛΑΤΕΣ ΜΕ ΣΤΟΧΑΣΤΙΚΕΣ ΔΙΑΚΡΙΤΕΣ ΖΗΤΗΣΕΙΣ ΚΑΙ ΠΟΙΝΕΣ ΓΙΑ ΤΗ ΜΕΡΙΚΗ ΙΚΑΝΟΠΟΙΗΣΗ ΤΩΝ ΖΗΤΗΣΕΩΝ

Θεοδόσης Δ. Δημητράκος¹, Επαμεινώνδας Γ. Κοριακίδης²

¹Τμήμα Μαθηματικών, Πανεπιστήμιο Αιγαίου

²Τμήμα Στατιστικής, Οικονομικό Πανεπιστήμιο Αθηνών

dimitheo@aegean.gr, ekyriak@aueb.gr

ΠΕΡΙΛΗΨΗ

Θεωρούμε το πρόβλημα της διανομής ενός προϊόντος σε διάφορους πελάτες με τη βοήθεια ενός οχήματος. Υποθέτουμε ότι η ζήτηση του προϊόντος για κάθε πελάτη είναι μία διακριτή τυχαία μεταβλητή και ότι οι πελάτες εξυπηρετούνται από το όχημα σύμφωνα με μία προκαθορισμένη διάταξη. Το όχημα ξεκινάει τη διαδρομή του από μία αποθήκη με μία συγκεκριμένη ποσότητα του προϊόντος και υπάρχει ένα κόστος για τη μετακίνηση του οχήματος από τον έναν πελάτη προς τον επόμενο πελάτη. Επίσης υπάρχει ένα κόστος για τη μετακίνηση του οχήματος από τον κάθε πελάτη προς την αποθήκη. Υποθέτουμε επίσης ότι υπάρχει ένα κόστος ως ποινή (penalty cost) ανά τεμάχιο του προϊόντος που δεν διανέμεται, σε κάποιον πελάτη, αν η ζήτηση του πελάτη ικανοποιείται μερικώς ή δεν ικανοποιείται. Κατά την άφιξη του οχήματος σε κάθε πελάτη, το όχημα έχει διάφορες επιλογές ενεργειών, ανάλογα με την ποσότητα του προϊόντος που υπάρχει στο όχημα και την ποσότητα του προϊόντος που ζητείται από τον κάθε πελάτη. Το πρόβλημα είναι η εύρεση της πολιτικής που ελαχιστοποιεί το συνολικό αναμενόμενο κόστος. Αποδεικνύουμε ότι η μορφή της βέλτιστης πολιτικής για κάθε πελάτη είναι τύπου κατωφλίου (threshold-type structure) και χαρακτηρίζεται από τρεις κρίσιμους αριθμούς. Ένας κατάλληλος αλγόριθμος δυναμικού προγραμματισμού κατασκευάζεται για τον υπολογισμό της βέλτιστης πολιτικής. Δύο αριθμητικά παραδείγματα επίσης παρουσιάζονται στα οποία η ζήτηση του προϊόντος για κάθε πελάτη είναι μία διακριτή τυχαία μεταβλητή που ακολουθεί τη διωνυμική και την ομοιόμορφη κατανομή, αντίστοιχα.

Λέξεις-κλειδιά: πρόβλημα βέλτιστης δρομολόγησης, μερική εξυπηρέτηση, δυναμικός προγραμματισμός, ελάχιστο συνολικό αναμενόμενο κόστος

1. ΕΙΣΑΓΩΓΗ

Το πρόβλημα της δρομολόγησης ενός οχήματος, γνωστό στη βιβλιογραφία ως Vehicle Routing Problem (VRP) είναι ένα πρόβλημα συνδυαστικής βελτιστοποίησης με πολλές εφαρμογές στις μεταφορές και στις διανομές προϊόντων σε πελάτες. Τα βιβλία των Toth and Vigo (2002), Simchi-Levi, Chen and Bramel (2005) περιέχουν μία αναλυτική παρουσίαση αποτελεσμάτων για διάφορες μορφές του προβλήματος. Παρακάτω, περιγράφουμε ένα μοντέλο δρομολόγησης στο οποίο ένα όχημα με περιορισμένη χωρητικότητα διανέμει ένα προϊόν σε N πελάτες σύμφωνα με μία προκαθορισμένη διάταξη. Θεωρούμε ένα σύνολο κορυφών $V = \{0, \dots, N\}$, όπου η κορυφή 0 αναπαριστά την αποθήκη και οι κορυφές $1, \dots, N$ αντιστοιχούν στους πελάτες. Οι πελάτες εξυπηρετούνται από ένα όχημα σύμφωνα με την προκαθορισμένη διάταξη $1, \dots, N$. Έστω $\xi_j, j \in \{1, \dots, N\}$, ο αριθμός των τεμαχίων του προϊόντος που ζητά ο πελάτης j . Υποθέτουμε ότι η ξ_j είναι μία διακριτή τυχαία μεταβλητή με δυνατές τιμές $0, \dots, Q$, όπου Q είναι η χωρητικότητα του οχήματος. Θεωρούμε ότι η κατανομή της τυχαίας μεταβλητής $\xi_j, j \in \{1, \dots, N\}$ είναι γνωστή. Υποθέτουμε ότι η ζήτηση κάθε πελάτη γίνεται γνωστή κατά την άφιξη του οχήματος σ' αυτόν. Έστω $c_{j,j+1}, j = 1, \dots, N-1$, το κόστος της μετακίνησης από τον πελάτη j στον πελάτη $j+1$ και $c_{j0}, c_{0j}, j = 1, \dots, N$, το κόστος της μετακίνησης από τον πελάτη j προς την αποθήκη και το κόστος της μετακίνησης από την αποθήκη προς τον πελάτη j , αντίστοιχα. Είναι λογικό να υποθέσουμε ότι αυτά τα κόστη είναι συμμετρικά και ικανοποιούν την τριγωνική ανισότητα, δηλαδή ισχύει ότι: $c_{j0} = c_{0j}, j = 1, \dots, N$ και $c_{j,j+1} \leq c_{j0} + c_{0,j+1}$. Υποθέτουμε ότι το όχημα ξεκινά τη διαδρομή του από την αποθήκη με Q τεμάχια του προϊόντος. Κατά την άφιξη του οχήματος σε κάθε πελάτη, θεωρούμε ότι είναι αποδεκτό το όχημα να ικανοποιήσει μερικώς (partial satisfaction) ή να μην ικανοποιήσει τη ζήτηση του κάθε πελάτη. Σ' αυτήν την περίπτωση ένα κόστος εισάγεται ως ποινή (penalty cost) ανά τεμάχιο προϊόντος που δεν διανέμεται για κάθε πελάτη. Το όχημα μετά τη μερική ή ολική ικανοποίηση των ζητήσεων όλων των πελατών επιστρέφει στην αποθήκη. Ο στόχος μας είναι να βρούμε τη στρατηγική δρομολόγησης του οχήματος που ελαχιστοποιεί το συνολικό αναμενόμενο κόστος για μία ολοκληρωμένη διαδρομή του οχήματος. Αυτό το κόστος περιλαμβάνει: (i) κόστη μετακίνησης μεταξύ διαδοχικών πελατών, (ii) κόστη μετακίνησης μεταξύ των πελατών και της αποθήκης και (iii) κόστη εξαιτίας της μερικής ικανοποίησης των ζητήσεων. Οι Yang et al. (2000) και Pandelis et al. (2012, 2013) μελέτησαν διάφορες μορφές στοχαστικών προβλημάτων βέλτιστης δρομολόγησης ενός οχήματος. Στην παρούσα εργασία ένας κατάλληλος αλγόριθμος του δυναμικού προγραμματισμού κατασκευάζεται με τον οποίον υπολογίζουμε τη βέλτιστη πολιτική δρομολόγησης του οχήματος. Αποδεικνύουμε ότι η βέλτιστη πολιτική έχει μία ειδική μορφή κατωφλίου (threshold-type structure) η οποία χαρακτηρίζεται από τρεις κρίσιμους αριθμούς. Στο επόμενο εδάφιο παρουσιάζουμε τις εξισώσεις δυναμικού προγραμματισμού και τη

μορφή της βέλτιστης πολιτικής δρομολόγησης του οχήματος. Στο Εδάφιο 3, παρουσιάζουμε δύο αριθμητικά παραδείγματα στα οποία η ζήτηση του προϊόντος για κάθε πελάτη είναι μία διακριτή τυχαία μεταβλητή που ακολουθεί τη διωνυμική και την ομοιόμορφη κατανομή, αντίστοιχα. Επισημαίνεται ότι το περιεχόμενο της παρούσας εργασίας μαζί με διάφορες επεκτάσεις της, έχει υποβληθεί προς δημοσίευση από τους Kyriakidis and Dimitrakos (2013).

2. Η ΒΕΛΤΙΣΤΗ ΠΟΛΙΤΙΚΗ

Θεωρούμε ότι το όχημα ξεκινά τη διαδρομή του από την αποθήκη με Q τεμάχια του προϊόντος. Έστω ότι κατά την άφιξη του οχήματος στον πελάτη j , ο αριθμός των τεμαχίων του προϊόντος που υπάρχουν στο όχημα είναι ίσος με z και η ζήτηση του προϊόντος είναι ίση με s . Για $j \in \{1, 2, \dots, N-1\}$, αν $z \geq s$, το όχημα έχει δύο επιλογές: (i) να ταξιδέψει κατευθείαν προς τον επόμενο πελάτη $j+1$ για να τον εξυπηρετήσει (Ενέργεια 1) ή (ii) να επιστρέψει στην αποθήκη για να ανανεώσει το απόθεμά του με Q τεμάχια του προϊόντος και εν συνεχεία να ταξιδέψει προς τον επόμενο πελάτη $j+1$ για να τον εξυπηρετήσει (Ενέργεια 2). Για $j \in \{1, 2, \dots, N-1\}$, αν $z < s$, το όχημα έχει τέσσερις επιλογές: (i) να ταξιδέψει κατευθείαν προς τον επόμενο πελάτη $j+1$ έχοντας υποστεί ένα κόστος λόγω της μη-ικανοποίησης της ζήτησης του προϊόντος ίσο με $\pi_j(s-z)$ (Ενέργεια 1*), (ii) να επιστρέψει στην αποθήκη για να ανανεώσει το απόθεμά του με Q τεμάχια του προϊόντος και εν συνεχεία να ταξιδέψει προς τον επόμενο πελάτη $j+1$ έχοντας υποστεί ένα κόστος λόγω της μη-ικανοποίησης της ζήτησης του προϊόντος ίσο με $\pi_j(s-z)$ (Ενέργεια 2*), (iii) να επιστρέψει στην αποθήκη για να ανανεώσει το απόθεμά του με Q τεμάχια του προϊόντος, να επιστρέψει στον πελάτη j για να του δώσει $\theta \in \{1, \dots, s-z\}$ τεμάχια του προϊόντος και εν συνεχεία να ταξιδέψει προς τον επόμενο πελάτη $j+1$ με τα υπόλοιπα $Q-\theta$ τεμάχια του προϊόντος έχοντας υποστεί ένα κόστος λόγω της μη-ικανοποίησης της ζήτησης του προϊόντος ίσο με $\pi_j(s-z-\theta)$ (Ενέργεια 3), (iv) να επιστρέψει στην αποθήκη για να ανανεώσει το απόθεμά του με την οφειλόμενη ποσότητα $s-z$, να επιστρέψει στον πελάτη j για να του δώσει την οφειλόμενη ποσότητα, να κάνει ένα δεύτερο ταξίδι στην αποθήκη για να ανανεώσει το απόθεμά του με Q τεμάχια του προϊόντος και εν συνεχεία να ταξιδέψει προς τον επόμενο πελάτη $j+1$ (Ενέργεια 4). Για $j = N$, αν $z \geq s$, το όχημα ταξιδεύει προς την αποθήκη για να ολοκληρώσει τη διαδρομή του. Αν $z < s$, το όχημα έχει δύο επιλογές: (i) να ταξιδέψει προς την αποθήκη για να ολοκληρώσει τη διαδρομή του ή (ii) να ταξιδέψει προς την αποθήκη για να ανανεώσει το απόθεμά του με την οφειλόμενη ποσότητα $s-z$, να επιστρέψει στον πελάτη N για να του δώσει την οφειλόμενη ποσότητα και εν συνεχεία να ταξιδέψει προς την αποθήκη για να ολοκληρώσει τη διαδρομή του έχοντας υποστεί ένα κόστος λόγω της μη-

ικανοποίησης της ζήτησης του προϊόντος ίσο με $\pi_N(s-z)$. Αναζητούμε την πολιτική που ελαχιστοποιεί το συνολικό αναμενόμενο κόστος για μία ολοκληρωμένη διαδρομή του οχήματος. Έστω $f_j(z, s), j \in \{1, \dots, N\}, z \in \{0, \dots, Q\}, s \in \{0, \dots, Q\}$, το ελάχιστο συνολικό αναμενόμενο μελλοντικό κόστος μέχρι το όχημα να ολοκληρώσει τη διαδρομή του, όταν κατά την άφιξη του οχήματος στον πελάτη j , ο αριθμός των τεμαχίων του προϊόντος που υπάρχουν στο όχημα είναι ίσος με z και η ζήτηση του προϊόντος είναι ίση με s . Μία βέλτιστη στρατηγική δρομολόγησης του οχήματος μπορεί να υπολογιστεί από τις παρακάτω εξισώσεις δυναμικού προγραμματισμού:

Για $j \in \{1, 2, \dots, N-1\}$, αν $z < s$, έχουμε ότι:

$$f_j(z, s) = \min\{H_j(z, s), R_j\}, \quad (1)$$

όπου,

$$H_j(z, s) = c_{j,j+1} + \min_{0 \leq \theta \leq s} \{\pi_j(s - \theta) + Ef_{j+1}(z - \theta, \xi_{j+1})\},$$

$$R_j = c_{0j} + c_{j+1,0} + Ef_{j+1}(Q, \xi_{j+1}).$$

Για $j \in \{1, 2, \dots, N-1\}$, αν $z < s$, έχουμε ότι:

$$f_j(z, s) = \min\{A_j(z, s), B_j(z, s), C_j(z, s), D_j\}, \quad (2)$$

όπου,

$$A_j(z, s) = c_{j,j+1} + \pi_j(s - z) + Ef_{j+1}(0, \xi_{j+1}),$$

$$B_j(z, s) = c_{0j} + c_{0,j+1} + \pi_j(s - z) + Ef_{j+1}(Q, \xi_{j+1}),$$

$$C_j(z, s) = 2c_{0j} + c_{j,j+1} + \min_{0 < \theta \leq s-z} \{\pi_j(s - z - \theta) + Ef_{j+1}(Q - \theta, \xi_{j+1})\},$$

$$D_j = 3c_{0j} + c_{0,j+1} + Ef_{j+1}(Q, \xi_{j+1}).$$

Η οριακή συνθήκη είναι:

$$f_N(z, s) = c_{0N} + \min\{2c_{0N}, \pi_N(s - z)\}1(z < s).$$

Το ελάχιστο συνολικό αναμενόμενο κόστος είναι ίσο με $f_0 = c_{01} + Ef_1(Q, \xi_1)$. Στις παραπάνω εξισώσεις, οι αναμενόμενες τιμές υπολογίζονται ως προς τις τυχαίες μεταβλητές $\xi_j, j = 1, \dots, N$. Ο πρώτος όρος ανάμεσα στα άγκιστρα στην Εξίσωση (1) αντιστοιχεί στην Ενέργεια 1 ενώ ο δεύτερος όρος αντιστοιχεί στην Ενέργεια 2. Ο πρώτος όρος ανάμεσα στα άγκιστρα στην Εξίσωση (2) αντιστοιχεί στην Ενέργεια 1*, ο δεύτερος όρος αντιστοιχεί στην Ενέργεια 2*, ο τρίτος όρος αντιστοιχεί στην Ενέργεια 3 και ο τέταρτος όρος αντιστοιχεί στην Ενέργεια 4.

Είναι δυνατόν να αποδειχθεί το παρακάτω θεώρημα που δίνει τη μορφή της βέλτιστης πολιτικής δρομολόγησης του οχήματος.

Θεώρημα. (i) Για κάθε πελάτη $j \in \{1, \dots, N-1\}$ και για κάθε ζήτηση προϊόντος $s \in \{0, \dots, Q\}$, αν $z \geq s$, υπάρχει ένας κρίσιμος ακέραιος $h_1(j, s)$ τέτοιος ώστε η βέλτιστη ενέργεια είναι η Ενέργεια 1 αν $z \geq h_1(j, s)$, ενώ αν $z < h_1(j, s)$ η βέλτιστη ενέργεια είναι η Ενέργεια 2.
(ii) Για κάθε πελάτη $j \in \{1, \dots, N-1\}$ και για κάθε ζήτηση προϊόντος $s \in \{0, \dots, Q\}$, αν $z < s$, υπάρχουν δύο κρίσιμοι ακέραιοι αριθμοί $h_2(j, s), h_3(j, s)$ με $h_3(j, s) \leq h_2(j, s)$ τέτοιοι ώστε αν $h_2(j, s) \leq z$, η βέλτιστη ενέργεια είναι μία εκ των Ενεργειών 1* ή 2*, αν $h_3(j, s) \leq z < h_2(j, s)$ η βέλτιστη ενέργεια είναι η Ενέργεια 3 και αν $z < h_3(j, s)$ η βέλτιστη ενέργεια είναι η Ενέργεια 4.

Η απόδειξη του θεωρήματος γίνεται με επαγωγή ως προς j και παρουσιάζεται αναλυτικά στην εργασία των Kyriakidis and Dimitrakos (2013).

3. ΑΡΙΘΜΗΤΙΚΑ ΠΑΡΑΔΕΙΓΜΑΤΑ

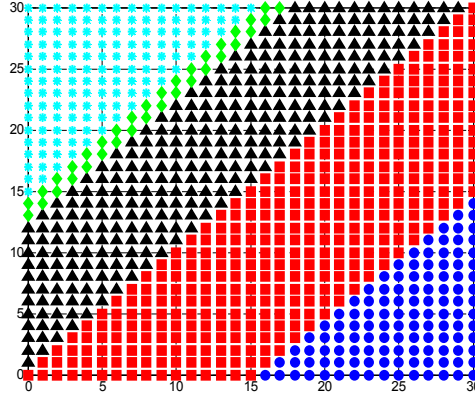
Στα παρακάτω παραδείγματα, η ζήτηση του προϊόντος για κάθε πελάτη είναι μία διακριτή τυχαία μεταβλητή που ακολουθεί τη διωνυμική και την ομοιόμορφη κατανομή, αντίστοιχα. Ο αλγόριθμος του δυναμικού προγραμματισμού υλοποιήθηκε με το πακέτο Matlab σε προσωπικό υπολογιστή τύπου Intel Core 2 Duo, 2.5 GHz και 4 GB RAM. Στα σχήματα που ακολουθούν, ο οριζόντιος άξονας περιέχει τις τιμές του αριθμού των τεμαχίων z του προϊόντος που υπάρχουν στο όχημα και ο κάθετος άξονας περιέχει τις τιμές της ζήτησης s του προϊόντος.

Παράδειγμα 1. Υποθέτουμε ότι $N = 15$ και $Q = 30$. Τα κόστη μεταξύ των πελατών j και $j+1$, $j \in \{1, \dots, 14\}$, είναι: $c_{12} = 19$, $c_{23} = 23$, $c_{34} = 19$, $c_{45} = 27$, $c_{56} = 24$, $c_{67} = 30$, $c_{78} = 23$, $c_{89} = 26$, $c_{9,10} = 19$, $c_{10,11} = 17$, $c_{11,12} = 23$, $c_{12,13} = 22$, $c_{13,14} = 25$ και $c_{14,15} = 28$. Τα κόστη μεταξύ των πελατών j , $j \in \{1, \dots, 15\}$, και της αποθήκης (κορυφή 0) είναι: $c_{10} = 23$, $c_{20} = 21$, $c_{30} = 18$, $c_{40} = 20$, $c_{50} = 25$, $c_{60} = 22$, $c_{70} = 27$, $c_{80} = 19$, $c_{90} = 24$, $c_{10,0} = 25$, $c_{11,0} = 20$, $c_{12,0} = 26$, $c_{13,0} = 18$, $c_{14,0} = 27$ και $c_{15,0} = 30$. Παρατηρούμε ότι αυτά τα κόστη ικανοποιούν την τριγωνική ανισότητα. Θεωρούμε ότι, για κάθε πελάτη $j \in \{1, \dots, 15\}$, το κόστος π_j είναι ίσο με 3. Επιπλέον υποθέτουμε ότι, για κάθε πελάτη $j \in \{1, \dots, 15\}$, η ζήτηση ξ_j του προϊόντος είναι μία διακριτή τυχαία μεταβλητή που ακολουθεί τη διωνυμική κατανομή $Bin(Q, p)$,

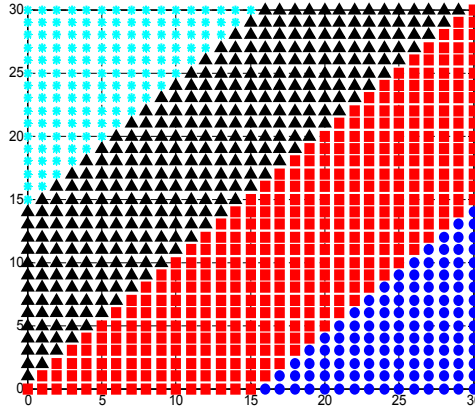
δηλαδή ισχύει ότι: $\Pr(\xi_j = x) = \binom{Q}{x} p^x (1-p)^{Q-x}, x = 0, \dots, Q$. Για $p = 0.4$, στα

Σχήματα 1 και 2, παρουσιάζουμε τη μορφή της βέλτιστης πολιτικής για τον Πελάτη 4 και τον Πελάτη 6, αντίστοιχα. Για $z \geq s$, η Ενέργεια 1 σημειώνεται με κύκλους και η Ενέργεια 2 σημειώνεται με τετράγωνα. Για $z < s$, χρησιμοποιούμε: (i) τα σύμβολα “x” για να αναπαραστήσουμε την Ενέργεια 1* που αντιστοιχεί στην ποσότητα $A_j(z, s)$, (ii) τα άνω τρίγωνα για να αναπαραστήσουμε την Ενέργεια 2* που αντιστοιχεί στην ποσότητα $B_j(z, s)$, (iii) τους ρόμβους για να αναπαραστήσουμε την Ενέργεια 3 που αντιστοιχεί στην ποσότητα $C_j(z, s)$ και (iv) τους αστερίσκους για να αναπαραστήσουμε την Ενέργεια 4 που αντιστοιχεί στην ποσότητα D_j .

Σχήμα 1. Οι βέλτιστες αποφάσεις για τον Πελάτη 4

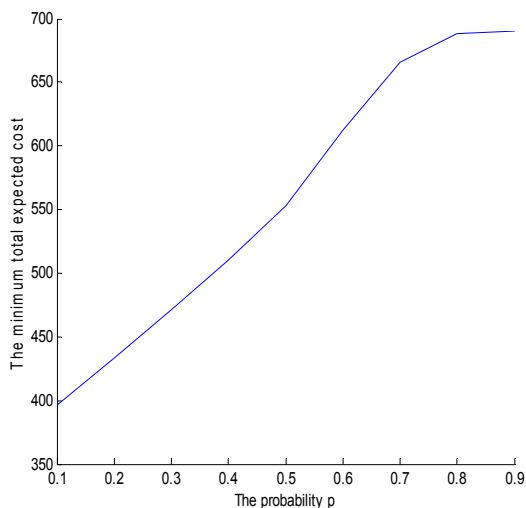


Σχήμα 2. Οι βέλτιστες αποφάσεις για τον Πελάτη 6



Παρατηρούμε ότι στα Σχήματα 1 και 2 δεν εμφανίζεται ως βέλτιστη ενέργεια η Ενέργεια 1*. Επισημαίνουμε ότι, στο συγκεκριμένο παράδειγμα, η Ενέργεια 1* δεν εμφανίζεται ως βέλτιστη ενέργεια για κανέναν πελάτη. Το γεγονός αυτό οφείλεται στη δομή κόστους που επιλέχθηκε στο συγκεκριμένο παράδειγμα. Η τιμή του ελάχιστου συνολικού αναμενόμενου κόστους f_0 βρέθηκε, κατά προσέγγιση, ίση με 507.43. Στο Σχήμα 3, παρουσιάζεται ένα γράφημα που δείχνει τη μεταβολή του ελάχιστου συνολικού αναμενόμενου κόστους f_0 καθώς η πιθανότητα p παίρνει τιμές στο σύνολο $\{0.1, \dots, 0.9\}$.

Σχήμα 3. Το ελάχιστο συνολικό αναμενόμενο κόστος f_0 καθώς μεταβάλλεται η πιθανότητα p



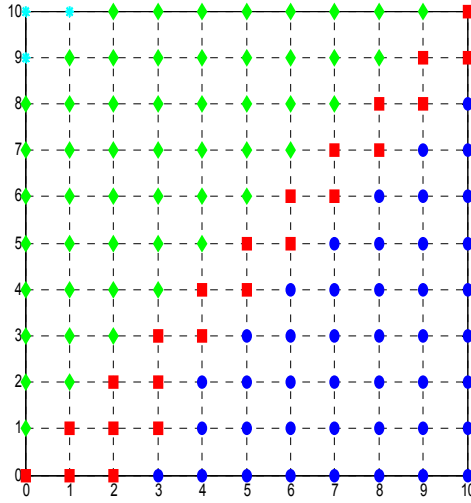
Παρατηρούμε ότι καθώς η πιθανότητα p παίρνει τιμές στο σύνολο $\{0.1, \dots, 0.6\}$ το ελάχιστο συνολικό αναμενόμενο κόστος αυξάνεται, κατά προσέγγιση, γραμμικά. Όταν η πιθανότητα p παίρνει τιμές στο σύνολο $\{0.7, 0.8, 0.9\}$ το ελάχιστο συνολικό αναμενόμενο κόστος αυξάνεται αργά. ■

Παράδειγμα 2. Έστω ότι $N = 5$ και $Q = 10$. Τα κόστη μεταξύ των πελατών j και $j+1$, $j \in \{1, \dots, 5\}$, είναι: $c_{12} = 4$, $c_{23} = 6$, $c_{34} = 5$ και $c_{45} = 4$. Τα κόστη μεταξύ των πελατών j , $j \in \{1, \dots, 5\}$ και της αποθήκης (κορυφή 0) είναι: $c_{10} = 6$, $c_{20} = 10$, $c_{30} = 8$, $c_{40} = 7$ και $c_{50} = 9$. Παρατηρούμε ότι αυτά τα κόστη ικανοποιούν την τριγωνική ανισότητα.

Θεωρούμε ότι, για κάθε πελάτη $j \in \{1, \dots, 5\}$, τα κόστη π_j δίνονται από το ακόλουθο διάνυσμα: $(\pi_1, \pi_2, \pi_3, \pi_4, \pi_5) = (1.5, 2, 1.7, 1.9, 2)$. Επιπλέον υποθέτουμε

ότι, για κάθε πελάτη $j \in \{1, \dots, 5\}$, η ζήτηση ξ_j του προϊόντος είναι μία διακριτή τυχαία μεταβλητή που ακολουθεί την ομοιόμορφη κατανομή, δηλαδή ισχύει ότι: $\Pr(\xi_j = x) = (Q+1)^{-1}, x \in \{0, \dots, Q\}$. Στο Σχήμα 4, παρουσιάζουμε τη μορφή της βέλτιστης πολιτικής για τον Πελάτη 1.

Σχήμα 4. Οι βέλτιστες αποφάσεις για τον Πελάτη 1



Στον Πίνακα 1, για τις καταστάσεις (z, s) , για τις οποίες $s = 3$ και $z \in \{5, \dots, 10\}$, όπου η ενέργεια που επιλέγεται από τη βέλτιστη πολιτική είναι η Ενέργεια 1, παρουσιάζονται οι βέλτιστες τιμές της ποσότητας θ .

Πίνακας 1. Οι βέλτιστες τιμές του θ όταν η Ενέργεια 1 είναι βέλτιστη

Καταστάσεις (z, s)	Οφειλόμενη ποσότητα s	Βέλτιστη τιμή του θ
(5,3)	3	0
(6,3)	3	0
(7,3)	3	0
(8,3)	3	1
(9,3)	3	2
(10,3)	3	3

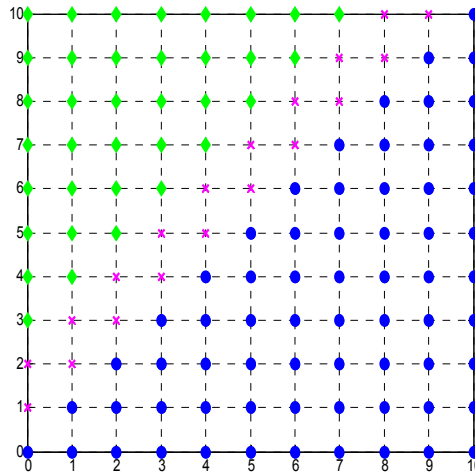
Στον Πίνακα 2, για τις καταστάσεις (z, s) , για τις οποίες $s = 6$ και $z \in \{0, \dots, 5\}$, όπου η ενέργεια που επιλέγεται από τη βέλτιστη πολιτική είναι η Ενέργεια 3, παρουσιάζονται οι βέλτιστες τιμές της ποσότητας θ .

Πίνακας 2. Οι βέλτιστες τιμές του θ όταν η Ενέργεια 3 είναι βέλτιστη

Καταστάσεις (z, s)	Οφειλόμενη ποσότητα $s - z$	Βέλτιστη τιμή του θ
(0,6)	6	3
(1,6)	5	3
(2,6)	4	3
(3,6)	3	3
(4,6)	2	2
(5,6)	1	1

Στο Σχήμα 5, παρουσιάζουμε τη μορφή της βέλτιστης πολιτικής για τον Πελάτη 4.

Σχήμα 5. Οι βέλτιστες αποφάσεις για τον Πελάτη 4



Όπως φαίνεται στο Σχήμα 5, για τον Πελάτη 4, παρατηρούμε ότι η Ενέργεια 1* εμφανίζεται ως βέλτιστη ενέργεια. Η τιμή του ελάχιστου συνολικού αναμενόμενου κόστους f_0 βρέθηκε, κατά προσέγγιση, ίση με 56.23. ■

ABSTRACT

We consider a stochastic vehicle routing problem in which the customers are served according to a predefined sequence and the demands of the customers are discrete random variables. It is assumed that a penalty cost is imposed if a customer's demand is not satisfied or if it is satisfied partially. The objective is the determination of the policy that serves the customers with the minimum total expected cost. A suitable dynamic programming algorithm is developed for the determination of the optimal policy. It is proved that the optimal policy has a specific threshold-type structure which is characterized by three critical numbers. Two numerical examples are also presented in which the demands of the customers for the product are discrete random variables which follow the binomial and the uniform distribution, respectively.

Keywords: optimal routing problem, partial service, dynamic programming, minimum total expected cost

ΑΝΑΦΟΡΕΣ

- Kyriakidis E. G. and Dimitrakos T. D. (2013). A vehicle routing problem with a predefined customer sequence, stochastic demands and penalties for unsatisfied demands, Submitted for publication.
- Pandelis D. G., Kyriakidis E. G. and Dimitrakos T. D. (2012). Single vehicle routing problems with a predefined customer sequence, compartmentalized load and stochastic demands, *European Journal of Operational Research* **217**, 324-332.
- Pandelis D. G., Karamatsoukis C. C. and Kyriakidis E. G. (2013). Single vehicle routing problems with a predefined customer order, unified load and stochastic discrete demands, *Probability in the Engineering and Informational Sciences* **27** (1), 1-23.
- Simchi-Levi D., Chen X. and Bramel J. (2005). The Logic of Logistics. Theory, Algorithms and Applications for Logistics and Supply Chain Management, 2nd Edition, Springer, New York.
- Toth P. and Vigo D. (Eds.), (2002). The Vehicle Routing Problem, Siam, Philadelphia.
- Yang W-H., Mathur K. and Ballou R. H. (2000). Stochastic vehicle routing problem with restocking, *Transportation Science* **34**, 99-112.



ΠΟΙΟΤΙΚΟΣ ΕΛΕΓΧΟΣ ΚΑΙ ΑΠΟΔΟΤΙΚΟΤΗΤΑ ΤΩΝ ΜΕΤΟΧΩΝ ΤΩΝ ΕΛΛΗΝΙΚΩΝ ΕΤΑΙΡΕΙΩΝ ΜΕΤΑΠΟΙΗΤΙΚΟΥ ΚΛΑΔΟΥ

Α. Δονάτου, Ι.Λεβεντίδης

Τμήμα Οικονομικών Επιστημών, Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών

adonatu@yahoo.gr , ylevent@econ.uoa.gr

Περίληψη

Στόχος της εργασίας είναι να ερευνηθεί η σχέση μεταξύ της ποιότητας της λειτουργίας των συστημάτων και διαδικασιών ελέγχου και της αποδοτικότητας των μεγάλων ελληνικών εταιρειών του μεταποιητικού κλάδου, σε βραχυχρόνια και μακροχρόνια περίοδο. Μεταξύ των άλλων εξετάστηκε η σχέση της πιθανότητας λάθους λογιστικών χειρισμών στους ισολογισμούς και της μεταβλητότητας απόδοσης των μετοχών, τόσο συνολικά για εταιρείες του μεταποιητικού κλάδου όσο και για εταιρείες κατηγοριών του κλάδου (καλές, μέτριες και κακές, ως προς τους λογιστικούς χειρισμούς).

Λέξεις κλειδιά: Ποιοτικός έλεγχος, εταιρείες μεταποιητικού κλάδου, αποδοτικότητα μετοχών

1. ΕΙΣΑΓΩΓΗ

Η ποιότητα του ελέγχου εξαρτάται σε μεγάλο βαθμό από το πώς οι ελεγκτές εφαρμόζουν τα ελεγκτικά πρότυπα (Krishnan and Schauer, 2000; McConnell and Banks, 1998; Aldhizer et al., 1995; Cook, 1987). Όταν η ποιότητα του ελέγχου είναι ικανοποιητική τότε, όπως είναι αναμενόμενο, τα αποτελέσματα, που απορρέουν από τις χρηματοοικονομικές καταστάσεις, λειτουργούν ως αξιόπιστη πληροφόρηση, η οποία κατευθύνει τα κεφάλαια της εταιρείας σε επενδυτικά προγράμματα, που αυξάνουν την αξία των μετοχών (Ball et al., 2000). Η καταγραφή ορθών ή λανθασμένων λογιστικών χειρισμών (False Financial Statements, (FFS)) στους Ισολογισμούς μιας εταιρείας, που αφορά κυρίως τον, αντίστοιχο, χειρισμό οικονομικών δεδομένων, μπορεί να θεωρηθεί ως σοβαρή ένδειξη για την ποιότητα των συστημάτων και διαδικασιών εσωτερικού ή εξωτερικού ελέγχου. Σύμφωνα με τα Πρότυπα Ελεγκτικών Καταστάσεων (SAS), που

διατυπώθηκαν από το Αμερικάνικο Ινστιτούτο Πιστοποιημένων Λογιστών (AICPA), ως λάθος λογιστικοί χειρισμοί αναφέρονται οι ακούσιες λανθασμένες διατυπώσεις ή παραλείψεις των ποσών ή των κοινοποιήσεων των οικονομικών καταστάσεων. Η αλλοίωση των στοιχείων εμφανίζεται με διάφορες μορφές, όπως η υπερτίμηση των παγίων, των πωλήσεων και κερδών ή η μείωση των υποχρεώσεων, των εξόδων, ή των πιθανών απωλειών που σημειώθηκαν σε μια λογιστική χρήση. Στην Ελλάδα, η εμφάνιση λάθους λογιστικών χειρισμών βρίσκεται στο επίκεντρο του ενδιαφέροντος, κυρίως από το έτος 2000, που αποτέλεσε ένα κρίσιμο σημείο αναφοράς για το Ελληνικό Χρηματιστήριο, το οποίο έπασχε από στασιμότητα τόσο από πλευράς τιμών των μετοχών όσο και από πλευράς εμπορευσιμότητας.

Η σχέση της μεταβλητότητας της μετοχής με την οικονομική κατάσταση της επιχείρησης είναι ένα θέμα που έχει απασχολήσει και στο παρελθόν το ενδιαφέρον των ερευνητών. Μία από τις σημαντικότερες προσεγγίσεις του θέματος αυτού, είναι οι εργασίες του Merton (1974), σχετικά με την πρόβλεψη αθέτησης υποχρεώσεων των επιχειρήσεων καθώς και τον υπολογισμό της αντίστοιχης πιθανότητας αθέτησης. Η μεταβλητότητα της αξίας της επιχείρησης (σ_v) δεν είναι μια άμεσα παρατηρήσιμη ποσότητα, αφού η αξία της επιχείρησης δεν αποτελεί διαπραγματεύσιμο μέγεθος σε κάποια αντίστοιχη αγορά. Ένα βασικό βήμα στην επίλυση αυτού του προβλήματος, είναι η σύνδεση της μεταβλητότητας σ_v με την μεταβλητότητα των αποδόσεων της μετοχής (σ_E). Η σχέση αυτή οδηγεί στη σχέση της πιθανότητας αθέτησης (PD) με τη μεταβλητότητα σ_E : $PD = f(\sigma_E)$. Αυτό σημαίνει, ότι η μεταβλητότητα των αποδόσεων της μετοχής αποτελεί προβλεπτικό παράγοντα για το κατά πόσο η επιχείρηση θα οδηγηθεί σε αθέτηση των υποχρεώσεων της. Επίσης, σημαίνει, ότι η μεταβλητότητα σ_E μπορεί να οδηγήσει στον υπολογισμό της PD, δηλαδή της αντίστοιχης πιθανότητας αθέτησης.

Στο πλαίσιο αυτό, ιδιαίτερη σημασία έχει να ερευνηθεί η σχέση της μεταβλητότητας της απόδοσης της μετοχής με την πιθανότητα ύπαρξης λανθασμένων λογιστικών χειρισμών. Η μεγαλύτερη σ_E σημαίνει υψηλότερη PD (και άρα χειρότερη πιστοληπτική αξιολόγηση της επιχείρησης και επιβάρυνση των βασικών οικονομικών δεικτών) και, συνεπώς, ενδεχόμενη προσπάθεια «ωραιοποίησης» των οικονομικών στοιχείων της επιχείρησης. Η «ωραιοποίηση» των λογιστικών καταστάσεων από μια επιχείρηση, προκειμένου να δείχνει αυξημένα ίδια κεφάλαια ή αυξημένη κερδοφορία, πιθανώς σημαίνει ότι οι μετοχές της είναι ή ήταν υπερτιμημένες ή ότι η οικονομική ευρωστία της είναι πλασματική, με πιθανή συγκάλυψη εμφάνισης ζημιών, χαμηλής αποδοτικότητας και πιστωτικών προβλημάτων. Βέβαια, η ύπαρξη σοβαρών λανθασμένων λογιστικών χειρισμών είναι ένδειξη προβληματικής και ανέντιμης Διοίκησης, η οποία και δεν εφαρμόζει τις αρχές της διαφάνειας, της δεοντολογίας και της εταιρικής διακυβέρνησης, αλλά και αποτελεί από μόνη της σοβαρό λόγο για την έλλειψη εμπιστοσύνης της αγοράς προς την επιχείρηση.

Στην εργασία αυτή, επιχειρήθηκε να εξετασθεί κατά πόσο η ποιότητα των συστημάτων και διαδικασιών ελέγχου συνδέεται με την αποδοτικότητα των μετοχών των εταιρειών, με βάση δείγμα εισηγμένων στο χρηματιστήριο ελληνικών εταιρειών του μεταποιητικού κλάδου.

Ειδικότερα:

(α) Εξετάζεται αν η πιθανότητα λάθους λογιστικών χειρισμών στους ισολογισμούς των εταιρειών, για τα οικονομικά έτη 2000, 2001, 2002, 2003 και 2004 συνδέεται με τη μεταβλητότητα των αποδόσεων των μετοχών για βραχυχρόνια και μακροχρόνια περίοδο.

(β) Ερευνάται η σχέση της μεταβλητότητας της απόδοσης των μετοχών και του “ποιοτικού ελέγχου” που χαρακτηρίζει τις τρεις κατηγορίες εταιρειών, στις οποίες διαχωρίστηκε το δείγμα των εταιρειών του μεταποιητικού κλάδου (κακές, μέτριες και καλές ως προς τους λογιστικούς χειρισμούς), με χρήση κατάλληλων κριτηρίων. Ειδικότερα, μελετάται για βραχυχρόνια και μακροχρόνια περίοδο:

(i) Η συμπεριφορά της συνάρτησης μεταβλητότητας των αποδόσεων των μετοχών, σε σχέση με την πιθανότητα ύπαρξης λάθους λογιστικών χειρισμών στους ισολογισμούς, για την υπό εξέταση χρονική περίοδο, ως προς τις ομάδες των εταιρειών που προέκυψαν από την κατηγοριοποίηση.

(ii) Η συνολική επίδραση των πιθανοτήτων κακής ποιότητας λογιστικών χειρισμών στη μεταβλητότητα απόδοσης των εταιρειών των τριών ομάδων.

2. ΕΜΠΕΙΡΙΚΗ ΑΝΑΛΥΣΗ

2.1. Υπόδειγμα και δεδομένα

Στην έρευνα αυτή χρησιμοποιήθηκαν εβδομαδιαία δεδομένα της χρονικής περιόδου 2000-2004 για 61 Ελληνικές μεταποιητικές εταιρείες, επειδή μόνο για αυτές υπήρχαν τα απαιτούμενα στοιχεία. Τα στοιχεία έχουν ληφθεί από τους ισολογισμούς των εταιρειών του δείγματος των ετών 2000, 2001, 2002, 2003 και 2004. Για τη λήψη δεδομένων χρησιμοποιήθηκαν οι βάσεις: site ase.gr (για τιμές μετοχών) και Hellastat (για λοιπά οικονομικά στοιχεία). Αρχικά, υπολογίστηκαν οι αριθμοδείκτες, οι οποίοι βοηθούν στη διάγνωση της ποιότητας λογιστικών χειρισμών, και εκτιμήθηκαν οι πιθανότητες λανθασμένων λογιστικών χειρισμών για την υπό εξέταση περίοδο. Ειδικότερα, οι πιθανότητες πραγματοποίησης λάθους λογιστικών χειρισμών, prob2000, prob2001, prob2002, prob2003, prob2004, προέκυψαν από τη χρησιμοποίηση των αριθμοδεικτών επενδύσεις / πωλήσεις (investments/sales), καθαρά κέρδη / πωλήσεις (net profit/sales), κεφάλαιο κινήσεως / σύνολο ενεργητικού (working capital / total assets). Στη συνέχεια, πολλαπλασιάστηκαν οι τιμές, που προέκυψαν, με τον αντίστοιχο συντελεστή βαρύτητας (beta), ο οποίος βρέθηκε από την εφαρμογή του υποδείγματος του Spathis (2002), και ο οποίος συμβάλλει στην ανίχνευση λάθους ή ψευδών δηλώσεων στις οικονομικές καταστάσεις. Το αντίστοιχο άθροισμα των σχετικών τιμών, που υπολογίστηκαν, δίνει πιθανότητες για κάθε έτος και για κάθε εξεταζόμενη εταιρεία του κλάδου (Spathis (2002)). Ακόμη, υπολογίστηκαν οι τυπικές αποκλίσεις των εβδομαδιαίων αποδόσεων των μετοχών για τα δύο πρώτα έτη και για όλα τα έτη της υπό εξέτασης χρονικής περιόδου. Αναλυτικά, προκειμένου να ερευνηθεί η σχέση που υπάρχει, μεταξύ της πορείας των τιμών των μετοχών της εταιρείας και του ποιοτικού ελέγχου, εφαρμόστηκε ένα γραμμικό στοχαστικό υπόδειγμα. Ως εξαρτημένη μεταβλητή χρησιμοποιήθηκε η

τυπική απόκλιση των εβδομαδιαίων αποδόσεων της μετοχής και ως ανεξάρτητες μεταβλητές οι πιθανότητες για ύπαρξη λάθους λογιστικών χειρισμών στους ισολογισμούς των εταιρειών του δείγματος.

Σημειώνεται ότι η χρονική περίοδος 1999-2000 ήταν μια κρίσιμη μεταβατική περίοδος για το Ελληνικό Χρηματιστήριο και για το λόγο αυτό εξετάστηκε και η βραχυχρόνια και η μακροχρόνια κατάσταση. Η τυπική απόκλιση των αποδόσεων των μετοχών (std2) των δύο πρώτων ετών (2000, 2001) που χρησιμοποιήθηκε, διαδοχικά, ως εξαρτημένη μεταβλητή, απεικονίζει βραχυπρόθεσμη κατάσταση και η τυπική απόκλιση των αποδόσεων των μετοχών (std5) των πέντε ετών (2000-2004) προσεγγίζει μακροπρόθεσμη κατάσταση.

2.2. Εκτιμητική μεθοδολογία και αποτελέσματα

2.2.1. Ανάλυση παλινδρόμησης κατά στάδια

Η εκτιμητική διαδικασία, που εφαρμόστηκε, ακολουθεί, σε πρώτη φάση, τα εξής στάδια: Αρχικά, εφαρμόστηκε, ένα υπόδειγμα logit, για την πιθανότητα να υπάρχουν λάθη λογιστικών χειρισμών τα έτη 2000 έως και 2004, με τη χρήση αριθμοδεικτών. Ακολούθως, για το γραμμικό υπόδειγμα εκτιμήθηκαν οι συντελεστές των αντιστοίχων πιθανοτήτων με τη μέθοδο των ελαχίστων τετραγώνων. Ειδικότερα, αφού χρησιμοποιήθηκε η μέθοδος «ανάλυση παλινδρόμησης κατά στάδια», (Drapper and Smith, 1966), βρέθηκαν οι στατιστικά σημαντικότερες ερμηνευτικές μεταβλητές, για τις εξαρτημένες μεταβλητές std5 ή std2.

Το πρώτο υπόδειγμα, έχει τη μορφή:

$$\text{std5} = a_0 + a_1 \text{prob2000} + a_2 \text{prob2001} + a_3 \text{prob2002} + a_4 \text{prob2003} + a_5 \text{prob2004} + \varepsilon_1 \quad (1)$$

Δηλαδή, εξετάζεται η σχέση μεταξύ της τυπικής απόκλισης των αποδόσεων των μετοχών των 5 ετών (std5) και των πιθανοτήτων να εμφανισθούν λάθος λογιστικοί χειρισμοί στα έτη 2000, 2001, 2002, 2003 και 2004 (prob2000, prob2001, prob2002, prob2003, αντίστοιχα). Όπου ε:τυχαίο σφάλμα. Εφαρμόζοντας τη μέθοδο παλινδρόμησης κατά στάδια βρέθηκε ότι η τυπική απόκλιση των αποδόσεων των μετοχών, για την πενταετία, έχει τη στατιστικά σημαντικότερη γραμμική σχέση με την πιθανότητα ύπαρξης λάθους χειρισμών για το έτος 2004 και με εκτίμηση του αντίστοιχου συντελεστή ίση με 0,03752 (για $\alpha=0,01$: $F=33,658 > 7,08$, $t=5,802 > 2,58$).

Το δεύτερο υπόδειγμα, έχει τη μορφή:

$$\text{std2} = b_0 + b_1 \text{prob2000} + b_2 \text{prob2001} + b_3 \text{prob2002} + b_4 \text{prob2003} + b_5 \text{prob2004} + \varepsilon_2 \quad (2)$$

Δηλαδή, εξετάζεται η σχέση μεταξύ της τυπικής απόκλισης των αποδόσεων των μετοχών των 2 πρώτων ετών (std2) και των πιθανοτήτων να εμφανισθούν λάθη λογιστικών χειρισμών στα έτη 2000, 2001, 2002, 2003 και 2004. Εφαρμόζοντας τη διαδικασία παλινδρόμησης κατά στάδια βρέθηκε ότι η τυπική απόκλιση των αποδόσεων των μετοχών, για την πρώτη διετία, έχει τη στατιστικά σημαντικότερη γραμμική σχέση με την πιθανότητα ύπαρξης λάθους λογιστικών χειρισμών για το

έτος 2002 και με εκτίμηση του αντίστοιχου συντελεστή ίση με 0,0537 (για $\alpha=0,01$: $F=44,665>7,08$, $t=6,683>2,58$).

2.2.2. Κατηγοριοποίηση των εταιρειών

Προκειμένου να πραγματοποιηθεί η δεύτερη φάση της εκτιμητικής διαδικασίας επιχειρήθηκε να διαχωρισθούν οι ελληνικές εταιρείες μεταποιητικού κλάδου σε τρεις κατηγορίες: (α) Εταιρείες με μικρές παρατηρήσεις (καλές ως προς τους λογιστικούς χειρισμούς). (β) Εταιρείες με μεσαίες παρατηρήσεις (μέτριες ως προς τους λογιστικούς χειρισμούς). (γ) Εταιρείες με πολλές παρατηρήσεις (κακές ως προς τους λογιστικούς χειρισμούς).

Για να κατηγοριοποιηθούν οι εταιρείες του δείγματος, με βάση τη συμμόρφωσή τους με τις γενικά παραδεκτές λογιστικές αρχές, έγινε χρήση των πιστοποιητικών των ορκωτών ελεγκτών για τις δημοσιευμένες λογιστικές καταστάσεις για τα έτη 2000 - 2004. Πιο συγκεκριμένα, με βάση τις παρατηρήσεις επί των μεγεθών του ισολογισμού, που προέκυψαν από τον έλεγχο, υπολογίσθηκε η επίπτωση που θα είχε η συνεπής τήρηση των λογιστικών αρχών στο σύνολο των ιδίων κεφαλαίων. Ειδικότερα, για να επιτευχθεί ο διαχωρισμός χρησιμοποιήθηκε ο αριθμοδείκτης που ισούται με το λόγο: (Συνολικό ποσό της επίπτωσης των παρατηρήσεων των ορκωτών ελεγκτών στην καθαρή θέση που θα είχε η κατάρτιση των ετήσιων λογιστικών καταστάσεων, με βάση τις γενικές παραδεκτές λογιστικές αρχές), προς το (Σύνολο ιδίων κεφαλαίων, όπως προκύπτει από τους ισολογισμούς). Με βάση τα στοιχεία που αφορούν τις τιμές του αριθμοδείκτη, οι επιχειρήσεις του δείγματος κατηγοριοποιήθηκαν σε τρεις κατηγορίες με χρήση του στατιστικού πακέτου SPSS (Spathis 2002).

Σημειώνεται ότι εφαρμόζοντας την πολλαπλή διαχωριστική ανάλυση, με χρήση κατάλληλων αριθμοδεικτών (Spathis 2002), μπορεί να διαμορφωθούν κανόνες για την περαιτέρω διάκριση και πρόβλεψη των μεταποιητικών εταιρειών των οποίων παραποιοούνται στοιχεία των οικονομικών καταστάσεων (Λεβεντίδης και Δονάτου (2012)).

2.2.3. Ανάλυση συνδιακύμανσης

Σε δεύτερη φάση της εκτιμητικής διαδικασίας, εφαρμόσθηκε η «ανάλυση συνδιακύμανσης», (Johnson, 1974), όπου χρησιμοποιήθηκαν οι προηγούμενες διακριτές μεταβλητές ως προς τη μέση πιθανότητα για λανθασμένους λογιστικούς χειρισμούς στους ισολογισμούς, για τρεις κατηγορίες ελληνικών μεταποιητικών εταιρειών. Η πρώτη κατηγορία περιλαμβάνει 17 εταιρείες καλές ως προς τους λογιστικούς χειρισμούς, η δεύτερη κατηγορία περιλαμβάνει 23 εταιρείες μέτριες ως προς τους λογιστικούς χειρισμούς και η τρίτη κατηγορία περιλαμβάνει 21 εταιρείες κακές ως προς τους λογιστικούς χειρισμούς. Δηλαδή, στην ανάλυση εμφανίζονται τρεις κατηγορίες (στρώματα) και συνολικά μια σειρά 61 τιμών δεδομένων, με 17, 23, 21 τιμές δεδομένων για κάθε κατηγορία, αντίστοιχα.

Αρχικά, εφαρμόστηκε το ακόλουθο υπόδειγμα παλινδρόμησης:

$$Y_{ij} = \beta_0 + \beta_1 X_{ij,1} + \beta_2 X_{ij,2} + \beta_3 X_{ij,3} + \beta_4 X_{ij,4} + \beta_5 X_{ij,5} + \varepsilon_{ij} \quad (3)$$

όπου: Y_{ij} = η j τιμή της εξαρτημένης μεταβλητής στην κατηγορία i

$X_{ij,k}$ = η j τιμή της ερμηνευτικής μεταβλητής k στην κατηγορία i

ε_{ij} = ο διαταρακτικός όρος της j τιμής στην κατηγορία i ,

για $i = 1, 2, 3$, κατηγορίες μεταποιητικών εταιρειών

$j = 1, 2, 3, \dots, 61$ μεταποιητικές εταιρείες

$k = 1, 2, 3, 4, 5$ πιθανότητες για λανθασμένους λογιστικούς χειρισμούς.

Σημειώνεται ότι: (α) Ως εξαρτημένη μεταβλητή χρησιμοποιείται, εναλλάξ, η τυπική απόκλιση των εβδομαδιαίων αποδόσεων των μετοχών $std2$ που αφορά τα έτη 2000 και 2001 και $std5$ που αφορά τα έτη 2000-2004. (β) Ως ερμηνευτικές μεταβλητές X_1, X_2, X_3, X_4, X_5 χρησιμοποιούνται, αντίστοιχα, οι ετήσιες πιθανότητες $prob2000, prob2001, prob2002, prob2003, prob2004$.

Αν υποθέσουμε ότι ο σταθερός όρος διαφέρει από κατηγορία σε κατηγορία, τότε το υπόδειγμα γράφεται ως εξής:

$$Y_{ij} = \beta_0 + \gamma_2 D_{2j} + \gamma_3 D_{3j} + \beta_1 X_{ij,1} + \beta_2 X_{ij,2} + \dots + \beta_5 X_{ij,5} + u_{ij} \quad (4)$$

όπου: $D_{2j} = 1$ αν οι τιμές δεδομένων προέρχονται από την κατηγορία 2

$= 0$ για κάθε άλλη περίπτωση

$D_{3j} = 1$ αν οι τιμές δεδομένων προέρχονται από την κατηγορία 3

$= 0$ για κάθε άλλη περίπτωση.

Αν υποθέσουμε ότι και οι συντελεστές των ερμηνευτικών μεταβλητών διαφέρουν από κατηγορία σε κατηγορία, τότε έχουμε 3 διαφορετικές συναρτήσεις, όσες δηλαδή οι κατηγορίες εταιρειών, και, επομένως, τις ακόλουθες 3 παλινδρομήσεις:

$$\begin{aligned} \text{Κατηγορία 1: } Y_{1p} &= \beta_{10} + \beta_{11} X_{1p,1} + \beta_{12} X_{1p,2} + \dots + \beta_{15} X_{1p,5} + u_{1p} \\ \text{Κατηγορία 2: } Y_{2l} &= \beta_{20} + \beta_{21} X_{2l,1} + \beta_{22} X_{2l,2} + \dots + \beta_{25} X_{2l,5} + u_{2l} \\ \text{Κατηγορία 3: } Y_{3m} &= \beta_{30} + \beta_{31} X_{3m,1} + \beta_{32} X_{3m,2} + \dots + \beta_{35} X_{3m,5} + u_{3m} \end{aligned} \quad (5)$$

για $p = 1, 2, \dots, 17$ $l = 1, 2, \dots, 23$ $m = 1, 2, \dots, 21$.

Τα προαναφερόμενα υποδείγματα εκτιμήθηκαν με τη μέθοδο ελαχίστων τετραγώνων και μετά εφαρμόστηκε ο έλεγχος, με κριτήριο τη στατιστική F , των εξής

υποθέσεων: (α) Έλεγχος της υπόθεσης ότι όλοι οι συντελεστές είναι ίσοι. (β) Έλεγχος της υπόθεσης ότι οι σταθεροί όροι είναι ίσοι. (γ) Έλεγχος της υπόθεσης ότι οι συντελεστές (κλίσεις) των ερμηνευτικών μεταβλητών δεν διαφέρουν από κατηγορία σε κατηγορία.

Με βάση τα αποτελέσματα της ανάλυσης συνδιακύμανσης, που προέκυψαν από την εκτίμηση των αντίστοιχων υποδειγμάτων, για την εξαρτημένη μεταβλητή std2, η εφαρμογή σε επίπεδο στατιστικής σημαντικότητας $\alpha = 0,01$ του αντίστοιχου στατιστικού κριτηρίου F:

(i) Τόσο για τον έλεγχο ισότητας των συντελεστών των ερμηνευτικών μεταβλητών ($F = 4,77 > F_{0,01, 10, 165} = 2,43$) όσο και για τον έλεγχο ισότητας όλων των συντελεστών ($F = 4,77 > F_{0,01, 12, 165} = 2,29$), έδειξε ότι υπάρχει στατιστικά σημαντική διαφορά για τις τρεις κατηγορίες των εταιρειών.

(ii) Για τον έλεγχο ισότητας των σταθερών όρων ($F = 0,14 < F_{0,01, 2, 175} = 4,7$), έδειξε ότι δεν υπάρχει στατιστικά σημαντική διαφορά μεταξύ των σταθερών όρων για τις τρεις κατηγορίες των εταιρειών.

Στη συνέχεια, σύμφωνα με τα στοιχεία της ανάλυσης συνδιακύμανσης των στοιχείων, που προέκυψαν από την εκτίμηση των αντίστοιχων υποδειγμάτων, που έχουν ως εξαρτημένη μεταβλητή την std5, η εφαρμογή σε επίπεδο στατιστικής σημαντικότητας $\alpha=0,01$ του αντίστοιχου στατιστικού κριτηρίου F:

(i) Τόσο για τον έλεγχο ισότητας των συντελεστών των ερμηνευτικών μεταβλητών ($F = 5,23 > F_{0,01, 10, 165} = 2,43$) όσο και για τον έλεγχο ισότητας όλων των συντελεστών ($F = 6,31 > F_{0,01, 12, 165} = 2,29$), έδειξε ότι υπάρχει στατιστικά σημαντική διαφορά για τις τρεις κατηγορίες των εταιρειών.

(ii) Για τον έλεγχο ισότητας των σταθερών όρων ($F = 0,01 < F_{0,01, 2, 175} = 4,73$), έδειξε ότι δεν υπάρχει στατιστικά σημαντική διαφορά μεταξύ των σταθερών όρων για τις τρεις κατηγορίες των εταιρειών.

Για κάθε παλινδρόμηση χρησιμοποιήθηκε ο έλεγχος κανονικότητας κατά Jarque-Bera. Ειδικότερα το μέτρο που εφαρμόστηκε για τον έλεγχο κανονικότητας είναι:

$$JB = \frac{N-k}{6} \left(Skew^2 + \frac{1}{4} \cdot (Kurt - 3)^2 \right) \quad (6)$$

όπου: N ο αριθμός των παρατηρήσεων, k ο αριθμός των ανεξάρτητων μεταβλητών που χρησιμοποιήθηκαν για την εξαγωγή της σειράς, Skew η ασυμμετρία και Kurt η κύρτωση της σειράς, αντιστοίχως. Κάτω από την μηδενική υπόθεση της κανονικής κατανομής το μέτρο JB ακολουθεί την χ^2 κατανομή με δύο βαθμούς ελευθερίας.

Στους ακόλουθους Πίνακες 1 και 2 δίνονται τα αποτελέσματα του ελέγχου κανονικότητας των σφαλμάτων των υποδειγμάτων που εκτιμήθηκαν και χρησιμοποιήθηκαν για τη διενέργεια της ανάλυσης συνδιακύμανσης των τριών κατηγοριών εταιρειών:

Πίνακας 1 . Έλεγχος κανονικότητας σφαλμάτων υποδειγμάτων με εξαρτημένη μεταβλητή sd2

Υπόδειγμα	R ²	Έλεγχος κανονικότητας σφαλμάτων
(3)	0,50	JB-pvalue = 0.053,σφάλματα κανονικά
(4)	0,50	JB-pvalue = 0.058,σφάλματα κανονικά
(5) για την κατηγορία 1	0,73	JB-pvalue = 0.476,σφάλματα κανονικά
(5) για την κατηγορία 2	0,43	JB-pvalue = 0.707,σφάλματα κανονικά
(5) για την κατηγορία 3	0,69	JB-pvalue = 0.780,σφάλματα κανονικά

Πίνακας 2 . Έλεγχος κανονικότητας σφαλμάτων υποδειγμάτων με εξαρτημένη μεταβλητή sd5

Υπόδειγμα	R ²	Έλεγχος κανονικότητας σφαλμάτων
(3)	0,41	JB-pvalue = 0.849,σφάλματα κανονικά
(4)	0,41	JB-pvalue = 0.847,σφάλματα κανονικά
(5) για την κατηγορία 1	0,64	JB-pvalue = 0.684,σφάλματα κανονικά
(5) για την κατηγορία 2	0,38	JB-pvalue = 0.889,σφάλματα κανονικά
(5) για την κατηγορία 3	0,75	JB-pvalue = 0.585,σφάλματα κανονικά

Στον ακόλουθο Πίνακα 3 δίνονται τα αποτελέσματα του ελέγχου αν οι ανεξάρτητες μεταβλητές (prob2000, prob2001, prob2002, prob2003, prob2004) ακολουθούν την κανονική κατανομή.

Πίνακας 3 . Έλεγχος κανονικότητας ανεξαρτήτων μεταβλητών

Ανεξάρτητη μεταβλητή	Έλεγχος κανονικότητας
Prob2000	JB-pvalue = 0.000, δεν ακολουθεί την κανονική κατανομή
Prob2001	JB-pvalue = 0.055, ακολουθεί την κανονική κατανομή
Prob2002	JB-pvalue = 0.058, ακολουθεί την κανονική κατανομή
Prob2003	JB-pvalue = 0.109, ακολουθεί την κανονική κατανομή
Prob2004	JB-pvalue = 0.061, ακολουθεί την κανονική κατανομή

Σημείωση: Για το κρίσιμο έτος 2000 θα πρέπει να ληφθεί υπόψη ο προβληματισμός που υπάρχει για την ποιότητα των οικονομικών στοιχείων. Όλες οι ανεξάρτητες μεταβλητές ακολουθούν ικανοποιητικά την κανονική κατανομή για επίπεδο στατιστικής σημαντικότητας 1%.

Σύμφωνα με την ανάλυση συνδιακύμανσης, που εφαρμόστηκε για τη χρονική περίοδο 2000-2004 και για επίπεδο στατιστικής σημαντικότητας 1%, συνάγονται, σε βραχυχρόνια και μακροχρόνια βάση, τα εξής:

(i) Ο παράγων που αφορά την πιθανότητα ύπαρξης λάθους λογιστικών χειρισμών ερμηνεύει συνολικά τη μεταβλητότητα των μετοχών των ελληνικών εταιρειών του μεταποιητικού κλάδου για βραχυχρόνια και μακροχρόνια περίοδο και διαφορετικά για κάθε μία από τις τρεις κατηγορίες εταιρειών του κλάδου (καλές, μέτριες και κακές ως προς τους λογιστικούς χειρισμούς).

(ii) Η συνάρτηση μεταβλητότητας της απόδοσης των μετοχών έδειξε ότι παραμένει σταθερή σε σχέση με τις τρεις κατηγορίες των εταιρειών του κλάδου.

3. ΣΥΜΠΕΡΑΣΜΑΤΑ

Η σημασία της πληροφόρησης για την «πορεία» των επιχειρήσεων εξηγεί, εν μέρει, την ύπαρξη λογιστικών χειρισμών στις οικονομικές καταστάσεις ορισμένων επιχειρήσεων προς την κατεύθυνση της «ωραιοποίησης» των αποτελεσμάτων και της «καθαρής» θέσης τους. Οι χειρισμοί αυτοί μπορεί να έχουν επίδραση στην αναμενόμενη κερδοφορία ή στην αποτύπωση των κεφαλαίων ή και στα δύο. Προβλέψεις της κερδοφορίας για τα επόμενα έτη, σε συνδυασμό με άλλες παραμέτρους χρησιμοποιούνται, ευρέως, για τον υπολογισμό της εσωτερικής αξίας της μετοχής. Ένα, λοιπόν, ενδιαφέρον προς εξέταση ερώτημα είναι: Κατά πόσο η ύπαρξη λογιστικών χειρισμών, όπως περιγράφεται στις σημειώσεις των ορκωτών λογιστών, επηρεάζει τις αποδόσεις των μετοχών ή επηρεάζεται από τις προηγούμενες αποδόσεις τους.

Στην εργασία αυτή η εφαρμογή κατάλληλης στατιστικής ανάλυσης βοήθησε στην επίτευξη ικανοποιητικών, ως προς την πληροφόρηση, και αξιόπιστων, ως προς τη στατιστική σημαντικότητα, ευρημάτων που αφορούν τη σχέση της ποιότητας των συστημάτων και των διαδικασιών ελέγχου και της αποδοτικότητας των μετοχών ελληνικών εταιριών του μεταποιητικού κλάδου.

Αρχικά, η ανάλυση που πραγματοποιήθηκε έδειξε ότι, με πιθανότητα 99%, η πιθανότητα ύπαρξης λάθους λογιστικών χειρισμών, για την υπό εξέταση χρονική περίοδο και κυρίως για τα έτη 2002 και 2004, ερμηνεύει συνολικά τη μεταβλητότητα αποδοτικότητας των μετοχών των εταιρειών του μεταποιητικού κλάδου, αντίστοιχα, σε βραχυχρόνια και μακροχρόνια περίοδο. Ακολούθως, για την περίοδο που εξετάστηκε και σε βραχυχρόνια και σε μακροχρόνια βάση, βρέθηκε ότι, με πιθανότητα 99%, η πιθανότητα κακής ποιότητας λογιστικών χειρισμών, για τα έτη 2000-2004, ερμηνεύει τη μεταβλητότητα των αποδόσεων των μετοχών διαφορετικά για κάθε μία από τις τρεις κατηγορίες εταιρειών, που ανέδειξε η στατιστική διαδικασία (καλές, μέτριες και κακές ως προς τους λογιστικούς χειρισμούς). Αντίθετα, η συνάρτηση μεταβλητότητας αποδόσεως των μετοχών έδειξε ότι είναι σταθερή, σχετικά με την κατηγοριοποίηση των εταιρειών.

ABSTRACT

The purpose of this paper is the examination of the relationship between the operational quality of audit systems and of procedures of the efficiency of big greek manufacturing enterprises, in the short-term as well as in the long-term. In addition to that, the authors investigated the effect on the volatility of shares, of accounting malpractice that alters the reliability of the balance sheet. This effect was examined for the total of number of listed manufacturing companies, as well as for specific subgroups of the sector according to the quality of their accounting practices (good, medium and poor).

ΑΝΑΦΟΡΕΣ

- Aldhizer G. R., Miller J. R., and Moraglio J. F., (1995). Common Attributes of Quality Audits, *Journal of Accountancy*, **179**, 61-68.
- Ball R., Kothari S.P., Robin A., (2000). The Effect of International Institutional Factors on Properties of Accounting Earnings, *Journal of Accounting and Economics*, **29**, 1-51.
- Draper N. and Smith H., (1998). *Applied Regression Analysis*, New York, Third Edition, Wiley-Interscience.
- Johnston J., (1984). *Econometric Methods*, New York, Third edition, McGraw-Hill College.
- Krishman J. and Schauer P.C., (2000). The Differentiation of Quality Among Auditors: Evidence from the Not-for-Profit Sector, *Auditing: A Journal of Practice and Theory*, **19**, 9-25.
- Λεβεντίδης Ι. και Δονάτου Α., (2012). Η Εφαρμογή της Πολλαπλής Διαχωριστικής Ανάλυσης για την Πρόβλεψη των Εταιρειών ενός Κλάδου των οποίων Παραποιοούνται Στοιχεία των Οικονομικών Καταστάσεων, Τιμητικός Τόμος Πέτρου Γέμτου, Εκδόσεις Α. Σάκκουλα.
- McConnell D. K. Jr. and Banks G. Y., (1998). A Common Peer Review Problem, *Journal of Accountancy*, **185(6)**, 39-44.
- Merton R.C., (1974). On the Pricing of Corporate Debt, the Risk Structure of Interest Rates, *Journal of Finance*, **29**, 449-470.
- Spathis C.T., (2002). Detecting False Financial Statements Using Published Data: Some Evidence from Greece, *Managerial Auditing Journal*, **17(4)**, 179-191.

ΑΙΤΙΑΚΕΣ ΣΧΕΣΕΙΣ ΜΕΤΑΞΥ ΤΟΥ ΕΛΛΕΙΜΜΑΤΟΣ ΙΣΟΖΥΓΙΟΥ ΤΡΕΧΟΥΣΩΝ ΣΥΝΑΛΛΑΓΩΝ ΚΑΙ ΤΩΝ ΔΗΜΟΣΙΟΝΟΜΙΚΩΝ ΕΛΛΕΙΜΜΑΤΩΝ: Μια Εμπειρική Έρευνα για την Ελλάδα

Νίκος Δριτσάκης

Τμήμα Εφαρμοσμένης Πληροφορικής, Πανεπιστήμιο Μακεδονίας
drits@uom.gr

ΠΕΡΙΛΗΨΗ

Στην εργασία αυτή ερευνούμε την αιτιακή σχέση ανάμεσα στο έλλειμμα του κρατικού προϋπολογισμού και το ισοζύγιο τρεχουσών συναλλαγών για την Ελλάδα την περίοδο 1995 - 2011. Για τη διερεύνηση της σχέσης αυτής χρησιμοποιούμε τη μεθοδολογία του επαυξημένου αυτοπαλίνδρομου διανύσματος που αναπτύχθηκε από τους Toda and Yamamoto (1995). Τα εμπειρικά αποτελέσματα έδειξαν ότι δεν υπάρχουν αιτιακές σχέσεις μεταξύ των μεταβλητών που εξετάζουμε.

Λέξεις Κλειδιά: Έλλειμμα του κρατικού προϋπολογισμού, Έλλειμμα ισοζυγίου τρεχουσών συναλλαγών, Συνολοκλήρωση, Έλεγχος αιτιότητας των Toda and Yamamoto.

JEL classification: C12, C23, F15

1. ΕΙΣΑΓΩΓΗ

Η υπόθεση των διδύμων ελλειμμάτων ισχυρίζεται ότι η αύξηση του ελλείμματος του κρατικού προϋπολογισμού θα προκαλέσει αύξηση του ισοζυγίου τρεχουσών συναλλαγών και αντίστροφα. Σύμφωνα με την κεϋνσιανή θεωρία σε μια ανοικτή οικονομία, μια αύξηση του ελλείμματος του κρατικού προϋπολογισμού θα προκαλέσει αύξηση της συνολικής ζήτησης και των πραγματικών επιτοκίων. Τα υψηλά επιτόκια θα οδηγήσουν σε εισροή κεφαλαίων με αποτέλεσμα την ανατίμηση του εγχώριου νομίσματος. Υψηλότερη τιμή του εγχώριου νομίσματος θα επηρεάσει αρνητικά τις εξαγωγές, και επομένως θα έχουμε επιδείνωση στο έλλειμμα τρεχουσών συναλλαγών.

Μια χώρα με ένα έλλειμμα τρεχουσών συναλλαγών θα δανειστεί πόρους από τον υπόλοιπο κόσμο που θα πρέπει να επιστρέψει στο μέλλον. Εδώ θα πρέπει να σημειώσουμε ότι ένα έλλειμμα τρεχουσών συναλλαγών δεν είναι κατ' ανάγκη αρνητικό φαινόμενο για μια χώρα με οικονομική ανάπτυξη. Αν η χώρα με τους

πόρους που θα δανειστεί δημιουργήσει μια κερδοφόρα επένδυση υψηλής απόδοσης που να καλύπτει τα κεφάλαια και τους τόκους των δανείων, τότε η χώρα θα επωφεληθεί από το χρέος της στο μέλλον. Από την άλλη πλευρά, αν το έλλειμμα τρεχουσών συναλλαγών δημιουργείται με σκοπό την αύξηση της κατανάλωσης, χωρίς καμία βελτίωση του μετοχικού κεφαλαίου, θα προκαλέσει στη χώρα δυσκολία στο μέλλον να εξοφλήσει το χρέος της.

Μεγάλα ελλείμματα στον κρατικό προϋπολογισμό και στο ισοζύγιο τρεχουσών συναλλαγών δεν είναι νέο φαινόμενο. Οι πιο ανεπτυγμένες χώρες όπως οι ΗΠΑ στη δεκαετία του 1980 έχει βιώσει μια ισχυρή και ταυτόχρονη διεύρυνση του ελλείμματος τρεχουσών συναλλαγών, καθώς και του κρατικού προϋπολογισμού.

Ο Laney (1984) εξετάζει την αιτιακή σχέση ανάμεσα στα δίδυμα ελλείμματα για τις ανεπτυγμένες και τις αναπτυσσόμενες χώρες, και το αποτέλεσμα της έρευνας του δείχνει ότι η σχέση μεταξύ αυτών των δύο μεταβλητών είναι πολύ ισχυρότερη στις αναπτυσσόμενες χώρες. Παρά το γεγονός ότι έχουν υπάρξει πολλές εμπειρικές έρευνες για τη σχέση μεταξύ του ελλείμματος του προϋπολογισμού και του ελλείμματος ισοζυγίου τρεχουσών συναλλαγών, στις ανεπτυγμένες οικονομίες, έχει δοθεί λίγη προσοχή στις υπό ανάπτυξη οικονομίες.

Στη εργασία αυτή, εξετάζουμε εμπειρικά τη δυναμική μακροχρόνια αιτιακή σχέση μεταξύ του προϋπολογισμού και του ισοζυγίου τρεχουσών συναλλαγών σε μια μικρή ανοικτή οικονομία της Ελλάδος. Για το σκοπό αυτό, χρησιμοποιήσουμε τον έλεγχο συνολοκλήρωσης μεγίστης πιθανοφάνειας των Johansen και Juselius (1990) και τον τροποποιημένο έλεγχο αιτιότητας που πρότειναν οι Toda και Yamamoto (1995).

Η εργασία αυτή είναι οργανωμένη σε πέντε τμήματα. Το τμήμα 2 παρουσιάζει τη θεωρία για την υπόθεση των δίδυμων ελλειμμάτων. Το τμήμα 3 παρουσιάζει την ανασκόπηση της βιβλιογραφίας. Η παράγραφος 4 καθορίζει τυπικά το αναλυτικό πλαίσιο και την οικονομετρική μεθοδολογία. Τα εμπειρικά αποτελέσματα παρουσιάζονται στο τμήμα 5, και τέλος τα καταληκτικά σχόλια που περιέχονται στο τμήμα 6.

2. Η ΘΕΩΡΙΑ ΓΙΑ ΤΗΝ ΥΠΟΘΕΣΗ ΤΩΝ ΔΙΔΥΜΩΝ ΕΛΛΕΙΜΜΑΤΩΝ

Σύμφωνα με τη θεωρία των «διδύμων ελλειμμάτων» (ελλείμματα του δημοσίου και ελλείμματα τρεχουσών συναλλαγών) τα δημοσιονομικά ελλείμματα προκαλούν τα ελλείμματα τρεχουσών συναλλαγών και αντίστροφα. Για να διευκρινίσουμε τη σχέση αυτή ξεκινάμε από την ταυτότητα του εθνικού εισοδήματος σε μια ανοικτή οικονομία.

Σε μια ανοικτή οικονομία το εθνικό εισόδημα δίνεται από την παρακάτω ταυτότητα:

$$Y = C + I + G + X - IM \quad (1)$$

όπου:

Y = Το εθνικό εισόδημα.

C = Η ιδιωτική κατανάλωση

I = Η πραγματική επένδυση

G = Κρατικές δαπάνες

X = Εξαγωγές

IM = Εισαγωγές

Η εξίσωση τρεχουσών συναλλαγών (CAD) ορίζεται από τη διαφορά μεταξύ εξαγωγών (X) και εισαγωγών (IM) και μπορεί να παρασταθεί ως εξής:

$$CAD = X - IM \quad (2)$$

Σύμφωνα με την ταυτότητα (1) η εξίσωση (2) μπορεί να γραφεί ως εξής:

$$CAD = Y - (C + I + G) \quad (3)$$

όπου ($C + I + G$) είναι οι εσωτερικές δαπάνες μιας χώρας. Σε μια κλειστή οικονομία (χωρίς διεθνές εμπόριο) η αποταμίευση (S) είναι ίση με την επένδυση (I). Σε μια όμως ανοιχτή οικονομία η αποταμίευση μπορεί να οριστεί ως εξής:

$$S = I + CAD \quad (4)$$

Από την εξίσωση (4) γνωρίζουμε ότι σε μια ελεύθερη οικονομία, κάθε χώρα αναζητά κεφάλαια για επενδύσεις τόσο σε τοπικό όσο και σε διεθνές επίπεδο προκειμένου να αυξηθεί το εισόδημα. Η εθνική αποταμίευση μπορεί να χωριστεί σε δύο μέρη την ιδιωτική και την κρατική αποταμίευση. Η ιδιωτική αποταμίευση, συμβολίζεται με S_p και η κρατική με S_g και δίνονται από τις παρακάτω εξισώσεις:

$$\begin{aligned} S_p &= Y - T - C \\ (5) \quad S_g &= T - G \end{aligned} \quad (6)$$

όπου T είναι τα φορολογικά έσοδα.

Η εξίσωση (5) με βάση την εξίσωση (3) μπορεί να διαμορφωθεί ως εξής:

$$S_p = CAD + I + G - T \quad (7)$$

$$\text{ή}$$
$$CAD = S_p - I - (G - T)$$

$$\text{ή}$$
$$CAD = S_p - I - BD \quad (8)$$

Από την εξίσωση (8) το δημόσιο έλλειμμα του προϋπολογισμού (BD) ορίζεται ως η διαφορά $G - T$ έχει αρνητικό πρόσημο και είναι ανεξάρτητο από τις κρατικές αποταμιεύσεις. Σύμφωνα, με την κεϋνσιανή θεωρία, από την εξίσωση (8) μπορούμε επίσης να δούμε ότι μια αύξηση στο δημόσιο έλλειμμα του προϋπολογισμού (BD) θα προκαλέσει αύξηση στο έλλειμμα τρεχουσών συναλλαγών (CAD) στη περίπτωση που οι ιδιωτικές αποταμιεύσεις (S_p) και οι επενδύσεις (I) παραμένουν σταθερές.

Αντιθέτως, ο Summers (1988) υποστήριξε ότι μια αντίθετη αιτιακή σχέση μπορεί να υπάρχει από το έλλειμμα τρεχουσών συναλλαγών (CAD) προς το δημόσιο έλλειμμα του προϋπολογισμού (BD), γεγονός που οδηγεί σε βραδύτερους ρυθμούς την οικονομική ανάπτυξη και στη συνέχεια αυξάνει το δημόσιο έλλειμμα του προϋπολογισμού (BD).

Σύμφωνα με τη θεωρία του Ρικάρντο όταν μια κυβέρνηση μειώσει τους φόρους θα αυξήσει το δημόσιο έλλειμμα (BD). Η αύξηση του δημοσίου ελλείμματος

θα επιφέρει στη συνέχεια υψηλότερους φόρους για την αποπληρωμή του χρέους. Η αύξηση όμως του φόρου στους πολίτες θα μειώσει τις δαπάνες τους για κατανάλωση και θα αυξήσει την ιδιωτική αποταμίευση σε βάρος της κρατικής αποταμίευσης, οπότε το δημοσιονομικό έλλειμμα (BD) δεν θα έχει καμία επίδραση στο έλλειμμα τρεχουσών συναλλαγών (CAD).

3. ΒΙΒΛΙΟΓΡΑΦΙΚΗ ΑΝΑΣΚΟΠΗΣΗ

Η βιβλιογραφία με το θέμα των διδύμων ελλειμμάτων επικεντρώνεται κυρίως σε δύο θεωρητικές προσεγγίσεις, την κεϋνσιανή και την ρικαρδιανή υπόθεση ισοδυναμίας. Ωστόσο, η «υπόθεση των διδύμων ελλειμμάτων» βρίσκεται κάτω από τις τέσσερις παρακάτω υποθέσεις.

Σύμφωνα με την πρώτη υπόθεση ο Barro (1974) υποστηρίζει ότι δεν υπάρχει συσχετισμός μεταξύ του ελλείμματος του κρατικού προϋπολογισμού και του ελλείμματος τρεχουσών συναλλαγών. Αυτό μπορεί να γίνει κατανοητό με τη μείωση της δημόσιας αποταμίευσης λόγω του μεγάλου δημοσιονομικού ελλείμματος που θα συνδυάζεται με ίση αύξηση των ιδιωτικών καταθέσεων (Barro, 1989). Δηλαδή οι καταναλωτές σκέπτονται ότι αν ένας φόρος μειωθεί ή καταργηθεί σήμερα θα οδηγήσει σε μεγαλύτερο δημοσιονομικό έλλειμμα το οποίο θα οδηγήσει σε μελλοντικές αυξήσεις των φόρων για την εξυπηρέτηση του δημόσιου χρέους. Με τον τρόπο αυτό εξοικονομούμε χρήματα σήμερα και τα πληρώνουμε στο μέλλον.

Η δεύτερη υπόθεση που υποστηρίζεται από τον Laney (1984), αναφέρει ότι υπάρχει μονόδρομη σχέση αιτιότητας από το έλλειμμα του προϋπολογισμού, προς το έλλειμμα τρεχουσών συναλλαγών. Άλλες μελέτες που τηρείται η κεϋνσιανή πρόταση στην οποία αύξηση του ελλείμματος του προϋπολογισμού οδηγεί σε μια επιδείνωση του ελλείμματος τρεχουσών συναλλαγών περιλαμβάνουν τις εργασίες των Ahmed (1986), Abell (1990), και Vamvoukas (1997),

Η τρίτη υπόθεση υποστηρίζει ότι υπάρχει μονόδρομη σχέση αιτιότητας από το έλλειμμα τρεχουσών συναλλαγών προς το έλλειμμα του προϋπολογισμού. Αυτό το αποτέλεσμα εμφανίζεται όταν η επιδείνωση των τρεχουσών συναλλαγών οδηγεί σε έναν πιο αργό ρυθμό την οικονομική ανάπτυξη και ως εκ τούτου αυξάνει το έλλειμμα του προϋπολογισμού. Αυτό είναι ιδιαίτερα εμφανές σε μια μικρή ανοικτή οικονομία που εξαρτάται από ξένες εισροές κεφαλαίων (ξένες άμεσες επενδύσεις) που χρηματοδοτούν την οικονομική ανάπτυξη. Με άλλα λόγια, η δημοσιονομική κατάσταση μιας χώρας θα επηρεαστεί από τις μεγάλες εισροές κεφαλαίων ή μέσω συσσώρευσης του χρέους με αποτέλεσμα η χώρα να λειτουργήσει με έναν ελλειμματικό προϋπολογισμό. Τα άρθρα των Anoruo και Ramchander (1998), Khalid και Teo (1999), και Hatemi και Shukur (2002), υποστηρίζουν την υπόθεση αυτή.

Τέλος, ερευνητές όπως ο Darrat (1988), ο Islam (1998), και ο Normadin (1999), στις εμπειρικές μελέτες τους υποστηρίζουν την αμφίδρομη σχέση μεταξύ των διδύμων ελλειμμάτων.

4. ΑΝΑΛΥΤΙΚΟ ΠΛΑΙΣΙΟ ΚΑΙ ΟΙΚΟΝΟΜΕΤΡΙΚΗ ΜΕΘΟΔΟΛΟΓΙΑ

Στην ενότητα αυτή περιγράφουμε το οικονομετρικό υπόδειγμα, καθώς και τα δεδομένα που χρησιμοποιούμε στην εργασία, που καλύπτουν μια περίοδο 17 ετών (1995-2011). Τα δεδομένα αυτά περιλαμβάνουν το δημοσιονομικό έλλειμμα (BD) που προκύπτει αφαιρώντας τις δημόσιες δαπάνες από τα δημόσια έσοδα, και το έλλειμμα τρεχουσών συναλλαγών (CAD) που προκύπτει αφαιρώντας τις συνολικές εισαγωγές από συνολικές εξαγωγικές. Και οι δύο μεταβλητές BD και CAD εκφράζεται ως ποσοστό του ονομαστικού ΑΕΠ. Οι παραπάνω μεταβλητές έχουν συνταχθεί από το Διεθνές Νομισματικό Ταμείο (ΔΝΤ) και OECD.

Για τη σχέση ανάμεσα στο δημοσιονομικό έλλειμμα και στο έλλειμμα τρεχουσών συναλλαγών, στην περίπτωση της Ελλάδας, χρησιμοποιούμε τους παρακάτω ελέγχους:

- Έλεγχος μοναδιαίας ρίζας των χρονικών σειρών.
- Έλεγχος της συνολοκλήρωσης για την μακροχρόνια σχέση μεταξύ των μεταβλητών.
- Έλεγχος της αιτιακής σχέσης ανάμεσα στις μεταβλητές.

4.1 ΕΛΕΓΧΟΣ ΜΟΝΑΔΙΑΙΑΣ ΡΙΖΑΣ (ΤΑΞΗ ΟΛΟΚΛΗΡΩΣΗΣ)

Για τον έλεγχο της μοναδιαίας ρίζας των μεταβλητών χρησιμοποιούμε τον επαυξημένο έλεγχο των Dickey-Fuller (ADF) (1979), τον έλεγχο των Phillips-Perron (PP) (1988), καθώς και τον έλεγχο των Kwiatkowski, Philips, Schmidt and Shin (KPSS) (1992).

4.2 ΕΛΕΓΧΟΣ ΤΗΣ ΣΥΝΟΛΟΚΛΗΡΩΣΗΣ - Η ΠΡΟΣΕΓΓΙΣΗ ΤΩΝ JOHANSEN - JUSELIUS

Αφού καθοριστεί η τάξη ολοκλήρωσης των χρονικών σειρών, κάνουμε τον έλεγχο της συνολοκλήρωσης χρησιμοποιώντας την τεχνική των Johansen και Juselius (1990). Οι ερευνητές αυτοί έχουν αναπτύξει δύο στατιστικούς ελέγχους για να προσδιορίσουν τον αριθμό των συνολοκληρωμένων διανυσμάτων, τη μέγιστη ιδιοτιμή και το ίχνος. Το μήκος της βέλτιστης υστέρησης του δυναμικού VAR υποδείγματος προσδιορίζεται από τον τροποποιημένο στατιστικό έλεγχο (LR), το τελικό σφάλμα της πρόβλεψης (FPE), καθώς και τα κριτήρια των Akaike (AIC), Schwarz (SC), και Hannan-Quinn (HQ).

4.3 Η ΔΙΑΔΙΚΑΣΙΑ ΤΩΝ TODA AND YAMAMOTO ΣΤΑ ΑΥΤΟΠΑΛΙΝΔΡΟ-ΜΑ ΔΙΑΝΥΣΜΑΤΑ

Προκειμένου να εκτιμηθεί η αιτιακή σχέση ανάμεσα στο δημοσιονομικό έλλειμμα και στο έλλειμμα τρεχουσών συναλλαγών, χρησιμοποιούμε τη μέθοδο των Toda-Yamamoto (1995), που είναι μια επαυξημένη αιτιακή μέθοδος του Granger

(1969). Η προσέγγιση των Toda και Yamamoto είναι ένας εναλλακτικός έλεγχος αιτιότητας που βασίζεται στην εξίσωση της μη αιτιότητας του Granger αλλά με επιπλέον προσδιοριστικές υστερήσεις που καθορίζονται από την τάξη ολοκλήρωσης των χρονικών σειρών που ελέγχουμε στην αιτιότητα.

Οι Toda και Yamamoto για τον έλεγχο της μη αιτιότητας του Granger στις χρονικές σειρές προτείνουν τη χρήση ενός τροποποιημένου ελέγχου του Wald του MWald. Ο έλεγχος αυτός μάς επιτρέπει να έχουμε συμπεράσματα από τις αιτιακές σχέσεις ενός επαυξημένου συστήματος VAR ανεξάρτητα από το αν οι μεταβλητές είναι ολοκληρωμένες της ίδιας τάξης ή αν υπάρχουν ή όχι συνολοκληρωμένες σχέσεις (αν υπάρχουν συνολοκληρωμένα διανύσματα). Αρκεί η τάξη ολοκλήρωσης κάποιων από τις μεταβλητές να μην είναι μεγαλύτερες από τις χρονικές υστερήσεις του υποδείγματος VAR.

Ο έλεγχος MWald ακολουθεί ασυμπτωτικά την κατανομή X^2 όταν εκτιμούμε το επαυξημένο σύστημα VAR ($k + d_{\max}$), όπου $d_{\max}$, είναι η μέγιστη τάξη ολοκλήρωσης. Σύμφωνα με τους Rambaldi και Doran (1996) οι έλεγχοι MWald στους ελέγχους αιτιότητας του Granger αυξάνουν την αποτελεσματικότητά τους όταν εκτιμούμε τα υποδείγματα με τη μέθοδο των φαινομενικά ασυσχέτιστων παλινδρομήσεων (SUR). Ο έλεγχος των Toda and Yamamoto για μη αιτιότητα του Granger που χρησιμοποιείται στην εργασία αυτή με την εκτίμηση του παρακάτω διμεταβλητού συστήματος VAR και τη μέθοδο SUR δίνεται από τις παρακάτω εξισώσεις:

$$BD_t = \alpha_0 + \sum_{i=1}^{k+d} \alpha_{1i} BD_{t-i} + \sum_{j=1}^{k+d} \alpha_{2j} CAD_{t-j} + e_t \quad (9)$$

$$CAD_t = \beta_0 + \sum_{i=1}^{k+d} \beta_{1i} CAD_{t-i} + \sum_{j=1}^{k+d} \beta_{2j} BD_{t-j} + u_t \quad (10)$$

Η βέλτιστη τάξη υστέρησης είναι k , και d είναι η μέγιστη τάξη ολοκλήρωσης των μεταβλητών στο σύστημα ($d_{\max}$), e_t και u_t είναι οι όροι σφάλματος που υποθέτουμε ότι είναι λευκοί θόρυβοι. Κάθε μεταβλητή παλινδρομείται από την άλλη μεταβλητή με χρονικές υστερήσεις από 1 μέχρι $k + d_{\max}$, χρησιμοποιώντας τη μέθοδο SUR.

Από την εξίσωση (9) εξετάζουμε αν η μεταβλητή CAD δεν αιτιάζεται κατά Granger τη μεταβλητή BD. Η μηδενική υπόθεση που θα δημιουργηθεί θα είναι: $H_0 : \alpha_{2j} = 0$ έναντι της εναλλακτικής $H_0 : \alpha_{2j} \neq 0$ όπου $j \leq k$

Ομοίως, από την εξίσωση (10), εξετάζουμε αν η μεταβλητή BD δεν αιτιάζεται κατά Granger τη μεταβλητή CAD. Η μηδενική υπόθεση που θα δημιουργηθεί θα είναι: $H_0 : \beta_{2j} = 0$ έναντι της εναλλακτικής $H_0 : \beta_{2j} \neq 0$ όπου $j \leq k$

5. ΕΜΠΕΙΡΙΚΑ ΑΠΟΤΕΛΕΣΜΑΤΑ

Το πρώτο βήμα στην εργασία αυτή είναι να καθορίσουμε το βαθμό ολοκλήρωσης της κάθε μεταβλητής. Για το σκοπό αυτό χρησιμοποιούμε ελέγχους που προτείνονται από τους Dickey and Fuller (1979, 1981), ADF, τους Phillips και

Perron (1988) PP, καθώς και από τους Kwiatkowski, Phillips, Schmidt και Shin (1992), KPSS. Ο Πίνακας 1 παρουσιάζει τα αποτελέσματα από τους ελέγχους των ADF, PP, και KPSS για όλες τις μεταβλητές.

Πίνακας 1: Έλεγχοι μοναδιαίας ρίζας

Variables	ADF		PP		KPSS	
	Const	Const, Tr.	Const	Con, Tr.	Const	Const, Tr
Level						
BD	-1.504(0)	-2.713(1)*	-1.610[1]	-3.308[8]	0.357[3]*	0.156[5]**
CAD	-3.802(2)*	-1.855(3)	-1.490[0]	-1.245[0]	0.473[1]**	0.129[1]*
First Differences						
ΔBD	-2.742(1)*	-3.092(1)	-3.835[5]**	-3.729[7]	0.185[3]	0.186[5]*
ΔCAD	-4.239(2)**	-4.69(2)**	-2.825[1]*	-2.681[2]*	0.258[0]	0.055[1]

Σημειώσεις: 1. ***, **, * για επίπεδα σημαντικότητας 1, 5 και 10% αντίστοιχα. 2. Οι αριθμοί μέσα στις παρενθέσεις για το ADF, αντιπροσωπεύουν τον αριθμό των χρονικών υστερήσεων της εξαρτημένης μεταβλητής που χρησιμοποιείται για τα σφάλματα του λευκού θορύβου. 3. Ο αριθμός των χρονικών υστερήσεων για το ADF επιλέχτηκε χρησιμοποιώντας το κριτήριο του Akaike (AIC). 4. Οι αριθμοί μέσα στις αγκύλες για τα PP και KPSS αναφέρονται στον αριθμό των περιόδων της αυτοσυσχέτισης βασισμένο στον εκτιμητή των Newey-West (1994) χρησιμοποιώντας τη στατιστική συνάρτηση του Bartlett. 5. Δ=Πρώτες διαφορές

Τα αποτελέσματα του Πίνακα 1 αποκαλύπτουν ότι και οι δύο μεταβλητές είναι μη στάσιμες στα επίπεδα των δεδομένων τους. Γίνονται στάσιμες στις πρώτες διαφορές τους, ως εκ τούτου θα μπορούσαν να περιγραφούν ως ολοκληρωμένες πρώτης τάξης I(1). Δεδομένου ότι οι σειρές είναι ολοκληρωμένες πρώτης τάξης I(1), η μέθοδος συνολοκλή-ρωσης των Johansen and Juselius μπορεί να εφαρμοστεί.

Το μήκος της βέλτιστης υστέρησης είναι σημαντικό για να προσδιοριστεί η πραγματική δυναμική του υποδείγματος VAR. Τα αποτελέσματα για την επιλογή της βέλτιστης υστέρησης του υποδείγματος VAR για τις δύο μεταβλητές παρουσιάζονται στον Πίνακα 2.

Πίνακας 2: Κριτήρια για τις υστερήσεις των μεταβλητών του VAR υποδείγματος

Υστερήσεις	LR	FPE	AIC	SC	HQ
0	NA	5.31e+25	64.90917	65.00047	64.90072
1	24.55455*	1.02e+25	63.24837	63.52225	63.22302
2	7.797689	8.02e+24*	62.95339*	63.40986*	62.91113*
3	2.128305	1.19e+25	63.22077	63.85983	63.16162

Σημειώσεις:* δηλώνει τον αριθμό των υστερήσεων που επιλέγονται από τα κριτήρια LR: τροποποιημένος στατιστικός έλεγχος, FPE: Τελικό σφάλμα πρόβλεψης, AIC: Κριτήριο του Akaike, SC: Κριτήριο του Schwarz, HQ: Κριτήριο των Hannan-Quinn.

Από τα αποτελέσματα του Πίνακα 2 παρατηρούμε ότι όλα τα κριτήρια (εκτός του LR) μας δείχνουν 2 υστερήσεις στο δυναμικό υπόδειγμα VAR. Επομένως η τάξη του υποδείγματος VAR είναι k=2.

Στη συνέχεια, χρησιμοποιούμε την τεχνική του Johansen (1988, 1991), και των Johansen και Juselius (1990) με την πλήρη μέθοδο της μέγιστης πιθανοφάνειας (ML) για να διαπιστώσουμε αν υπάρχουν συνολοκληρωμένα διανύσματα μεταξύ των χρονικών σειρών. Ο Πίνακας 3 παρουσιάζει τα αποτελέσματα από τους ελέγχους συνολοκλήρωσης του Johansen.

Πίνακας 3: Έλεγχος Συνολοκλήρωσης του Johansen

Null Hypothesis	Statistics		5% critical value	
	Trace test	Max-Eigen	Trace test	Max-Eigen
BD CAD (Order VAR k = 2)				
$r = 0$	13.188	12.645	15.49	14.26
$r \leq 1$	0.542	0.542	3.841	3.841

Σημειώσεις: 1. Οι κρίσιμες τιμές προέρχονται από τους Osterwald – Lenum (1992). 2. Το r δηλώνει τον αριθμό των συνολοκληρωμένων διανυσμάτων. 3. Για την τάξη του υποδείγματος VAR, χρησιμοποιούμε τα κριτήρια LR, FPE, AIC, SC, και HQ.

Τα εμπειρικά αποτελέσματα από τον Πίνακα 3 δείχνουν ότι τόσο η μέγιστη ιδιοτιμή όσο και το ίχνος έχουν τιμές μικρότερες από τις κρίσιμες τιμές σε επίπεδο σημαντικότητας 5%. Ως εκ τούτου, η μηδενική υπόθεση των μη συνολοκληρωμένων διανυσμάτων δεν απορρίπτεται. Άρα μπορούμε να πούμε ότι δεν υπάρχει συνολοκληρωμένο διάνυσμα και επομένως ούτε και μακροχρόνια σχέση ισορροπίας μεταξύ των μεταβλητών.

Ο Πίνακας 4 παρουσιάζει τα αποτελέσματα της αιτιότητας των Toda και Yamamoto με τη βέλτιστη υστέρηση (k), των σειρών σε ένα δυναμικό υπόδειγμα VAR ($k + d_{\max}$).

Πίνακας 4: Έλεγχος μη αιτιότητας των Toda-Yamamoto

Μηδενική Υπόθεση	Lag(k)	$k + d_{\max}$	MWald	Συμπέρασμα
BD δεν αιτιάζεται κατά Granger CAD	2	2+1	2.246 (0.170)	Δεν απορρίπτεται η H_0
CAD δεν αιτιάζεται κατά Granger BD	2	2+1	1.391 (0.322)	Δεν απορρίπτεται η H_0

Σημειώσεις: 1. Το $k + d_{\max}$ δηλώνει την τάξη του υποδείγματος VAR. 2. Η επιλογή του μήκους της υστέρησης βασίστηκε στα κριτήρια LR, FPE, AIC, SC, και HQ. 3. ***, ** και * δηλώνει το επίπεδο σημαντικότητας 1%, 5% και 10%.

Τα αποτελέσματα του Πίνακα 4 μας δείχνουν ότι δεν υπάρχουν αιτιακές σχέσεις ανάμεσα στο έλλειμμα του κρατικού προϋπολογισμού και το ισοζύγιο τρεχουσών συναλλαγών για την Ελλάδα την περίοδο που μελετάμε.

6. ΣΥΜΕΡΑΣΜΑΤΑ

Η παραπάνω έρευνα μας οδηγεί στα ακόλουθα βασικά συμπεράσματα.

1) Σύμφωνα με τα αποτελέσματα του Πίνακα 3 δεν υπάρχει μακροχρόνια ισορροπία μεταξύ του ελλείμματος του προϋπολογισμού και του ελλείμματος του ισοζυγίου τρεχουσών συναλλαγών.

2) Χρησιμοποιώντας το έλεγχο μη αιτιότητας των Toda και Yamamoto βρήκαμε ότι καμιά αιτιακή σχέση δεν υπάρχει ανάμεσα στα δίδυμα ελλείμματα.

Τα ευρήματα αυτά για την Ελλάδα επιβεβαιώνουν την ισχύ της θεωρίας του Ρικάρντο (ρικαρδιανού ισοδυνάμου), καθώς και την υπόθεση του Barro. Επομένως η δημοσιονομική προσαρμογή επιβάλλεται για τους παρακάτω λόγους:

(i) Ενίσχυση του ελέγχου των δαπανών.

(ii) Ενίσχυση στην αποδοτικότητα των εσόδων.

(iii) Θέσπιση μέτρων για την αντιστάθμιση της αστάθειας των εσόδων που προέρχονται από μη φορολογικά έσοδα.

(iv) Επαναπροσανατολισμός των πόρων για ενίσχυση της ανάπτυξης και των επενδύσεις και εξορθολογισμό της δημόσιας διοίκησης.

(v) Ενίσχυση της διαχείρισης του χρέους.

ABSTRACT

In this paper we investigate the causal relationship between the budget deficit and the current account deficit for Greece for the period 1995-2011. We explore this relationship using augmented vector autoregressive (VAR) methodology for causality developed by Toda and Yamamoto (1995). The empirical results suggest that causal relationships do not exist between the examined variables.

Keywords: Budget Deficit, Current Account Deficit, Cointegration, Toda and Yamamoto Causality Test

JEL Classification: C22, E31, E50

ΑΝΑΦΟΡΕΣ

Abell, J.D. (1990). Twin deficits during the 1980's: An empirical investigation. *Journal of Macroeconomics*, 12, 81-96.

Ahmed, S. (1986). Temporary and permanent government spending in an open economy: Some evidence for the United Kingdom. *Journal of Monetary Economics*, 17, 197-224.

Akaike, H. (1974). A New Look at the Statistical Model Identification, *IEEE Transaction on Automatic Control*, 19(6), 716 - 723.

Anoruo, A., and Ramchander, S. (1998). Current account and fiscal deficits: Evidence from five developing economies of Asia. *Journal of Asian Economics*, 9(3), 487-501.

Barro, R.J. (1974). Are government bonds net wealth? *Journal of Political Economy*, 82, 1095-1117.

Barro, R.J. (1989). The Ricardian approach to budget deficits. *Journal of Economic Perspectives*, 3, 37-54.

- Darrat, A.F. (1988). Have large budget deficits caused rising trade deficits? *Southern Economic Journal*, 54, 879-886.
- Dickey, D. A. and Fuller, W.A. (1979). Distribution of Estimators for Autoregressive Time Series with a Unit Root, *Journal of the American Statistical Association*, 74(366), 427-431.
- Hatemi, A., and Shukur, G. (2002). Multivariate-based causality tests of twin deficits in the US. *Journal of Applied Statistics*, 29, 817-824.
- Granger, C.W.J. (1969). Investigating Causal Relations by Econometric Models and Cross-spectral Methods, *Econometrica* 37(3), 24-36.
- Islam, M.F. (1998). Brazil's twin deficits: An empirical examination. *Atlantic Economic Journal*, 26, 121-128.
- Johansen, S. and Juselius, K. (1990). Maximum Likelihood Estimation and Inference on Cointegration-with Application to the Demand for Money, *Oxford Bulletin of Economics and Statistics*, 52(2), 169-210.
- Khalid, A.M., & Teo, W.G. (1999). Causality tests of budget and current account deficits: Cross-country comparisons. *Empirical Economics*, 24, 389-402.
- Kwiatkowski, D., Phillips, P., Schmidt, P., & Shin, Y. (1992). Testing the null hypothesis of Stationarity against the alternative of a unit root. *Journal of Econometrics*, 54 (1-3), 159-178.
- Laney, O.L. (1984). The strong dollar, the current account, and the federal deficits: Cause and effect. Federal Reserve Bank of Dallas, *Economic Review*, January, 1-14.
- MacKinnon, J.G. (1991). Critical Value for Cointegration Test. In R.F. Engle and C.W.J. Granger (Eds) *Long-run Economic Relationship: Reading in Cointegration*. Oxford: Oxford University Press.
- Newey, W.K. and K.D. West (1994). Automatic lag selection in covariance matrix estimation, *Review of Economic Studies*, 61(4), 631-653.
- Normadin, M. (1999). Budget deficit persistence and the twin deficits hypothesis. *Journal of International Economics*, 49, 171-193.
- Osterwald-Lenum, M. (1992). A Note with Quantiles of the Asymptotic Distribution of the Maximum Likelihood Cointegration Rank Test Statistics, *Oxford Bulletin of Economics and Statistics* 54(3), 461-471.
- Phillips, P.C.B. and Perron, P. (1988). Testing for a Unit Root in Time Series Regression, *Biometrika*, 75(2), 335-346.
- Rambaldi, A.N. and Doran, H. (1996). Testing for Granger non-causality in cointegrated systems made easy,"Working Paper in Econometrics and Applied Statistics No. 88, Department of Econometrics, University of New England, Armidale.
- Summers, L.H., (1988). Tax policy and international competitiveness, NBER Chapters, in: *International Aspects of Fiscal Policies* (pp. 349-386). National Bureau of Economic Research, Inc.
- Toda, H.Y., and Yamamoto, T. (1995). Statistical inference in vector autoregressive with possibly integrated processes. *Journal of Econometric*, 66, 225-250.
- Vamvoukas, G.A. (1997). Have large budget deficits caused increasing trade deficits? Evidence from a developing country. *Atlantic Economic Journal*, 25(1), 80-90.



ΣΧΕΔΙΑΣΜΟΙ ΛΑΤΙΝΙΚΩΝ ΥΠΕΡΚΥΒΩΝ

Ευθυμίου Ι., Γεωργίου Στ.

Τμήμα Στατιστικής και Αναλογιστικών- Χρηματοοικονομικών Μαθηματικών,
Πανεπιστήμιο Αιγαίου
{iefthimiu, stgeorgiou}@aegean.gr

ΠΕΡΙΛΗΨΗ

Οι σχεδιασμοί Λατινικών υπερκύβων (LHDs) χρησιμοποιούνται ευρέως στο σχεδιασμό πειραμάτων με υπολογιστές [1-9]. Σ' αυτή την εργασία, παρουσιάζεται συνοπτικά μια καινούργια μέθοδος για την κατασκευή ορθογώνιων σχεδιασμών Λατινικών υπερκύβων (OLHDs) με 12, 16, 20 και 24 παράγοντες και με ευέλικτο αριθμό πειραματικών εκτελέσεων. Αποδεικνύεται ότι αν υπάρχει ένας OLH (n, m), τότε υπάρχει και ένας OLH ($24n, 12m$), ένας OLH ($32n, 16m$), ένας OLH ($40n, 20m$), ένας OLH ($48n, 24m$), ένας OLH ($24n+1, 12m$), ένας OLH ($32n+1, 16m$), ένας OLH ($40n+1, 20m$) και ένας OLH ($48n+1, 24m$). Κατασκευάζονται OLHDs με την ιδιότητα ότι κάθε τετραγωνική επίδραση ή οποιαδήποτε αλληλεπίδραση δύο παραγόντων είναι ορθογώνια ως προς όλες τις κύριες επιδράσεις του σχεδιασμού. Τέλος, προτείνονται νέες μέθοδοι πολλαπλασιασμού και άλλες κατασκευές πολλών άπειρων οικογενειών ορθογώνιων σχεδιασμών Λατινικών υπερκύβων με πολλούς παράγοντες. Οι κλάσεις των OLHDs που προκύπτουν είναι νέες αφού σχεδιασμοί με αυτές τις παραμέτρους δεν μπορούν να κατασκευαστούν από άλλες γνωστές μεθόδους της βιβλιογραφίας.

Λέξεις Κλειδιά: Ορθογώνιοι σχεδιασμοί Λατινικών υπερκύβων, Περιοδική συνάρτηση Αυτοσυσχέτισης, Κυκλικοί πίνακες, Πειράματα με υπολογιστές.

1. ΕΙΣΑΓΩΓΗ

Ένας πειραματικός σχεδιασμός με n πειραματικές εκτελέσεις (παρατηρήσεις) και m παράγοντες είναι ένας $n \times m$ πίνακας $X = [x_1, x_2, \dots, x_m]$, όπου x_j είναι ο j -οστός παράγοντας και x_{ij} είναι το επίπεδο του j παράγοντα στην i -οστή πειραματική εκτέλεση. Πιο συγκεκριμένα, το $x_{ij} = 1$ αν ο παράγοντας j μετέχει στην i -οστή παρατήρηση και $x_{ij} = 0$ αν δεν μετέχει, δηλαδή ο πίνακας X έχει στοιχεία 0 και 1. Θα έχει στοιχεία ± 1 αν κάθε παράγοντας έχει δύο επίπεδα.

Ένας σχεδιασμός $X = L(n, m)$ θα καλείται σχεδιασμός Λατινικού υπερκύβου (LHD) με n πειραματικές εκτελέσεις (παρατηρήσεις) και m παράγοντες αν κάθε μεταβλητή του σχεδιασμού περιλαμβάνει n επίπεδα που κατανομούνται ομοιόμορφα, δηλαδή κάθε στήλη έχει $n/2$ στοιχεία ίσα με 1 και $n/2$ ίσα με 0 ή ± 1 , οπότε n να είναι άρτιος.

Συνήθως στην ανάλυση παλινδρόμησης ένα πολυωνυμικό μοντέλο βαθμού k με m παράγοντες είναι της μορφής:

$$Y = \beta_0 + \sum_{i \leq m} \beta_i x_i + \sum_{i_1 \leq i_2 \leq m} \beta_{i_1 i_2} x_{i_1} x_{i_2} + \dots + \sum_{i_1 \leq \dots \leq i_k \leq m} \beta_{i_1 \dots i_k} x_{i_1} \dots x_{i_k} + \varepsilon,$$

όπου x_i είναι οι ανεξάρτητες μεταβλητές, β_i είναι οι γραμμικές επιδράσεις των x_i , $\beta_{i_1 \dots i_t}$ είναι η αλληλεπίδραση t -τάξης των $x_{i_1}, \dots, x_{i_t}$ και ανεξάρτητα $\varepsilon \sim N(0, \sigma^2)$. Προφανώς τα β_{ii} αντιστοιχούν στην τετραγωνική επίδραση του παράγοντα x_i και το $\beta_{i_1 i_2}$ αντιστοιχεί στην αλληλεπίδραση δεύτερης τάξης των παραγόντων x_{i_1}, x_{i_2} , για $i_1 \neq i_2$.

Για να μπορούμε να χρησιμοποιήσουμε τους LHDs που κατασκευάσαμε σε ένα μοντέλο δεύτερης τάξης για κρησάρισμα και να έχουμε ικανοποιητικές εκτιμήσεις των παραμέτρων, θα πρέπει οι LHDs να ικανοποιούν τις εξής ιδιότητες:

- Κάθε στήλη να είναι ορθογώνια ως προς τις άλλες στήλες του σχεδιασμού.
- Κάθε στήλη να είναι ορθογώνια με οποιαδήποτε στήλη που αντιστοιχεί σε τετραγωνική επίδραση ή αλληλεπίδραση δύο παραγόντων.

Κάθε σχεδιασμός Λατινικού Υπερκύβου $L(n, m)$ που ικανοποιεί την ιδιότητα a) καλείται ορθογώνιος και συμβολίζεται με OLHD (n, m).

Έστω $A = \{A_j : A_j = (a_{j0}, a_{j1}, \dots, a_{j(n-1)}), j = 1, \dots, \ell\}$, ένα σύνολο από ℓ διανύσματα μήκους n . Η περιοδική συνάρτηση αυτοσυσχέτισης $P_A(s)$ (θα την συμβολίζουμε ως PAF) ορίζεται, θεωρώντας τους δείκτες $i + s$ modulo n , ως εξής:

$$P_A(s) = \sum_{j=1}^{\ell} \sum_{i=0}^{n-1} a_{ji} a_{j, i+s}, \quad s = 0, 1, \dots, n-1. \quad (1)$$

Το σύνολο των διανυσμάτων A θα λέμε ότι έχει μηδενική PAF αν $P_A(s) = 0, \forall s = 1, 2, \dots, n-1$ και γενικότερα θα λέμε ότι έχει σταθερή PAF αν $P_A(s) = \gamma, \forall s = 1, 2, \dots, n-1$ για κάποιον ακέραιο αριθμό γ .

Σύνολα διανυσμάτων με σταθερή PAF χρησιμοποιούνται για να κατασκευάσουμε LHDs που ικανοποιούν τις ιδιότητες α) και β). Με R_k ορίζουμε των πίσω διαγώνιο ταυτοτικό πίνακα τάξης k . Ένας κυκλικός πίνακας ορίζεται ως ένας τετραγωνικός πίνακας $B = (b_{ij})$ τάξης n με την πρώτη του γραμμή $b_1 = (b_{1,0}, b_{1,1}, \dots, b_{1,n-1})$ και κάθε επόμενη γραμμή του να είναι κυκλική μετάθεση της προηγούμενης γραμμής, ($b_{ij} = b_{1,j-i+1}$, όπου $j-i+1$ λαμβάνεται modulo n , $i = 2, 3, \dots, n$ και $j = 0, 1, 2, \dots, n-1$). Στα ακόλουθα λήμματα χρησιμοποιούνται κυκλικοί πίνακες για την κατασκευή ορθογώνιων πινάκων.

Λήμμα 1. [Theorem 4.49, Geramita & Seberry (1979)]: Έστω τέσσερις κυκλικοί πίνακες A, B, C, D τάξης n που ικανοποιούν την σχέση

$$AA^T + BB^T + CC^T + DD^T = fI_n.$$

τότε ο Goethal-Seidel σχηματισμός

$$GS = GS(A, B, C, D) = \begin{pmatrix} A & BR_n & CR_n & DR_n \\ -BR_n & A & -D^T R_n & C^T R_n \\ -CR_n & D^T R_n & A & -B^T R_n \\ -DR_n & -C^T R_n & B^T R_n & A \end{pmatrix}$$

είναι ένας ορθογώνιος πίνακας τάξης $4n$.

Ένα σύνολο τετραγωνικών πινάκων $\{A_1, A_2, \dots, A_{2k}\}$ θα λέμε ότι είναι *amicable* αν ικανοποιεί την ακόλουθη σχέση:

$$\sum_{i=1}^k (A_{2i-1} A_{2i}^T - A_{2i} A_{2i-1}^T) = 0.$$

Λήμμα 2. [Kharaghani, 2000, Theorem 1]: Έστω $\{A_i\}_{i=1}^8$ amicable κυκλικοί πίνακες με $\sum_{i=1}^8 A_i A_i^T = fI_n$. Τότε ο Kharaghani σχηματισμός

$$\begin{bmatrix} A_1 & A_2 & A_4 R_n & A_3 R_n & A_6 R_n & A_5 R_n & A_8 R_n & A_7 R_n \\ -A_2 & A_1 & A_3 R_n & -A_4 R_n & A_5 R_n & -A_6 R_n & A_7 R_n & -A_8 R_n \\ -A_4 R_n & -A_3 R_n & A_1 & A_2 & -A_8^T R_n & A_7^T R_n & A_6^T R_n & -A_5^T R_n \\ -A_3 R_n & A_4 R_n & -A_2 & A_1 & A_7^T R_n & A_8^T R_n & -A_5^T R_n & -A_6^T R_n \\ -A_6 R_n & -A_5 R_n & A_8^T R_n & -A_7^T R_n & A_1 & A_2 & -A_4^T R_n & A_3^T R_n \\ -A_5 R_n & A_6 R_n & -A_7^T R_n & -A_8^T R_n & -A_2 & A_1 & A_3^T R_n & A_4^T R_n \\ -A_8 R_n & -A_7 R_n & -A_6^T R_n & A_5^T R_n & A_4^T R_n & -A_3^T R_n & A_1 & A_2 \\ -A_7 R_n & A_8 R_n & A_5^T R_n & A_6^T R_n & -A_3^T R_n & -A_4^T R_n & -A_2 & A_1 \end{bmatrix}$$

είναι ένας ορθογώνιος πίνακας τάξης $8n$.

2. ΑΞΙΟΛΟΓΗΣΗ ΠΕΙΡΑΜΑΤΩΝ ΜΕ ΥΠΟΛΟΓΙΣΤΕΣ

Έστω X_{int} ο $n \times (m(m-1)/2)$ πίνακας με όλες τις αλληλεπιδράσεις δύο παραγόντων και X_{quad} ο $n \times m$ πίνακας με όλους τους καθαρά τετραγωνικούς όρους (Steinberg και Lin, 2006, Georgiou, 2009). Οι πίνακες συσχετίσεων για την εκτίμηση του γραμμικού μοντέλου (πρώτης τάξης) υπό την ύπαρξη σημαντικών αλληλεπιδράσεων δύο παραγόντων ή καθαρών τετραγωνικών όρων δίδονται από την σχέση:

$$T = A_{int} = (X_1^T X_1)^{-1} X_1^T X_{int} \quad (2)$$

και

$$Q = A_{quad} = (X_1^T X_1)^{-1} X_1^T X_{quad} \quad (3)$$

αντίστοιχα. Η σχέση (2) χρησιμοποιείται όταν το μοντέλο έχει τις παραμέτρους $\beta_{i_1 i_2}$, $i_1 \neq i_2$ δευτέρης τάξης ενώ $\beta_{ii} = 0$ και το (3) όταν το μοντέλο έχει

μόνο τις τετραγωνικές β_{ii} ενώ οι υπόλοιπες $\beta_{i_1 i_2} = 0$ για $i_1 \neq i_2$.

Για την αξιολόγηση της απόδοσης ενός σχεδιασμού Λατινικού υπερκύβου ως προς τις αλληλεπιδράσεις δύο παραγόντων έχουμε τα εξής μέτρα:

$$ave(|t|) = E(|t|) = \frac{2 \sum_{i=1}^{m+1} \sum_{j=1}^{m(m-1)/2} |t_{ij}|}{m(m^2-1)}, \text{ and } \max_t t = \max_{i,j} |t_{ij}|. \quad (4)$$

Για την αξιολόγηση της απόδοσης ενός σχεδιασμού Λατινικού υπερκύβου ως προς τους τετραγωνικούς όρους έχουμε τα εξής μέτρα:

$$ave(|q|) = E(|q|) = \frac{\sum_{i=1}^{m+1} \sum_{j=1}^m |q_{ij}|}{m(m+1)}, \text{ and } \max q = \max_{i,j} |q_{ij}|. \quad (5)$$

Ένας σχεδιασμός D_1 θα λέγεται ότι είναι καλύτερος από έναν σχεδιασμό D_2 αν η τιμή του κριτηρίου που μας ενδιαφέρει στον D_1 είναι μικρότερη από την τιμή του ίδιου κριτηρίου στον D_2 .

Το ακόλουθο Λήμμα παρέχει ένα κάτω φράγμα για τα κριτήρια $E(|q|)$ και $\max |q_{ij}|$.

Λήμμα 3. (Georgiou, 2009): Έστω L_k ένας ορθογώνιος σχεδιασμός Λατινικού υπερκύβου με στοιχεία από τον μοναδιαίο κύβο $[-1,1]^{n \times m}$, με n πειραματικές εκτελέσεις (παρατηρήσεις) και m παράγοντες. Τότε,

$$E(|q|) \geq \frac{\gamma}{n(m+1)} = LB_{Eq} \text{ and } \max |q_{ij}| \geq \frac{\gamma}{n} = LB_{maxq}, \quad (6)$$

όπου

$$\gamma = \frac{n(n+1)}{3(n-1)}. \quad (7)$$

3. ΓΕΝΙΚΗ ΚΑΤΑΣΚΕΥΗ

Με την χρήση κατάλληλων διανυσμάτων με περιοδική συνάρτηση αυτοσυσχέτισης ίση με μηδέν κατασκευάζουμε νέες άπειρες οικογένειες ορθογώνιων σχεδιασμών Λατινικών υπερκύβων (OLHDs).

Θεώρημα 1. [Georgiou, S.D. and Efthimiou, I. (2012)]. Έστω D_b ένας ορθογώνιος $n \times n$ πίνακας με απόλυτη τιμή κάθε στήλης του να είναι μια μετάθεση του $(b+1, b+3, \dots, b+2n-1)$, όπου b ένας οποιοσδήποτε φυσικός αριθμός. Τότε υπάρχει ένας OLHD με $2nk$ γραμμές και n στήλες για κάθε $k=1, 2, \dots$

Απόδειξη. Εξ' ορισμού, έχουμε ότι ο D_b είναι ορθογώνιος σχεδιασμός, και έτσι:

$$D_b^T D_b = z_b I_n, \text{ όπου}$$

$$z_b = \sum_{i=1}^n (b+2i-1)^2 = \frac{n(4n^2 + 6bn + 3b^2 - 1)}{3}.$$

Για $k=1, 2, \dots$ ορίζουμε

$$L_k = \left(D_0^T, \dots, D_{2n(k-1)}^T, -D_0^T, \dots, -D_{2n(k-1)}^T \right)^T.$$

Είναι εύκολο να εξακριβωθεί ότι τα επίπεδα της κάθε στήλης του L_k είναι μια μετάθεση των $(1, 3, 5, \dots, 2nk-1, -1, -3, -5, \dots, -2nk-1)$. Οι στήλες του L_k είναι ανά δύο ορθογώνιες αφού οι στήλες του D_b είναι ανά δύο ορθογώνιες για κάθε φυσικό αριθμό b . Έτσι έχουμε:

$$\begin{aligned} L_k^T L_k &= \left(\sum_{\ell=0}^{k-1} D_{2n\ell}^T D_{2n\ell} \right) I_n = \left(2 \sum_{j=1}^{nk} (2j-1)^2 \right) I_n \\ &= \left(\frac{2nk(2nk-1)(2nk+1)}{3} \right) I_n \end{aligned}$$

Και έτσι ο πίνακας L_k είναι ο επιθυμητός OLHD που ψάχναμε.

3.2 ΚΑΤΑΣΚΕΥΕΣ ΝΕΩΝ ΣΧΕΔΙΑΣΜΩΝ

Πόρισμα 1. Υπάρχει ένας ορθογώνιος σχεδιασμός Λατινικού υπερκύβου (OLHD) με $n=24k$ πειραματικές εκτελέσεις και $m=12$ παράγοντες, με $n=32k$ πειραματικές εκτελέσεις και $m=16$ παράγοντες, με $n=40k$ πειραματικές εκτελέσεις και $m=20$ παράγοντες, με $n=48k$ πειραματικές εκτελέσεις και $m=24$ παράγοντες, για κάθε $k = 1, 2, \dots$.

Πόρισμα 2. Υπάρχει ένας ορθογώνιος σχεδιασμός Λατινικού υπερκύβου (OLHD) με $n=24k+1$ πειραματικές εκτελέσεις και $m=12$ παράγοντες, με $n=32k+1$ πειραματικές εκτελέσεις και $m=16$ παράγοντες, με $n=40k+1$ πειραματικές εκτελέσεις και $m=20$ παράγοντες, με $n=48k+1$ πειραματικές εκτελέσεις και $m=24$ παράγοντες, για κάθε $k = 1, 2, \dots$.

3.3 ΠΑΡΑΔΕΙΓΜΑΤΑ

Παράδειγμα. Ορίζουμε $k = 1$, $b = 40(k-1) \Rightarrow b = 0$. Χρησιμοποιώντας το Goethal-Seidel σχηματισμό και $b=0$, έχουμε έναν 20×20 ορθογώνιο πίνακα D_0 . Εφαρμόζοντας την μέθοδο που περιγράφεται στην απόδειξη του Θεωρήματος 1 έχουμε έναν ορθογώνιο σχεδιασμό Λατινικού υπερκύβου $L_1 = [D_0^T \ -D_0^T]^T$ με 40 πειραματικές εκτελέσεις (παρατηρήσεις) και 20 παράγοντες, όπου ο D_0 αναλυτικά είναι:

21	5	-27	29	23	-37	15	33	31	25	-9	-7	-3	1	39	-19	17	-15	13	11
29	21	5	-27	29	35	33	31	25	-37	-7	-3	1	39	-9	17	-15	13	11	-19
29	23	21	5	-27	33	31	25	-37	35	-3	1	39	-9	-7	-15	13	11	-19	17
-27	29	23	21	5	31	35	-37	35	33	1	39	-9	-7	-3	13	11	-19	17	-15
5	-27	29	23	21	25	-37	35	33	31	39	-9	-7	-3	1	11	-19	17	-15	13
27	-35	-33	-31	-25	21	5	-27	29	23	-13	15	-17	19	-11	1	-3	-7	-9	39
-35	-33	-31	-25	27	23	21	5	-27	29	15	-17	19	-11	-13	-3	-7	-9	39	1
-13	-31	-25	27	-35	29	23	21	5	-27	-17	19	-11	-13	15	-7	-9	39	1	3
-31	-25	27	-35	-33	-27	29	2	21	5	19	-11	-13	15	-17	-9	39	1	-3	-7
-35	37	-35	-33	-31	5	-27	29	23	21	-11	-13	15	-17	19	39	1	-3	-7	-9
9	7	3	-1	-39	13	-15	17	-19	11	21	5	-27	29	23	-31	-33	-35	37	-25
7	3	-1	-39	9	-15	17	-19	11	13	23	21	5	-27	29	-33	-35	37	-25	-31
3	-1	-39	9	7	17	-19	11	13	-15	29	23	21	5	-27	-35	37	-25	-31	-33
-1	-39	9	7	3	-19	11	13	-15	17	-27	29	23	21	5	37	-25	-31	-33	-35
-39	9	7	3	-1	11	13	-15	17	-19	5	-27	29	23	21	-25	-31	-33	-35	37
19	-17	15	-13	-11	-1	3	7	9	-39	31	33	35	-37	25	21	5	-27	29	23
-17	15	-13	-11	19	3	7	9	-39	-1	33	35	-37	25	31	23	21	5	-27	29
15	-13	-11	19	-17	7	9	-39	-1	3	35	-37	25	31	33	29	23	21	5	-27
-13	-11	19	-17	15	9	-39	-1	3	7	-37	25	31	33	35	-27	29	23	21	5
-11	19	-17	15	-13	-39	-1	3	7	9	25	31	33	35	-37	5	-27	29	23	21

Ο παραπάνω σχεδιασμός μας έδωσε τα εξής αποτελέσματα: $E(|t|) = \max |t_{ij}| = 0$, $\max |q_{ij}| = \frac{41}{117} = LB_{maxq}$, $J_R = [24, 40, 40, 64, 32, 40, 40, 40, 40, 16, 8, 8, 36, 64, 32, 72, 64, 24, 36, 8, 8, 8, 40, 20]$, $\Phi_{100}^R = 0.085$, $D_E = [3.744, 5.295]$, $J_E = [760, 20]$, $\Phi_{100}^E = 0.285$, $D_R = \left[\frac{160}{13}, \frac{484}{39}, \frac{164}{13}, \frac{512}{39}, \frac{172}{13}, \frac{40}{3}, \frac{524}{39}, \frac{532}{39}, \frac{536}{39}, \frac{548}{39}, \frac{556}{39}, \frac{560}{39}, \frac{188}{13}, \frac{192}{13}, \frac{580}{39}, \frac{200}{13}, \frac{800}{39} \right]$.

Παράδειγμα 2. Ορίζουμε $k=1$, $b=40k-39 \Rightarrow b=1$. Χρησιμοποιώντας το προηγούμενο πόρισμα και $b=1$, έχουμε έναν 20×20 ορθογώνιο πίνακα D_1 . Εφαρμόζοντας την μέθοδο που περιγράφεται στην απόδειξη του *Θεωρήματος 1* έχουμε έναν ορθογώνιο σχεδιασμό Λατινικού υπερκύβου $L_1 = [D_1 \ 0_{1 \times 12} - D_1]$ με 41 πειραματικές εκτελέσεις (παρατηρήσεις) και 20 παράγοντες, όπου D_1 είναι:

11	3	-14	15	12	-19	10	17	16	13	-5	-4	-2	1	20	-10	9	-8	7	6
12	11	9	-14	15	18	17	16	13	-19	-4	-2	1	20	-5	9	-8	7	6	-10
15	12	11	9	-14	17	16	13	-19	18	-2	1	20	-5	-4	-8	7	6	-10	9
14	15	12	11	9	16	13	19	18	17	1	20	5	4	21	7	6	10	9	8
3	-14	15	12	11	13	-19	18	17	16	20	-5	-4	-21	6	-10	9	9	-8	7
19	-18	-17	-16	-13	11	3	-14	15	12	-7	0	-9	10	-6	1	-2	-4	-5	20
-18	-17	-16	-18	19	12	11	3	-14	15	0	-9	10	-6	-7	-2	-4	-5	20	1
-17	-16	-13	19	-18	15	12	11	3	-14	-9	10	-6	-7	0	-4	-5	20	1	-2
-16	-13	19	-18	-17	-14	15	12	11	3	10	-6	-7	0	-9	-5	20	1	-2	-4
-13	19	-18	-17	-16	3	14	15	12	11	-6	-7	0	-9	10	20	1	-2	-4	-5
5	4	2	-1	-20	7	-8	9	-10	6	11	3	-14	15	12	-16	-17	-18	19	-13
4	2	-1	-20	5	-8	9	-10	6	7	12	11	3	-14	15	-17	-18	19	-13	-16
2	-1	-20	5	4	9	-10	6	7	-8	15	12	11	3	-14	-18	19	-13	-16	-17
-1	-20	5	4	2	-10	6	7	-8	9	-14	15	12	11	3	19	-13	-16	-17	-18
-20	5	4	2	-1	6	7	-8	9	10	3	-14	15	12	11	-13	-16	-17	-18	19
10	4	0	-7	-6	-1	2	4	5	-20	16	17	18	-19	13	11	3	-14	15	12
9	0	7	6	10	2	4	5	20	1	17	18	19	13	16	12	11	3	14	15
0	-7	-6	10	-9	4	5	-20	-1	2	18	-19	13	16	17	15	12	11	3	-14
-7	-6	10	-9	0	5	-20	-1	2	4	-19	13	16	17	18	-14	15	12	11	3
-6	10	-9	0	-7	-20	-1	2	4	5	12	16	17	18	-19	3	-14	15	12	11

4. ΤΕΧΝΙΚΕΣ ΠΟΛΛΑΠΛΑΣΙΑΣΜΟΥ - ΣΥΓΚΡΙΣΕΙΣ

Θεώρημα 2. [Georgiou, S.D. and Efthimiou, I. (2012)]. Υποθέτουμε ότι έχουμε ένα OLHD(n,m) με n πολλαπλάσιο του 4, και έστω ένας πίνακας Hadamard τάξης n . Τότε μπορούν να κατασκευαστούν:

- Οι ορθογώνιοι σχεδιασμοί Λατινικών υπερκύβων OLHD($2an$, am), για $a=12, 16, 20, 24$.
- Οι ορθογώνιοι σχεδιασμοί Λατινικών υπερκύβων OLHD($2an + 1$, am), για $a=12, 16, 20, 24$.

Η *απόδειξη* του παραπάνω θεωρήματος παρατίθεται στους Georgiou, S.D. and Efthimiou, I. (2012).

Ως παράδειγμα, στον επόμενο πίνακα υπολογίζουμε όλα τα κριτήρια που χρησιμοποιούνται σε αυτή την εργασία για τις τρεις μεθόδους κατασκευής του σχεδιασμού OLHD(192, 48). Τα αποτελέσματα δείχνουν ότι ο σχεδιασμός που έχει fold-over structure είναι πολύ καλύτερος από τους σχεδιασμούς που κατασκευάζονται με την μέθοδο του γινομένου Kronecker τόσο ως προς την ορθογωνιότητα ως και ως προς τα κριτήρια space-filling.

Πίνακας 1. Ιδιότητες των OLHD(192, 48).

	Our Theorem	Th.3, (Linetal2010)
$E(q)$	$\frac{9370175}{1035002451} = 0.0091$	$\frac{3037011}{345000817} = 0.0088$
$\max q_{ij} $	$\frac{193}{573} = 0.3368$	$\frac{193}{573} = 0.3368$
$E(t)$	$\frac{2906492}{2316434057} = 0.0013$	$\frac{606716585}{389160921576} = 0.0016$
$\max t_{ij} $	$\frac{257600}{7040833} = 0.0366$	$\frac{441333}{7040833} = 0.0627$
D_E	3.867, 3.920, 3.921	3.247, 3.253, 3.262
D_R	11.937, 12.188, 12.439	9.549, 9.633, 9.717
J_E	[2, 4, 2]	[1, 1, 1]
J_R	[2, 2, 20]	[4, 8, 4]
Φ_{100}^E	8.410140547087	8.401007308763
Φ_{100}^R	50.54701795254	50.50470181403

Πίνακας 2. Κατασκευή των OLHD από το Θεώρημα 2.

Runs	Factors	LBST Factors	SLL Factors	Construction		
				n	m	σ
24	12	8	4	1	1	12
32	16	12	16	1	1	16
40	20	—	4	1	1	20
48	24	12	8	1	1	24
64	16	32	32	2	1	16
80	20	12	8	2	1	20
96	24	24	16	2	1	24
128	32	48	64	4	2	16
160	40	24	16	4	2	20
192	48	48	32	8	4	12
256	64	192	128	8	4	16
320	80	48	32	8	4	20
384	144	48	64	8	6	24
512	128	—	256	16	8	16

576	144	—	32	12	6	24
640	160	96	64	20	10	16
768	192	96	128	16	8	24
768	288	96	128	16	12	24
960	240	24	32	20	10	24
1024	256	384	512	32	16	16
1152	288	96	64	24	12	24
1280	320	192	128	32	16	20
1536	384	192	64	32	16	24
1600	400	—	32	40	20	20
1920	480	48	64	40	20	24
2304	576	192	128	48	24	24
4608	1152	192	256	96	48	24
9216	2304	384	512	192	96	24

LBST: Lin, Bingham, Sitter, & Tang (2010), *Annal of Statistics*.

SLL: Sun, Liu, & Lin (2009), *Biometrika*.

5. ΣΥΝΟΨΗ

Σε αυτή την εργασία, προτείνουμε νέες διαδικασίες για την κατασκευή ορθογώνιων σχεδιασμών Λατινικών υπερκύβων (OLHDs) με 12, 16, 20 και 24 παράγοντες και ευέλικτο πλήθος πειραματικών εκτελέσεων (παρατηρήσεων), χωρίς τους περιορισμούς που απαιτούνται στους Butler (2001), Ye (1998) ή Lin et al (2006).

Επίσης δείξαμε ότι αν υπάρχει ένας OLH(n, m), τότε υπάρχουν και οι σχεδιασμοί OLH($24n, 12m$), OLH($32n, 16m$), OLH($40n, 20m$), OLH($48n, 24m$), OLH($24n + 1, 12m$), OLH($32n + 1, 16m$), OLH($40n + 1, 20m$) και OLH($48n + 1, 24m$).

ABSTRACT

Latin hypercube designs (LHDs) are commonly used in designing computer experiments. A number of methods have been proposed to construct LHDs with orthogonality among the main-effects. In this paper, we propose a new method for constructing orthogonal LHDs (OLHDs) with 12, 16, 20, and 24 factors having a flexible run size. Moreover, using these designs we provide new multiplication methods and further constructions for OLHDs. These constructions lead to infinite families of OLHD with many factors. For example, we show that when an OLHD(n, m) exists, there also exist OLHDs with (runs, factors) $\in \{(24n, 12m), (32n, 16m), (40n, 20m), (48n, 24m), (24n+1, 12m), (32n+1, 16m), (40n+1, 20m), (48n+1, 24m)\}$.

ΑΝΑΦΟΡΕΣ

- Butler, N.A. (2001). Optimal and orthogonal Latin hypercube designs for computer experiments. *Biometrika* 88, 847-857.
- Georgiou, S.D. (2009). Orthogonal Latin hypercube designs from generalized orthogonal designs. *J. Statist. Inference* 139, 1530-1540.
- Georgiou, S.D. and Efthimiou, I. (2012). Some classes of orthogonal Latin hypercube designs. *Statist. Sinica* (to appear), DOI:"10.5705/SS.2012.142".
- Geramita, A.V. and Seberry, J. (1979). Orthogonal Designs: Quadratic Forms and Hadamard Matrices. Marcel Dekker, New York-Basel.
- Kharaghani, H. (2000). Arrays for orthogonal designs, *J. Combin. Designs* 8, 166-173.
- Lin, C.D., Bingham, D., Sitter, R.R., and Tang, B. (2010). A new and flexible method for constructing designs for computer experiments. *Ann. Statist.* 38, 1460-1477.
- Steinberg, M. and Lin, D.K.J. (2006). A construction method for orthogonal Latin hypercube designs. *Biometrika* 93, 279-288.
- Sun, F.S., Liu, M.Q., and Lin, D.K.J. (2009). Construction of orthogonal Latin hypercube designs. *Biometrika* 96, 971-974.
- Ye, Kenny Q. (1998). Orthogonal column Latin hypercubes and their application in computer experiments. *J. American Statist. Assoc.* 93, 1430-1439.



ΔΙΕΡΕΥΝΗΣΗ ΕΝΟΣ ΣΤΟΧΑΣΤΙΚΟΥ ΜΟΝΤΕΛΟΥ ΓΙΑ ΤΟΝ ΔΙΑΧΩΡΙΣΜΟ ΤΟΥ ΘΟΡΥΒΟΥ ΑΠΟ ΤΗΝ ΑΦΙΞΗ ΠΛΗΡΟΦΟΡΙΑΣ ΣΤΗΝ ΕΞΕΛΙΞΗ ΤΗΣ ΤΙΜΗΣ ΜΙΑΣ ΜΕΤΟΧΗΣ.

Ο. Θεοδοσιάδου¹, Β. Πολυμένης², Γ. Τσακλίδης¹

¹Τμήμα Μαθηματικών, Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης
rania.theo@hotmail.com, tsaklidi@math.auth.gr

²Τμήμα Οικονομικών Επιστημών, Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης
polimenis@econ.auth.gr

ΠΕΡΙΛΗΨΗ

Στην παρούσα εργασία παρατίθενται τα αποτελέσματα της διερεύνησης ενός στοχαστικού μοντέλου που αποσκοπεί στο να διαχωρίσει τον θόρυβο από την άφιξη πληροφορίας στην εξέλιξη της τιμής μιας μετοχής. Πιο συγκεκριμένα, θεωρούμε πως η μεταβλητότητα της τιμής μίας μετοχής μπορεί να οφείλεται τόσο σε θόρυβο, ο οποίος δεν εμπεριέχει καμία πληροφορία και εξελίσσει τις τιμές κατά συνεχή τρόπο, όσο και σε άλματα στην τιμή της μετοχής, τα οποία δημιουργούνται κατά τη στιγμή άφιξης μίας νέας πληροφορίας στην αγορά. Η ανάλυση του μοντέλου στηρίζεται στη μελέτη των κεντρικών ροπών τρίτης τάξης και διερευνώνται δυο νέοι παράγοντες κινδύνου (γ και δ), που περιγράφουν την πιθανά ασύμμετρη έκθεση της τιμής μιας μετοχής στα -θετικά και αρνητικά- άλματα της αγοράς, τα οποία όμως δεν είναι άμεσα μετρήσιμα.

Λέξεις κλειδιά: θόρυβος, άφιξη πληροφορίας, άλματα τιμών, ροπές αποδόσεων.

1. ΕΙΣΑΓΩΓΗ

Το κλασικό β μετρά σωστά τον συστηματικό κίνδυνο μιας μετοχής μόνο υπό αυστηρούς όρους που στη σύγχρονη βιβλιογραφία αμφισβητούνται. Για παράδειγμα, οι επενδυτές πιθανά ανησυχούν για την έκθεση σε αρνητική λοξότητα και ως εκ τούτου απαιτούν υψηλότερη αποζημίωση για την έκθεσή τους σε πιθανό δυσμενές γεγονός (J. Pan, 2002). Επίσης, είναι μάλλον πρόθυμοι να δεχθούν μικρότερες αποδόσεις για να απολαύσουν έκθεση σε θετικά άλματα. Από τις αρχικές διαπιστώσεις των A. Kraus and R. Litzenberger (1976) και V. Bawa and E. Lindenberg (1977), η συλλοξότητα με τις αποδόσεις της αγοράς έχει αποδειχθεί ότι

είναι ένας σημαντικός παράγοντας στη διαμόρφωση των τιμών (ενδεικτικά βλ. V. Harlow and R. Rao (1989) και C. Harvey and A. Siddique (2000)).

Υπάρχουν επίσης ενδείξεις ασύμμετρης συμπεριφοράς των αποδόσεων στις ανοδικές και καθοδικές αγορές. Από τον R. Merton (1976), οι ερευνητές έχουν αποδεχθεί την ύπαρξη ασυνεχειών στις τιμές χρηματοοικονομικών τίτλων που εκθέτει τους επενδυτές σε άλματα. Έτσι, οι Carr et al. (2002) ερευνούν την ύπαρξη αλμάτων στις χρηματοπιστωτικές αγορές και βρίσκουν ότι τα άλματα μπορεί όχι μόνο να είναι μεγάλα και σπάνια τύπου Poisson, αλλά και μικρότερα που φτάνουν με γρήγορους ρυθμούς.

Στο Polimenis (2012) τα άλματα συμβαίνουν σε τυχαία χρονικά σημεία στα οποία υπάρχει άφιξη πληροφοριών για ολόκληρη την αγορά. Αυτό είναι ένα φυσικό μοντέλο που όμως είναι αρκετά γενικό ώστε να καλύπτει τα περισσότερα εμπειρικά ευρήματα. Σύμφωνα με το υπόδειγμα αυτό α) οι αποδόσεις των μετοχών εκτίθενται, τόσο σε μεταβλητότητα τύπου Brown, όσο και σε άλματα, β) αφίξεις αλμάτων διαφόρων μεγεθών μπορεί να συμβούν σε χαμηλές και υψηλές συχνότητες, γ) οι αποδόσεις των μετοχών μπορεί να παρουσιάζουν διαφορικές ευαισθησίες ως προς τον κίνδυνο άλματος από ότι κίνδυνο Brown και δ) ευαισθησίες στα άλματα είναι πιθανώς ασύμμετρες.

Εάν η αγορά αντιμετωπίζει ασύμμετρα ασυνέχειες τιμών με βάση το πρόσημό τους, η στατιστική ανάλυση σε αρνητικά και θετικά άλματα είναι μεγάλης σημασίας για την τιμολόγηση τίτλων. Με την προϋπόθεση πως τα άλματα δεν είναι ορατά, δεν υπάρχει μια απλή στατιστική διαδικασία για την εκτίμηση των βήτα συντελεστών ως προς αυτά. Ο Polimenis (2012) προτείνει τη χρήση ενός συνόλου ροπών για την εκτίμηση των συντελεστών ως προς τα άλματα. Αυτή η προσέγγιση βασίζεται στη συλλοξότητα, προκειμένου να φιλτράρει τη διεργασία Brown που δε συμμετέχει στις κεντρικές ροπές τρίτης τάξης.

Στην παρούσα εργασία διερευνάται περαιτέρω το στοχαστικό μοντέλο που προτάθηκε από τον Polimenis (2012) με στόχο τον διαχωρισμό του θορύβου από την άφιξη πληροφορίας στην εξέλιξη της τιμής μιας μετοχής. Διαπιστώνεται μεταξύ των άλλων η ευαισθησία των παραμέτρων ως προς την τρίτη (κεντρική) ροπή της απόδοσης της αγοράς για τη μετοχή Google, ενώ αναδεικνύονται και τα όρια των τιμών της τρίτης (κεντρικής) ροπής και της συλλοξότητας με την αγορά ώστε το σύστημα να έχει πραγματικές λύσεις.

2. TO ΜΟΝΤΕΛΟ

Η (λογαριθμική) απόδοση μιας μετοχής i στο χρονικό διάστημα $[t-1, t]$ θεωρούμε ότι δίνεται από τη σχέση

$$R_i(t) = a_i + \gamma_i X(t) - \delta_i Y(t) + \varepsilon_i(t), \quad (1)$$

ενώ η (λογαριθμική) απόδοση της αγοράς δίνεται από τη σχέση

$$R_m(t) = a_m + X(t) - Y(t) , \quad (2)$$

όπου,

a_i, a_m : οι τάσεις στις αποδόσεις της μετοχής i και της αγοράς αντίστοιχα,

$X(t), Y(t)$: οι (ανεξάρτητες) στοχαστικές διαδικασίες των αλμάτων στις αποδόσεις της αγοράς που οφείλονται στην άφιξη θετικών και αρνητικών ειδήσεων αντίστοιχα,

$\varepsilon_i(t)$: ο ιδιοσυγκρατικός θόρυβος της μετοχής i για τον οποίο δεχόμαστε ότι

$$\varepsilon_i(t) \sim N(0, q_i^2),$$

γ_i, δ_i : οι συντελεστές που υποδεικνύουν τον βαθμό επίδρασης της τιμής του θετικού και αρνητικού άλματος αντίστοιχα της απόδοσης της αγοράς στην απόδοση της μετοχής i .

Μέσω του παρακάτω συστήματος εξισώσεων που θα τεθεί προς ανάλυση γίνεται εκτίμηση των συντελεστών γ_i και δ_i μιας μετοχής i κάνοντας χρήση των (κεντρικών) ροπών τρίτης τάξης (Polimenis, 2012):

$$\left. \begin{aligned} s_m &= E(R_m - \mu_m)^3 = s_x - s_y \\ s_{i,m} &= E(R_m - \mu_m)^2 (R_i - \mu_i) = \gamma_i s_x - \delta_i s_y \\ s_{m,i} &= E(R_m - \mu_m)(R_i - \mu_i)^2 = \gamma_i^2 s_x - \delta_i^2 s_y \\ s_i &= E(R_i - \mu_i)^3 = \gamma_i^3 s_x - \delta_i^3 s_y \end{aligned} \right\} \quad (\Sigma 1)$$

όπου,

μ_m, μ_i : οι μέσες τιμές των αποδόσεων της αγοράς και της μετοχής i αντίστοιχα,

s_m, s_i : οι τρίτες ροπές των αποδόσεων της αγοράς και της μετοχής i αντίστοιχα,

$s_{i,m}, s_{m,i}$: οι συλλοξότητες των αποδόσεων της μετοχής i και της αγοράς,

s_x, s_y : οι τρίτες ροπές των αλμάτων στις αποδόσεις της αγοράς λόγω των θετικών και αρνητικών ειδήσεων αντίστοιχα.

Χρησιμοποιώντας τις εξισώσεις του (Σ1) καταλήγουμε στη σχέση

$$\delta_i = \frac{s_{m,i} - \gamma_i s_{i,m}}{s_{i,m} - \gamma_i s_m} = \frac{s_i - \gamma_i s_{m,i}}{s_{m,i} - \gamma_i s_{i,m}}, \quad (3)$$

η οποία μας δίνει τιμές για τα γ_i και δ_i ανεξάρτητες από τις μη παρατηρήσιμες ποσότητες s_x και s_y .

Μέσω του (Σ1) υπολογίστηκαν αριθμητικά οι τιμές των $\gamma_g, \delta_g, s_x, s_y$ για τη μετοχή Google (ο δείκτης i στο (Σ1) αντικαθίσταται από τον δείκτη g για τη συγκεκριμένη μετοχή), που ανήκει στον τεχνολογικό δείκτη NASDAQ,

χρησιμοποιώντας τις καθημερινές τιμές κλεισίματος της μετοχής για την περίοδο 2006-2008. Τα αποτελέσματα (Polimenis, 2012) δείχνουν ότι ο συντελεστής γ_g είναι αισθητά μεγαλύτερος του δ_g , όπως φαίνεται στον πίνακα που ακολουθεί.

Πίνακας 1. Οι τιμές των συντελεστών $\gamma_g, \delta_g, s_x, s_y$ για τη μετοχή Google.

	μ	σ	s	$s_{g,m}$	$s_{m,g}$	γ_g	δ_g	s_x	s_y
Market	-4.4 bp	1.70%	-0.93						
Google	-4 bp	2.50%	6.66	-0.57	0.86	4.1	1	0.11	1.04

Σημείωση. Οι τιμές των ροπών δεύτερης και τρίτης τάξης είναι οι αναγραφόμενες $\times 10^{-6}$.

3. ΑΝΑΛΥΣΗ ΕΥΑΙΣΘΗΣΙΑΣ

Ιδιαίτερο ενδιαφέρον παρουσιάζει η ανάλυση ευαισθησίας του (Σ1) στην περιοχή της λύσης ($s_m = -0.93 \times 10^{-6}$, $s_{g,m} = -0.57 \times 10^{-6}$, $s_{m,g} = 0.86 \times 10^{-6}$, $s_g = 6.66 \times 10^{-6}$). Για τον σκοπό αυτόν στην παρούσα ενότητα, θα διατηρήσουμε ενδεικτικά σταθερές τις τρεις τελευταίες παραμέτρους, μεταβάλλοντας την τέταρτη s_m . Δηλαδή, θα αναζητήσουμε τα χαρακτηριστικά της ευαισθησίας στην περιοχή της λύσης εξετάζοντας σημεία της ευθείας (s_m , $s_{g,m} = -0.57 \times 10^{-6}$, $s_{m,g} = 0.86 \times 10^{-6}$, $s_g = 6.66 \times 10^{-6}$). Στη συνέχεια ακολουθεί ανάλυση ευαισθησίας του (Σ1) με μεθόδους αριθμητικής ανάλυσης γύρω από τη λύση που προέκυψε για τη μετοχή Google ως προς τις μεταβλητές $\gamma_g, \delta_g, s_x, s_y$. Συγκεκριμένα, μεταβάλλουμε την παράμετρο s_m στο διάστημα $[-4 \times 10^{-6}, 4 \times 10^{-6}]$ και διατηρώντας σταθερές τις υπόλοιπες τρεις παραμέτρους του (Σ1) ($s_{g,m} = -0.57 \times 10^{-6}$, $s_{m,g} = 0.86 \times 10^{-6}$, $s_g = 6.66 \times 10^{-6}$), έχουμε τα αποτελέσματα του Πίνακα 2 σχετικά με τις αλλαγές που επιφέρει αυτή η μεταβολή στη λύση του (Σ1).

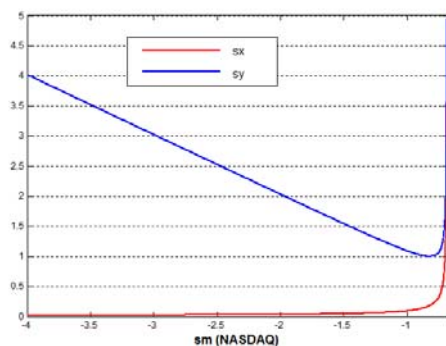
Πίνακας 2. Οι τιμές των $\gamma_g, \delta_g, s_x, s_y$ όταν $s_m \in [-4 \times 10^{-6}, 4 \times 10^{-6}]$ για τη Google.

Διάστημα	γ_g	δ_g	s_x	s_y
$[-4 \times 10^{-6}, -0.6860 \times 10^{-6}]$	> 0	> 0	> 0	> 0
$[-0.6850 \times 10^{-6}, 0.1860 \times 10^{-6}]$	το σύστημα	δίνει	μιγαδικές	ρίζες.

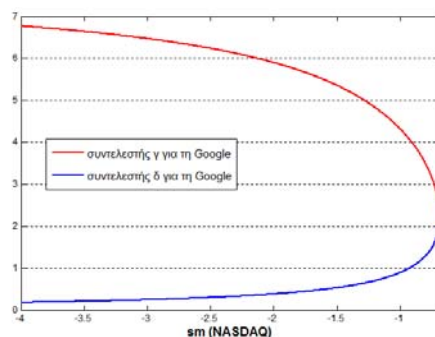
$[0.1870 \times 10^{-6}, 0.3770 \times 10^{-6}]$	< 0	< 0	> 0	> 0
$[0.3780 \times 10^{-6}, 4 \times 10^{-6}]$	το σύστημα	δίνει	λύσεις με	$s_y < 0$

Κατόπιν, τα γραφήματα που ακολουθούν αφορούν στο διάστημα $[-4 \times 10^{-6}, -0.6860 \times 10^{-6}]$, όπου $\gamma_g, \delta_g, s_x, s_y > 0$, καθώς αρνητικές τιμές των s_x, s_y στερούνται φυσικής σημασίας, ενώ αρνητικές τιμές των γ_g, δ_g υποδηλώνουν ότι η απόδοση της μετοχής κινείται προς την αντίθετη κατεύθυνση σε σχέση με την απόδοση της αγοράς, φαινόμενο που δεν απαντάται συχνά.

Σχήμα 1.(α) Μεταβολή των s_x, s_y σε σχέση με την παράμετρο s_m στο διάστημα $[-4 \times 10^{-6}, -0.6860 \times 10^{-6}]$. (β) Μεταβολή των γ_g, δ_g σε σχέση με την παράμετρο s_m στο διάστημα $[-4 \times 10^{-6}, -0.6860 \times 10^{-6}]$ για τη μετοχή Google (οι τιμές για τις τρίτες ροπές στους άξονες είναι πολλαπλασιασμένες επί 10^6).



(α)



(β)

Από το Σχήμα 1(α) προκύπτει ότι ο συνδυασμός της πολύ θετικής τιμής της τρίτης ροπής της μετοχής Google ($s_g = 6.66 \times 10^{-6}$) με αρκετά αρνητική τιμή της τρίτης ροπής της αγοράς ($s_m < -1.5 \times 10^{-6}$), υποδηλώνει πως πολύ μικρό μέρος της συνολικής διακύμανσης της αγοράς μπορεί να ερμηνευτεί από τις καλές ειδήσεις. Αντιθέτως, η μεταβλητή Y που αφορά στα αρνητικά νέα, θα πρέπει να εξηγήει εξ ολοκλήρου την τρίτη ροπή της αγοράς.

Με βάση το Σχήμα 1(β), λόγω της υψηλής –θετικής– τιμής της τρίτης ροπής ($s_g = 6.66 \times 10^{-6}$) που παρουσιάζει η μετοχή Google και με δεδομένη την αρνητική τιμή της τρίτης ροπής της αγοράς ($s_m \in [-4 \times 10^{-6}, -0.6860 \times 10^{-6}]$) προκύπτει μεγάλη ασυμμετρία στους συντελεστές γ_g και δ_g . Όταν η τιμή της τρίτης ροπής της αγοράς

είναι αρκετά αρνητική ($s_m < -1.5 \times 10^{-6}$), ο συντελεστής γ_g έχει αυξημένες τιμές. Δηλαδή, η απόδοση που προκύπτει από τις καλές ειδήσεις της αγοράς εισφέρει πολλαπλασιαστικά στην απόδοση της μετοχής Google. Αφετέρου, ο συντελεστής μεταφοράς δ_g της τιμής της μεταβλητής Y που αντιστοιχεί στις αρνητικές ειδήσεις είναι πολύ μικρός για αρκετά αρνητική τιμή της τρίτης ροπής της αγοράς και αυξάνεται αργά όσο αυξάνεται η τρίτη ροπή. Η περίπτωση $\gamma_g = \delta_g$ αποκλείεται, αφού από το (Σ1) εύκολα οδηγούμαστε σε άτοπο, αν λάβουμε υπόψη ότι $s_m < 0$, $s_{g,m} < 0$, ενώ $s_{m,g} > 0$ και $s_g > 0$.

Τα αλγεβρικά αποτελέσματα που ακολουθούν επεξηγούν τα παραπάνω γραφήματα. Πιο συγκεκριμένα, από τη δεύτερη εξίσωση του (Σ1) έχουμε ότι

$$\gamma_g = \frac{s_{g,m} + \delta_g s_y}{s_x} \quad (4)$$

Αντικαθιστώντας την (4) στην τρίτη εξίσωση του (Σ1) καταλήγουμε στην παρακάτω δευτεροβάθμια εξίσωση ως προς δ_g

$$s_y(s_y - s_x)\delta_g^2 + 2s_{g,m}s_y\delta_g + s_{g,m}^2 - s_{m,g}s_x = 0, \quad (5)$$

της οποίας η διακρίνουσα δίνεται από τη σχέση $\Delta = 4s_x s_y (s_{g,m}^2 - s_{m,g}s_m)$. Προκειμένου να έχει πραγματική λύση η (5), θα πρέπει $\Delta > 0$, δηλαδή $s_{g,m}^2 > s_{m,g}s_m$ εφόσον $s_x, s_y > 0$. Για $s_{g,m} = -0.57 \times 10^{-6}$ και $s_{m,g} = 0.86 \times 10^{-6}$ προκύπτει ότι $s_m < 0.3780 \times 10^{-6}$. Η συνθήκη αυτή είναι αναγκαία για να έχει λύση η εξίσωση (5) και κατ' επέκταση το (Σ1), ωστόσο δεν είναι απαραίτητα ικανή. Αυτό συμβαίνει διότι ο συγκεκριμένος περιορισμός διαμορφώθηκε με βάση τις τρεις πρώτες εξισώσεις του (Σ1) χωρίς να ληφθεί υπόψη η τέταρτη εξίσωση που αφορά την τρίτη ροπή της μετοχής. Σε κάθε περίπτωση το σύστημα δεν μπορεί να έχει πραγματική λύση για μια μετοχή που παρουσιάζει πολύ υψηλή τιμή τρίτης ροπής ($s_i \gg 0$) με ταυτόχρονα σημαντικά αρνητική τιμή της τρίτης ροπής της αγοράς. Ένας τρόπος για να αντιμετωπίσουμε το πρόβλημα αυτό θα ήταν να αποδεχτούμε την ύπαρξη ιδιοσυγκρατικής τρίτης ροπής για τη μετοχή και να εισάγουμε έναν όρο που να την αντιπροσωπεύει στην τέταρτη εξίσωση του (Σ1).

Στα Σχήματα 1(α) και 1(β) φαίνεται επίσης ότι πέραν της αναγκαίας συνθήκης $s_m < 0.3780 \times 10^{-6}$ για να έχει λύση το (Σ1), η λύση αυτή θα βρίσκεται έξω από την περιοχή του $s_m = 0$. Αυτό συμβαίνει διότι σε αυτήν την περιοχή είναι $s_x \cong s_y$. Τότε σύμφωνα με τη δεύτερη εξίσωση του (Σ1) έχουμε

$$-0.57 \times 10^{-6} \cong (\gamma_g - \delta_g) s_y \Rightarrow \gamma_g < \delta_g ,$$

εφόσον $s_y > 0$. Χρησιμοποιώντας όμως την τέταρτη εξίσωση του (Σ1) έχουμε ότι

$$6.66 \times 10^{-6} \cong (\gamma_g^3 - \delta_g^3) s_y \Rightarrow \gamma_g > \delta_g ,$$

καταλήγοντας επομένως σε άτοπο. Τελικά, όπως φαίνεται και στον Πίνακα 2, το (Σ1) δίνει λύσεις για $\gamma_g, \delta_g, s_x, s_y > 0$, όταν $-4 \times 10^{-6} \leq s_m \leq -0.6860 \times 10^{-6}$.

Η λύση της (5) δίνεται από τον τύπο

$$\delta_g = \frac{s_{g,m} - \left(\frac{s_x}{s_y} \right)^{1/2} \sqrt{s_{g,m}^2 - s_{m,g} s_m}}{s_m} , \quad (6)$$

ενώ αντικαθιστώντας τη σχέση (6) στην (4) προκύπτει ότι

$$\gamma_g = \frac{s_{g,m} - \left(\frac{s_y}{s_x} \right)^{1/2} \sqrt{s_{g,m}^2 - s_{m,g} s_m}}{s_m} . \quad (7)$$

Στο Σχήμα 1(α) παρατηρούμε ότι $\lim_{s_m \uparrow -0.6860 \times 10^{-6}} s_x = \lim_{s_m \uparrow -0.6860 \times 10^{-6}} s_y = \infty$ και χρησιμοποιώντας την πρώτη εξίσωση του (Σ1) έχουμε ότι

$$\lim_{s_m \uparrow -0.6860 \times 10^{-6}} \frac{s_x}{s_y} = \lim_{s_m \uparrow -0.6860 \times 10^{-6}} \frac{s_m + s_y}{s_y} = 1 .$$

Τότε, από τις (6) και (7) προκύπτει ότι

$$\lim_{s_m \uparrow -0.6860 \times 10^{-6}} \gamma_g = \lim_{s_m \uparrow -0.6860 \times 10^{-6}} \delta_g = 2.225 . \quad (8)$$

4. ΠΡΟΣΟΜΟΙΩΣΕΙΣ

Στη συνέχεια ακολουθούν προσομοιώσεις με βάση τις σχέσεις (1), (2) και (3) για την εκτίμηση των γ_g και δ_g στην περίπτωση της μετοχής Google. Σημειώνεται ότι σύμφωνα με τον Πίνακα 1 είναι $\gamma_g = 4.1$ και $\delta_g = 1$. Για τις προσομοιώσεις θεωρείται ότι $a_g = 0$, $a_m = 0$ και ότι οι διαδικασίες $X(t)$ και $Y(t)$ ακολουθούν Γάμμα κατανομές $\Gamma(k_x, \theta_x)$ και $\Gamma(k_y, \theta_y)$ αντίστοιχα, υπόθεση που απαντάται στη σχετική βιβλιογραφία για μοντελοποίηση αλμάτων στις τιμές. Για την εύρεση των τιμών των παραμέτρων των Γάμμα κατανομών παίρνουμε από τις (1) και (2) και με βάση τον Πίνακα 1 ότι

$$\sigma_m^2 = \sigma_x^2 + \sigma_y^2 \Rightarrow 290 \times 10^{-6} = \sigma_x^2 + \sigma_y^2 \quad (9)$$

$$\sigma_g^2 = \gamma_g^2 \sigma_x^2 + \delta_g^2 \sigma_y^2 + q_g^2 \Rightarrow 625 \times 10^{-6} = \gamma_g^2 \sigma_x^2 + \delta_g^2 \sigma_y^2 + q_g^2 \quad (10)$$

όπου,

σ_m^2, σ_g^2 : οι διασπορές των αποδόσεων της αγοράς και της μετοχής Google αντίστοιχα,

σ_x^2, σ_y^2 : οι διασπορές των αλμάτων στις αποδόσεις της αγοράς λόγω των θετικών και αρνητικών ειδήσεων αντίστοιχα.

Στη συνέχεια παρατίθενται προσομοιώσεις για δύο διαφορετικές τιμές του σ_x^2 , που αντιπροσωπεύουν με βάση τη (10) δύο διαφορετικά επίπεδα θορύβου, προκειμένου να ελεγχθεί αν τα προκύπτοντα σφάλματα των γ_g και δ_g ακολουθούν κανονικές κατανομές. Αρχικά παρατηρούμε ότι αντικαθιστώντας από την (9) το σ_y^2 στη (10) προκύπτει ότι $\sigma_x^2 < 21.19 \times 10^{-6}$.

(α) Θεωρούμε ότι $\sigma_x^2 = 21 \times 10^{-6}$, οπότε από την (9) προκύπτει ότι $\sigma_y^2 = 269 \times 10^{-6}$. Όσον αφορά τη διασπορά του θορύβου, δεδομένου ότι $\gamma_g = 4.1$, $\delta_g = 1$, $\sigma_x^2 = 21 \times 10^{-6}$ και $\sigma_y^2 = 269 \times 10^{-6}$, από τη (10) προκύπτει ότι $q_g^2 = 2.99 \times 10^{-6}$. Έτσι, για τον θόρυβο ισχύει $\varepsilon_g(t) \sim N(0, 2.99 \times 10^{-6})$ (χαμηλό επίπεδο θορύβου). Στη συνέχεια λαμβάνουμε υπόψη ότι για τη Γάμμα κατανομή ισχύει

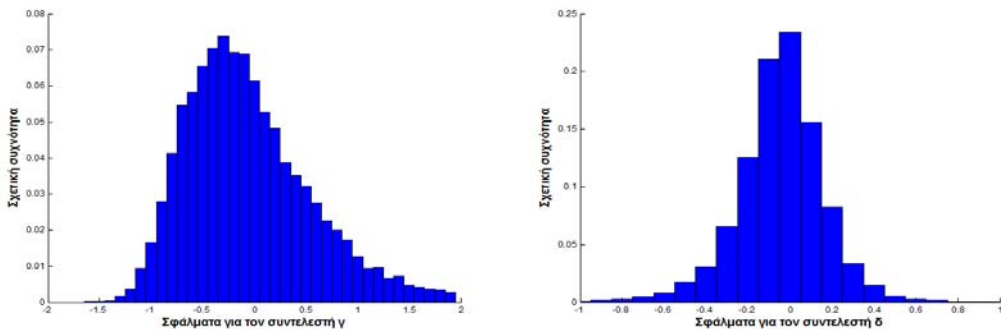
$$\sigma_x^2 = k_x \theta_x^2 \text{ και } s_x = 2k_x \theta_x^3. \quad (11)$$

Τότε προκύπτει

$$s_x = 2\sigma_x^2 \theta_x. \quad (12)$$

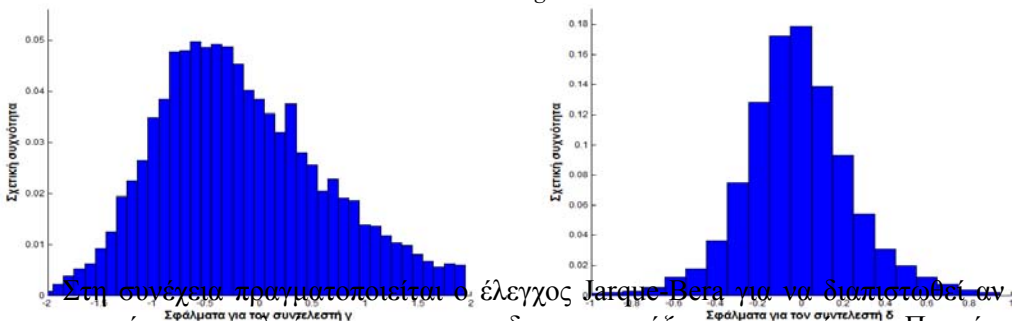
Από τον Πίνακα 1 έχουμε ότι $s_x = 0.11 \times 10^{-6}$ και αντικαθιστώντας την τιμή αυτή στη (12), δεδομένου ότι $\sigma_x^2 = 21 \times 10^{-6}$, προκύπτει ότι $\theta_x = 2.62 \times 10^{-3}$. Κατόπιν, από την (11) έχουμε ότι $k_x = 3.06$ και συνεπώς για την τυχαία μεταβλητή $X(t)$ ισχύει $X(t) \sim \Gamma(3.06, 2.62 \times 10^{-3})$. Όμοια προκύπτει ότι $Y(t) \sim \Gamma(71.99, 1.93 \times 10^{-3})$.

Σχήμα 2. Σφάλματα εκτίμησης των συντελεστών γ_g και δ_g αντίστοιχα, με βάση την προσομοίωση για τη μετοχή Google, όταν $\varepsilon_g(t) \sim N(0, 2.99 \times 10^{-6})$.



(β) Θεωρούμε ότι $\sigma_x^2 = 10 \times 10^{-6}$, οπότε προκύπτει $X(t) \sim \Gamma(0.33, 5.5 \times 10^{-3})$, $Y(t) \sim \Gamma(81.18, 1.86 \times 10^{-3})$ και $\varepsilon_g(t) \sim N(0, 176.9 \times 10^{-6})$ (υψηλό επίπεδο θορύβου).

Σχήμα 3. Σφάλματα εκτίμησης των συντελεστών γ_g και δ_g αντίστοιχα, με βάση την προσομοίωση για τη μετοχή Google, όταν $\varepsilon_g(t) \sim N(0, 176.9 \times 10^{-6})$.



Στη συνέχεια πραγματοποιείται ο έλεγχος Jarque-Bera για να διαπιστωθεί αν οι κατανομές των σφαλμάτων των γ_g και δ_g παρουσιάζουν κανονικότητα. Προκύπτει, όπως φαίνεται στον Πίνακα 3, ότι δεν παρουσιάζεται κανονικότητα σε καμία από τις κατανομές των σφαλμάτων (σε 5% επίπεδο σημαντικότητας).

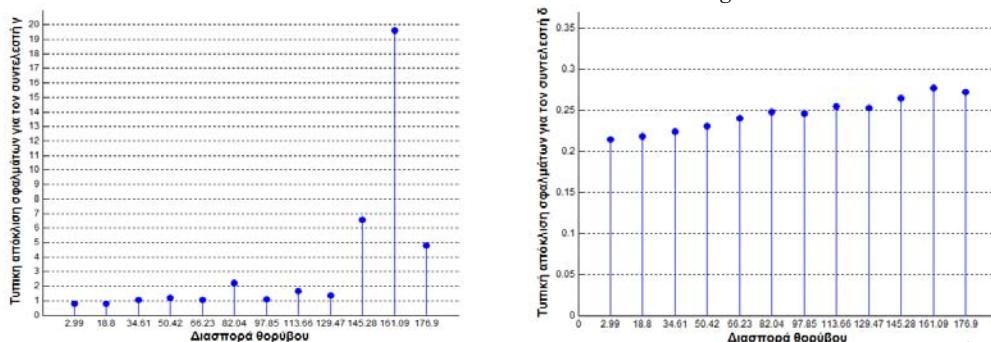
Πίνακας 3. Έλεγχος κανονικότητας.

	γ_g	δ_g
Jar-Bera (p-value)	$3.0829 \times 10^6 (< 0.001)$	$1.0545 \times 10^4 (< 0.001)$
	$1.5682 \times 10^9 (< 0.001)$	$3.6622 \times 10^3 (< 0.001)$

Επιπλέον, από το Σχήμα 4 που ακολουθεί παρατηρείται ότι καθώς αυξάνεται η διασπορά του γκαουσιανού θορύβου στις προσομοιώσεις, αυξάνεται (όχι μονότονα)

και η διασπορά στις κατανομές των σφαλμάτων εκτίμησης των συντελεστών γ_g και δ_g . Ενδεικτικά δίνονται τα παρακάτω γραφήματα που αποτυπώνουν τη μεταβολή στην τυπική απόκλιση των σφαλμάτων για αυξανόμενες τιμές της διασποράς του θορύβου.

Σχήμα 4. Τυπική απόκλιση των σφαλμάτων εκτίμησης των συντελεστών γ_g και δ_g αντίστοιχα για διάφορες τιμές της διασποράς του θορύβου $\varepsilon_g(t)$.



Σημείωση: Οι τιμές των διασπορών του θορύβου είναι οι αναγραφόμενες $\times 10^{-6}$.

5. ΣΥΜΠΕΡΑΣΜΑΤΑ

Εξετάστηκε ως προς την ευαισθησία μοντέλο σύμφωνα με το οποίο οι αποδόσεις μετοχών περιλαμβάνουν άλματα και θόρυβο. Διαπιστώθηκε η ευαισθησία των παραμέτρων ως προς την τρίτη (κεντρική) ροπή της αγοράς για τη μετοχή Google, ενώ αναδείχθηκαν και τα όρια μέσα στα οποία πρέπει να κινούνται η τρίτη ροπή της μετοχής και η συλλοξότητα με την αγορά, ώστε το (Σ1) να έχει πραγματικές λύσεις. Στην εργασία καταδεικνύεται ότι το (Σ1) δεν μπορεί να έχει πραγματική λύση για μια μετοχή που παρουσιάζει πολύ υψηλή τιμή τρίτης ροπής με ταυτόχρονα σημαντικά αρνητική τιμή τρίτης ροπής της αγοράς, και θα πρέπει το μοντέλο Polimenis (2012) να επεκταθεί ώστε να συμπεριλάβει την ύπαρξη ιδιοσυγκρατικής τρίτης ροπής για τη μετοχή. Το (Σ1) μπορεί να ερμηνεύσει τη θετική τιμή της τρίτης ροπής της μετοχής σε συνδυασμό με την αρνητική τιμή της τρίτης ροπής της αγοράς μόνο με επαρκή βαθμό ασυμμετρίας των συντελεστών γ_g και δ_g . Τέλος, προκύπτει ότι τα σφάλματα εκτίμησης των συντελεστών γ_g και δ_g με βάση τις προσομοιώσεις δεν ακολουθούν κανονικές κατανομές, ενώ παρατηρήθηκε ότι αυξάνεται η διασπορά τους, καθώς αυξάνεται η διασπορά του θορύβου.

ABSTRACT

In this study, we present the results of investigating a stochastic model which aims to distinguish true information from noise included in the stock price evolution. In particular, we assume that stock price is influenced not only by noise which carries no information and produces continuous paths, but also by jumps which occur when information arrives in the market. Provided that the jumps due to good or bad news are unobservable, the model proposes the use of the third central moments for the estimation of jump risk.

ΑΝΑΦΟΡΕΣ

- Bawa V. and Lindenberg E. (1977). Capital market equilibrium in a mean-lower partial moment framework, *J. Financ. Econ.* 5, p.189–200.
- Carr P., Geman H., Madan D.B., and Yor M. (2002). The fine structure of asset returns: An empirical investigation, *J. Bus.* 75, p.305–332.
- Carr P. and Wu L. (2004). Time-changed Levy processes and option pricing, *J. Financ. Econ.* 71, p.113–141.
- Harlow V. and Rao R. (1989). Asset pricing in a generalized mean-lower partial moment framework: Theory and evidence, *J. Financ. Quant. Anal.* 24, p.285–311.
- Harvey C. and Siddique A. (2000). Conditional skewness in asset pricing tests, *J. Finance* 55, p.1263–1295.
- Kraus A. and Litzenberger R. (1976). Skewness preference and the valuation of risk assets, *J. Finance* 31, p.1085–1100.
- Merton R. (1976). Option pricing when underlying stock returns are discontinuous, *J. Financ. Econ.* 3, p.125–144.
- Pan J. (2002). The jump-risk premia implicit in options: Evidence from an integrated time-series study, *J. Financ. Econ.* 63, p.3–50.
- Polimenis V. (2012). Information arrival as price jumps, *Optimization: A Journal of Mathematical Programming and Operations Research*, 61:10, p.1179–1190.



ΈΝΑ ΜΟΝΤΕΛΟ ΔΙΑΤΑΞΙΜΗΣ ΛΟΓΙΣΤΙΚΗΣ ΠΑΛΙΝΔΡΟΜΗΣΗΣ ΓΙΑ ΤΗ ΜΕΛΕΤΗ ΠΑΡΑΓΟΝΤΩΝ ΣΥΝΔΕΟΜΕΝΩΝ ΜΕ ΤΗ ΔΙΑΡΚΕΙΑ ΣΠΟΥΔΩΝ: ΜΙΑ ΕΦΑΡΜΟΓΗ ΣΕ ΔΕΔΟΜΕΝΑ ΤΟΥ ΠΑΝΤΕΙΟΥ ΠΑΝΕΠΙΣΤΗΜΙΟΥ

Δημήτριος Α. Καλαμάρας, Αγλαΐα Γ. Καλαματιανού

Τμήμα Κοινωνιολογίας

Πάντειο Πανεπιστήμιο Κοινωνικών και Πολιτικών Επιστημών

{dkalam, akalam}@panteion.gr

Περίληψη

Ζητήματα σχετικά με τη χρονική διάρκεια των πανεπιστημιακών σπουδών που παραπέμπουν σε θέματα παράτασης ή εγκατάλειψης των σπουδών, σε συνδυασμό με τους παράγοντες που συνδέονται με αυτά, έχουν προσελκύσει το ενδιαφέρον πολλών ερευνητών και από διαφορετικά επιστημονικά πεδία, σχεδόν από τα μέσα του 20 αιώνα. Στη μελέτη τέτοιων ζητημάτων η εμπλοκή στατιστικών μεθόδων είναι σχεδόν πάντα αναγκαία. Σε αυτό το πλαίσιο, για την ελληνική πραγματικότητα έχει προταθεί μια στατιστική μεθοδολογία από τη θεωρία της ανάλυσης επιβίωσης, με την οποία εκτιμάται η κατανομή της διάρκειας σπουδών φοιτητών κοινωνικών επιστημών. Στην παρούσα εργασία, δίνονται αποτελέσματα για τους παράγοντες που συνδέονται με αυτή τη κατανομή στη βάση ενός σχετικού κοινωνιολογικού προτύπου και με τη βοήθεια ενός μοντέλου διατάξιμης λογιστικής παλινδρόμησης.

1. ΕΙΣΑΓΩΓΗ: ΘΕΩΡΗΤΙΚΕΣ ΠΡΟΣΕΓΓΙΣΕΙΣ – ΤΟ ΖΗΤΟΥΜΕΝΟ ΤΗΣ ΕΡΓΑΣΙΑΣ

Το φαινόμενο της μη ολοκλήρωσης των σπουδών, στο αντικειμενικά προβλεπόμενο χρονικό διάστημα, από έναν αριθμό φοιτητών ενίοτε σημαντικό, μελετάται στη διεθνή βιβλιογραφία για την ανώτατη εκπαίδευση από τα μέσα του 20ου αιώνα, όπου συνήθως εξειδικεύεται είτε ως καθυστέρηση στην αποφοίτηση (χρονική παράταση σπουδών) είτε ως εγκατάλειψη των σπουδών (βλ. Eckland: 1964, Spady: 1970, De Francesco και Corrado:1977, Bayer: 1968, μεταξύ άλλων).

Για τη μελέτη του φαινομένου έχουν αναπτυχθεί διάφορες προσεγγίσεις: Στις πρώτες μελέτες της δεκαετίας του 1930 στις Ηνωμένες Πολιτείες, απλά καταγράφεται η αποτυχία ορισμένων φοιτητών να ολοκληρώσουν τις σπουδές τους η οποία αναφέρεται ως «φοιτητική θνησιμότητα», όρος που διατηρήθηκε ως τις αρχές της δεκαετίας του 1960 (Berger και Lyon: 2005). Ακολούθως εμφανίζονται οι πρώτες πιο σύνθετες προσεγγίσεις/ερμηνείες του φαινομένου που μπορούν να διακριθούν σε ψυχολογικές και κοινωνιολογικές. Αναλυτικότερα: Σύμφωνα με τις ψυχολογικές προσεγγίσεις (Astin: 1975), η απόφαση για εγκατάλειψη ή μη των σπουδών εξαρτάται αποκλειστικά από την «ποσότητα» της σωματικής και ψυχολογικής ενέργειας που ο φοιτητής αφιερώνει στις ακαδημαϊκές του εργασίες. Οι προσεγγίσεις αυτές επικεντρώνονται κυρίως στη μελέτη του φαινομένου σε σχέση με άμεσα μετρήσιμα χαρακτηριστικά των φοιτητών (όπως φύλο, ηλικία, καταγωγή κ.ά.) αλλά και του ακαδημαϊκού ιδρύματος φοίτησης (όπως τύπος ιδρύματος, χωροταξική θέση κ.ά). Σύμφωνα με τις κοινωνιολογικές προσεγγίσεις, η εγκατάλειψη των σπουδών οφείλεται στη μη «ενσωμάτωση» των φοιτητών στο ακαδημαϊκό και στο κοινωνικό περιβάλλον του ιδρύματος φοίτησης (student integration). Κύριος εκφραστής αυτών των προσεγγίσεων είναι ο Vincent Tinto (Tinto: 1975), που εξηγεί τη διαδικασία εγκατάλειψης των σπουδών στη βάση της θεωρίας περί «ενσωμάτωσης». Η βασική παραδοχή του μοντέλου του Tinto είναι ότι η απόφαση για εγκατάλειψη ή μη των σπουδών επηρεάζεται από τα ατομικά χαρακτηριστικά του φοιτητή σε συνδυασμό με τις δεσμεύσεις που έχει αναλάβει ως προς το ακαδημαϊκό ίδρυμα που εισήχθη. Η έγκαιρη και συνεχής αυτή δέσμευση επηρεάζει την «ενσωμάτωση» του φοιτητή σε ότι αυτός ονόμασε κοινωνικό και ακαδημαϊκό σύστημα του πανεπιστημίου. Οι δύο αυτοί τύποι «ενσωμάτωσης» θεωρούνται ως οι σημαντικότεροι παράγοντες για την παραμονή ή όχι των φοιτητών στο ακαδημαϊκό ίδρυμα και την αποφυγή της εγκατάλειψης των σπουδών.

Παράλληλα και εμπνευσμένες από τις κοινωνιολογικές προσεγγίσεις, αναπτύχθηκαν και προσεγγίσεις που εμπλέκουν παραμέτρους από τη θεωρία περί ικανοποίησης από την εργασία (job satisfaction). Σύμφωνα με αυτές, η εγκατάλειψη των σπουδών οφείλεται στην έλλειψη ικανοποίησης των φοιτητών από τις ακαδημαϊκές τους επιδόσεις και τις πιθανές προοπτικές επαγγελματικής αποκατάστασης που θα προσέφερε η ολοκλήρωση των σπουδών (Bean: 1980).

Η τάση που παρατηρείται τα τελευταία χρόνια στη μελέτη του φαινομένου είναι η ανάπτυξη πιο ολοκληρωμένων προτύπων όπου οι προαναφερόμενες θεωρίες δεν εκλαμβάνονται ως αμοιβαία αποκλειόμενες αλλά ως συμπληρωματικές (Cabreria κ.α.: 1990, 1993), υπό την έννοια, ότι η μελέτη του φαινομένου πρέπει να στηρίζεται στην από κοινού εξέταση ατομικών χαρακτηριστικών των φοιτητών, κοινωνικών παραγόντων και ακαδημαϊκών συνθηκών.

Ωστόσο, ανεξαρτήτως θεωρητικού πλαισίου προσέγγισης του φαινομένου, η εμπλοκή στατιστικών μεθόδων είναι πάντα αναγκαία, για την εμπειρική διερεύνηση των

υποθέσεων που τίθενται. (βλ. Cabrera κ.α.: 1993, Webster και Showers: 2011, Amraw και Jaeger: 2012 μεταξύ άλλων). Έτσι για παράδειγμα, η χρήση των μεθόδων της λογιστικής παλινδρόμησης και της διαχωριστικής ανάλυσης είναι εξαιρετικά συχνή στον προσδιορισμό παραγόντων συνδεόμενων με την εγκατάλειψη των σπουδών ή την παρατεταμένη φοίτηση (βλ. Aina: 2005, Cingano και Cipollone: 2007, Cabrera κ.α.: 1993, Webster και Showers: 2011 μεταξύ άλλων). Επίσης αποδεικνύεται ότι μέθοδοι από τη θεωρία της ανάλυσης επιβίωσης (βλ. Murtaugh και Schuster: 1999, Kalamatianou: 1999, Kalamatianou και McClean: 2003) και των μαρκοβιανών διαδικασιών (Herrera: 2006) αποτελούν χρήσιμα εργαλεία στη μελέτη της εξέλιξης των σπουδών των φοιτητών και της διάρκειας φοίτησης.

Στο σημείο αυτό είναι σημαντικό να αναφερθεί ότι οι διαφορετικές προϋποθέσεις που θέτει το εκπαιδευτικό σύστημα κάθε χώρας, για την εγγραφή και την ολοκλήρωση των σπουδών στην τριτοβάθμια εκπαίδευση, διαμορφώνουν διαφορετικά πλαίσια για την εξέλιξη του φαινομένου της παρατεταμένης φοίτησης ή εγκατάλειψης των σπουδών και συνεπώς και για τον τρόπο καταγραφής και ανάλυσής του.

Στην ελληνική πραγματικότητα, μια συστηματική προσέγγιση του θέματος της διάρκειας σπουδών, σε συγκεκριμένα αντικείμενα κοινωνικών σπουδών και στη βάση μεθοδολογίας από το χώρο της ανάλυσης επιβίωσης, περιλαμβάνεται στις εργασίες των Kalamatianou, (1999) και Kalamatianou και McClean, (2003). Συγκεκριμένα εκτιμάται η κατανομή της διάρκειας σπουδών από όπου προκύπτουν τρεις διακριτές κατηγορίες φοιτητών: Φοιτητές που αποφοιτούν στο ελάχιστο δυνατό χρονικό διάστημα ή λίγο αργότερα, φοιτητές που αποφοιτούν μετά από σχετικά παρατεταμένη φοίτηση -βραδεία αποφοίτηση- και φοιτητές που φαίνεται να μην αποφοιτούν -αιώνια φοίτηση-. Τα αποτελέσματα αυτά επιβεβαιώνονται και με νεότερα δεδομένα στην εργασία των Καλαμάρας και Καλαματιανού (2006). Στο πλαίσιο αυτών των αποτελεσμάτων προκύπτει εύλογα το ερώτημα σχετικά με τους παράγοντες που συνδέονται ή και προκαλούν την προαναφερόμενη κατανομή. Στην παρούσα εργασία, δίνονται μερικές πρώτες απαντήσεις σε αυτό το ερώτημα στη βάση ενός μοντέλου διατάξιμης λογιστικής παλινδρόμησης. Συνοπτικά εδώ, επεκτείνοντας τις υποθέσεις του Tinto έτσι ώστε να ενσωματωθούν οι ιδιαιτερότητες της ελληνικής πραγματικότητας, θεωρείται ότι η διάρκεια φοίτησης (μεταβλητή απόκρισης) συνδέεται με πέντε βασικές κατηγορίες παραγόντων που αφορούν σε δημογραφικά και σε ακαδημαϊκά χαρακτηριστικά του φοιτητή, σε κίνητρα και προσδοκίες σχετικά με το αντικείμενο σπουδών, σε εκτός ακαδημαϊκού πλαισίου -εξωγενή χαρακτηριστικά, και σε στάσεις του φοιτητή ως προς το θέμα της διάρκειας σπουδών. Τα χαρακτηριστικά αυτά (επεξηγηματικές μεταβλητές) περιγράφονται λεπτομερέστερα στην επόμενη ενότητα.

Ακολούθως η οργάνωση της εργασίας εξελίσσεται ως εξής: Στην επόμενη ενότητα περιγράφονται τα δεδομένα, οι μεταβλητές και το μοντέλο της διατάξιμης λογιστικής παλινδρόμησης. Η ενότητα 3 είναι αφιερωμένη στην παρουσίαση των αποτελεσμά-

των από την εφαρμογή του μοντέλου στο σύνολο του δείγματος και στις υποομάδες ανδρών και γυναικών ενώ στην τελευταία ενότητα 4, συνοψίζονται τα κυριότερα συμπεράσματα της εργασίας.

2. ΤΑ ΔΕΔΟΜΕΝΑ, ΟΙ ΜΕΤΑΒΛΗΤΕΣ ΚΑΙ ΤΟ ΜΟΝΤΕΛΟ ΤΗΣ ΔΙΑΤΑΞΙΜΗΣ ΛΟΓΙΣΤΙΚΗΣ ΠΑΛΙΝΔΡΟΜΗΣΗΣ

Η εφαρμογή βασίζεται σε δεδομένα, που συλλέχτηκαν με τη βοήθεια ειδικά κατασκευασμένου ερωτηματολογίου και αντιστοιχούν σε ένα αντιπροσωπευτικό δείγμα 1236 φοιτητών και πτυχιούχων (στο εξής «ερωτώμενοι») που εισήχθησαν στο Πάντειο Πανεπιστήμιο σε 17 συνεχή ακαδημαϊκά έτη. Η περίοδος παρακολούθησης για την πρώτη κοόρτη αντιστοιχεί σε 278 μήνες και για την τελευταία σε 86 μήνες.

Η **αποκριτική (εξαρτημένη) μεταβλητή** αφορά στη διάρκεια σπουδών των ερωτώμενων, μετρούμενη αρχικά σε μήνες (βλ. Kalamatianou και McClean: 2003). Για τις ανάγκες της παρούσας εργασίας οι τιμές της αποκριτικής μεταβλητής κωδικοποιήθηκαν, έτσι ώστε να αντανακλούν τις χρονικές φάσεις φοίτησης/αποφοίτησης, όπως προσεγγίστηκαν από τους Kalamatianou, (1999), Kalamatianou και McClean, (2003) Καλαμάρας και Καλαματιανού (2006), που περιγράφηκαν στην προηγούμενη ενότητα. Η κωδικοποίηση αυτή έχει ως εξής:

“1” ≈ πτυχίο στην πρώτη (χρονικά) δυνατή ευκαιρία (48 μήνες)

“2” ≈ πτυχίο με διετή καθυστέρηση (από 49 έως 72 μήνες)

“3” ≈ πτυχίο με μεγάλη καθυστέρηση ή “Βραδεία Αποφοίτηση” (από 73 έως 186 μήνες)

“4” ≈ «αιώνια φοίτηση» ή “εγκατάλειψη των σπουδών” (μεγαλύτερη των 186 μηνών)

Είναι σαφές ότι οι παραπάνω τιμές μπορούν να τεθούν σε διάταξη -από την μικρότερη ως τη μεγαλύτερη χρονικά διάρκεια φοίτησης- και συνεπώς η αποκριτική μεταβλητή μπορεί να θεωρηθεί ως μια διατάξιμη μεταβλητή με τέσσερις τιμές.

Πίνακας 1. Οι επεξηγηματικές (ανεξάρτητες) μεταβλητές:

<i>Δημογραφικά χαρακτηριστικά</i>	<i>Συμμετοχή στις παρατάξεις (1≈ ναι, 2≈ όχι)</i>
-Ηλικία εισαγωγής στο Πανεπιστήμιο	-Συναναστροφή με συμφοιτητές (1≈ ναι, 2≈ όχι)
-Τόπος καταγωγής (1≈Αθήνα, 2≈εκτός Αθήνας)	
-Τρόπος εισαγωγής στο πανεπιστήμιο (1≈ εισαγωγικές εξετάσεις, 2≈ άλλος τρόπος)	<i>Κίνητρα και Προσδοκίες</i>
-Κοινωνικοοικονομικό επίπεδο γονέων (1≈ υψηλό, 2≈ χαμηλό)	-Λόγος επιλογής του αντικειμένου σπουδών (1≈ μου άρεσε, 2≈ έτυχε)
<i>Ακαδημαϊκά χαρακτηριστικά του φοιτητή</i>	-«Δίαιλοι» επιρροής στην επιλογή του αντικειμένου σπουδών (1≈ επηρέασαν τρίτοι, 2≈ επιλογή του φοιτητή)
-Μέσος όρος βαθμολογίας Γ΄ τάξης λυκείου	-Προσδοκία επαγγελματικής αποκατάστασης με την απόκτηση του πτυχίου (1≈ ναι, 2≈ όχι)

-Βαθμολογία γενικών (εισαγωγικών) εξετάσεων	-Προσδοκία απόκτησης δεξιοτήτων από τις σπουδές για την αγορά εργασίας
-Μέσος όρος βαθμολογίας Α' και Β' εξαμήνου σπουδών	(1 ≈ ναι, 2 ≈ όχι)
-Συχνότητα μελέτης (1 ≈ καθόλου, 2 ≈ κατά την εξεταστική περίοδο, 3 ≈ συνεχώς)	-Απόκτηση γνώσης κατά τη διάρκεια των σπουδών (1 ≈ ναι, 2 ≈ όχι)
-Παρακολούθηση μαθημάτων (1 ≈ ναι, 2 ≈ όχι)	-Απόκτηση κύρους, κατά τη διάρκεια των σπουδών (1 ≈ ναι, 2 ≈ όχι)
-Συμμετοχή στο μάθημα (διατύπωση ερωτήσεων, αποριών) (1 ≈ ναι, 2 ≈ όχι)	-Σειρά του τμήματος σπουδών στο μηχανογραφικό δελτίο των εισαγωγικών εξετάσεων
-Ενδιαφέρον για το αντικείμενο σπουδών (πριν την εισαγωγή στο Πανεπιστήμιο)	Εξωγενή χαρακτηριστικά
(1 ≈ ναι, 0 ≈ όχι)	-Κάποιο σημαντικό αρνητικό γεγονός συνέβη στη διάρκεια των σπουδών (1 ≈ ναι, 2 ≈ όχι)
-Ικανοποίηση από το πρόγραμμα σπουδών (1 ≈ ναι, 2 ≈ όχι)	-Εργασία κατά τη διάρκεια των σπουδών (1 ≈ ναι, 2 ≈ όχι)
-Ικανοποίηση από τις συνθήκες διεξαγωγής των μαθημάτων (1 ≈ ναι, 2 ≈ όχι)	Στάσεις και πεποιθήσεις
-Συμμετοχή στις εκδηλώσεις του πανεπιστημίου (1 ≈ ναι, 2 ≈ όχι)	-Αιώνια φοίτηση (1 ≈ υπέρ, 2 ≈ κατά)
-Συμμετοχή στις φοιτητικές εκλογές (1 ≈ ποτέ, 2 ≈ κάποιες φορές, 3 ≈ συχνά)	-Επιβολή ανώτατου ορίου φοίτησης (1 ≈ υπέρ, 2 ≈ κατά)

Επιπλέον καταγράφηκε και το φύλο των ερωτωμένων που στην ανάλυση χρησιμοποιείται ως μεταβλητή ελέγχου.

Για τη διερεύνηση της σχέσης μεταξύ της αποκριτικής μεταβλητής και των επεξηγηματικών μεταβλητών θεωρούμε το παρακάτω μοντέλο διατάξιμης λογιστικής παλινδρόμησης.

$$\text{link}(Y_{ij}) = \theta_j - (\beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_p X_{ip}) \quad (1)$$

όπου,

Y συμβολίζει γενικά την αποκριτική μεταβλητή οι τιμές της οποίας δηλώνονται με j Εδώ έχουμε $j = 1, 2, 3, 4$ και αντιστοιχούν ακριβώς στις τιμές που περιγράφονται παραπάνω. Η τιμή $j = 4$ επιλέχτηκε ως η κατηγορία αναφοράς.

X συμβολίζει γενικά τις επεξηγηματικές μεταβλητές, το πλήθος των οποίων δηλώνεται με p , εδώ $p = 28$ και αριθμούνται με $g = 1, 2, \dots, p$.

i δηλώνει τον κάθε ερωτώμενο, $i = 1, 2, \dots, n$, όπου n το πλήθος τους και $n = 1236$.

Y_{ij} η τιμή της αποκριτικής μεταβλητής στον i ερωτώμενο, i και j όπως παραπάνω.

X_{ig} η τιμή της g επεξηγηματικής μεταβλητής, στον i ερωτώμενο, i και g όπως παραπάνω

$\theta_j = \frac{P(Y \leq j)}{P(Y > j)}$, cumulative odds, στην τιμή j της αποκριτικής μεταβλητής, είναι

$$\theta_1 = \frac{P(Y \leq 1)}{P(Y > 1)}, \theta_2 = \frac{P(Y \leq 2)}{P(Y > 2)}, \theta_3 = \frac{P(Y \leq 3)}{P(Y > 3)}$$

$link(Y_{ij}) = \tan[\pi(P(Y \leq j)) - 0.5]$ η συνάρτηση σύνδεσης *Cauchit*.

Σημειώνουμε ότι η συνάρτηση σύνδεσης *Cauchit* (Morgan και Smith 1992) επιλέχτηκε εδώ διότι παρουσιάζει καλύτερη εφαρμογή στα δεδομένα σε σχέση με άλλες συναρτήσεις σύνδεσης.

3. ΤΑ ΑΠΟΤΕΛΕΣΜΑΤΑ ΤΗΣ ΠΑΛΙΝΔΡΟΜΗΣΗΣ

Το μοντέλο (1) εφαρμόστηκε (SPSS v. 20) στο σύνολο του δείγματος αλλά και στις ομάδες των ανδρών και των γυναικών. Στοιχεία για την καταλληλότητα του μοντέλου δίνονται στον Πίνακα 2. Τα αποτελέσματα σε ότι αφορά τις τιμές των συντελεστών της παλινδρόμησης δίνονται στον Πίνακα 3.

Σκιαγραφώντας τα κυριότερα ευρήματα της ανάλυσης, σημειώνουμε τα ακόλουθα. Με αναφορά στο σύνολο του δείγματος, τα χαρακτηριστικά που έχουν στατιστικά σημαντική επίδραση στη διάρκεια σπουδών είναι: α) από τα ακαδημαϊκά χαρακτηριστικά, ο μέσος όρος βαθμολογίας του πρώτου έτους σπουδών και η συμμετοχή στα μαθήματα β) από τα κίνητρα και τις προσδοκίες, η επαγγελματική αποκατάσταση και η απόκτηση δεξιοτήτων από το αντικείμενο σπουδών για την αγορά εργασίας και γ) τα εξωγενή χαρακτηριστικά «εργασία στη διάρκεια των σπουδών» και «ύπαρξη κάποιου αρνητικού γεγονότος». Οι δύο πρώτες κατηγορίες χαρακτηριστικών έχουν θετική επίδραση στη διάρκεια σπουδών υπό την έννοια ότι υψηλός μέσος όρος και καταφατικές απαντήσεις ως προς τη συμμετοχή στα μαθήματα, τα κίνητρα και τις προσδοκίες συνδέονται με χαμηλές πιθανότητες ένταξης των ερωτωμένων στη κατηγορία φοιτητών αιώνιας φοίτησης. Ενδεικτικά αναφέρουμε ότι η «τύχη» των φοιτητών, που συμμετέχουν στα μαθήματα, να ενταχθούν στους φοιτητές αιώνιας φοίτησης, εκτιμάται περίπου 4 φορές μικρότερη ($e^{-1.444} = 0.236 \approx 23.6\%$) εκείνης των φοιτητών που δεν συμμετέχουν στα μαθήματα. Αντίθετα, τα εξωγενή χαρακτηριστικά έχουν αρνητική επίδραση στη διάρκεια σπουδών υπό την έννοια, ότι καταφατικές απαντήσεις (παρουσία αυτών των χαρακτηριστικών) συνδέονται με μακρά διάρκεια σπουδών ή και αιώνια φοίτηση. Ειδικότερα, υπογραμμίζουμε την αρνητική επίδραση της εργασίας κατά τη διάρκεια των σπουδών, στο χρόνο αποφοίτησης, που ενδέχεται να συνδέεται ως ένα βαθμό με την μέχρι σήμερα ανυπαρξία σπουδών μερικής φοίτησης στα ελληνικά πανεπιστήμια.

Πίνακας 2. Έλεγχος καταλληλότητας του μοντέλου

		Σύνολο δείγματος		Ανδρες		Γυναίκες	
		-2 Log Lik.	Chi-Sq.	-2 Log Lik.	Chi-Sq.	-2 Log Lik.	Chi-Sq.
Έλεγχος	Null Hypothesis	442.569		134.729		192.317	

<i>General</i>		361.974	80.595	111.319	23.410	132.136	60.181
<i>β.ε.</i>		94		76		92	
<i>επίπεδο σημαντικότητας</i>		0.836		1.000		0.996	
Καταλλη-	<i>Μοντέλο μόνο με σταθερά</i>	650.262		298.319		345.905	
λότητα του	<i>Συνολικό μοντέλο</i>	442.569	207.692	134.729	163.591	192.317	153.589
μοντέλου	<i>β.ε.</i>	47		38		46	
	<i>επίπεδο σημαντικότητας</i>	0.000*		0.000*		0.000*	
Συντελεστές	<i>Cox & Snell</i>	0.528		0.733		0.634	
προσδιο-	<i>Nagelkeks</i>	0.583		0.805		0.707	
ρισμού	<i>McFadden</i>	0.319		0.548		0.444	

Τα παραπάνω ευρήματα σε ότι αφορά τα ακαδημαϊκά χαρακτηριστικά και εκείνα των κινήτρων και προσδοκιών διαπιστώνονται επίσης και στις υποομάδες των ανδρών και των γυναικών. Ωστόσο, στην περίπτωση των ανδρών προκύπτει ότι σημαντική θετική επίδραση στη διάρκεια σπουδών έχουν μερικά ακόμα ακαδημαϊκά χαρακτηριστικά όπως η ικανοποίηση από τις συνθήκες διεξαγωγής των μαθημάτων και η συμμετοχή στις εκδηλώσεις του πανεπιστημίου και στα «κοινά» -φοιτητικές εκλογές), από την κατηγορία των κινήτρων και προσδοκιών, οι «διάυλοι» επιρροής για την επιλογή του αντικειμένου σπουδών (όταν η επιλογή είναι απόφαση του ίδιου του φοιτητή) και τέλος ο τόπος καταγωγής (Αθήνα) που εντάσσεται στα δημογραφικά χαρακτηριστικά.

Πίνακας 3. Αποτελέσματα διατάξιμης λογιστικής παλινδρόμησης στο σύνολο του δείγματος και στις ομάδες ανδρών και γυναικών

<i>Επεξηγηματικές μεταβλητές**</i>	<i>Σύνολο δείγματος</i>		<i>Άνδρες</i>		<i>Γυναίκες</i>	
	<i>β</i>	<i>e^β</i>	<i>β</i>	<i>e^β</i>	<i>β</i>	<i>e^β</i>
Δημογραφικά χαρακτηριστικά						
<i>Ηλικία εισαγωγής στο Πανεπιστήμιο</i>	0.156	1.169	-0.829	0.436	2.022*	7.553
<i>Τόπος καταγωγής _ «Αθήνα»</i>	-0.08	0.923	-4.541*	0.011	4.274*	71.808
<i>Τρόπος εισαγωγής _ «Πανελλήνιες»</i>	-1.302	0.272	-8.772	0.000	4.562	95.775
<i>Κοινων.οικον. επίπεδο γονέων _ «Υψηλό»</i>	-0.428	0.652	0.381	1.464	0.070	1.073
Ακαδημαϊκά χαρακτηριστικά						
<i>Μέσος όρος βαθ. Γ' τάξης λυκείου</i>	-0.156	0.856	0.132	1.141	-0.305	0.737
<i>Βαθμολογία γενικών εξετάσεων</i>	0.922	2.514	-0.289	0.749	-12.159	0.000
<i>Μέσος όρος βαθ. Α' και Β' εξαμήνου</i>	-1.541*	0.214	-7.433*	0.001	-2.011*	0.134
<i>Συχν. μελέτης _ «Καθόλου»</i>	-586.678	0.000	0.000	1.000	-643.344	0.000
<i>_ «Κατά την εξεταστική»</i>	-0.468	0.626	-1.010	0.364	-0.051	0.950
<i>Παρακολούθηση μαθημάτων _ «ναι»</i>	0.715	2.044	4.163	64.264	-1.814	0.163
<i>Συμμετοχή στο μάθημα _ «ναι»</i>	-1.444*	0.236	-5.487*	0.004	-6.573*	0.001
<i>Ενδ. για το αντικ. σπουδών _ «ναι»</i>	-0.219	0.803	-0.939	0.391	2.299	9.964
<i>Ικαν. από πρόγραμμα σπουδών _ «ναι»</i>	0.455	1.576	3.428	30.815	1.291	3.636
<i>Ικαν. από συν. διεξ. μαθημάτων _ «ναι»</i>	-0.915	0.401	-5.597*	0.004	-1.279	0.278
<i>Συμ. στις εκδηλώσεις του πανεπιστημίου _ «ναι»</i>	0.062	1.604	-4.198*	0.015	0.482	1.619
<i>Συμ. στις φοιτητικές εκλογές _ «ποτέ»</i>	0.504	1.655	6.098*	444.967	1.665	5.286
<i>_ «κάποιες φορές»</i>	-0.211	0.810	2.653	14.197	4.971*	144.171

Συμμετοχή στις παρατάξεις_«ναι»	0.195	1.215	9.139*	9311.45	-6.545*	0.001
Συναναστροφή με συμφοιτητές_«ναι»	0.136	1.146	-1.323	0.266	-8.782*	0.000
Κίνητρα και Προσδοκίες						
Λόγος επιλογής αντικ. σπουδών_«μου άρεσε»	0.931	2.537	0.103	1.108	1.163	3.200
«Διάλυλοι» επιρ. αντικ. σπουδών_«επιρ. τρίτων»	0.727	2.069	4.839*	126.343	3.676*	39.488
Προσδοκία επαγγελματικής αποκατάστασης με την απόκτηση του πτυχίου_«ναι»	500.068*	0.000	-877.24*	0.000	-602.08*	0.000
Απόκτηση δεξιοτήτων από τις σπουδές για την αγορά εργασίας_«ναι»	-2.822*	0.059	-11.319*	0.000	-10.235*	0.000
Απόκτηση. Γνώσης από τις σπουδές_«ναι»	-1.281*	0.278	-0.311	0.733	-6.666*	0.001
Απόκτηση κύρους από τις σπουδές_«ναι»	0.523	1.687	-2.756	0.064	4.235*	69.062
Σειρά του τμήματος στο μηχανογραφικό δελτίο	-0.009	0.991	-0.166	0.847	0.179*	1.196
Εξωγενή χαρακτηριστικά						
Υπαρξη αρν. γεγονότος κατά τις σπουδές_«ναι»	1.507*	4.513	1.247	3.480	3.448*	31.437
Εργασία κατά τη διάρκεια των σπουδών_«ναι»	1.946*	7.001	3.593	36.343	4.691*	108.962
Στάσεις και Πεποιθήσεις						
Αιώνια φοίτηση_«Υπέρ»	0.227	1.255	-4.111	0.016	16.445	1.584
_«Αδιάφορο»	0.24	1.271	1.183	3.264	8.199	1.779
Επιβολή ανώτατου ορίου φοίτησης_«Υπέρ»	0.893	2.314	-1.297	0.273	3.722*	41.347

**** Για τις ποιοτικές μεταβλητές αναφέρονται και οι τιμές στις οποίες αντιστοιχούν οι συντελεστές της παλινδρόμησης η ερμηνεία των οποίων γίνεται ως προς την κατηγορία/τιμή αναφοράς που δίνεται στην ενότητα 2.**

*** Σημαντικός συντελεστής στο 0.05**

Στην περίπτωση των γυναικών, πλέον των όσων αναφέρθηκαν για το σύνολο του δείγματος, σημαντική θετική επίδραση προκύπτει ότι έχουν από τα ακαδημαϊκά χαρακτηριστικά η συμμετοχή στα «κοινά» -φοιτητικές παρατάξεις- αλλά και η συναναστροφή με συμφοιτητές/τριες, από τα κίνητρα και τις προσδοκίες, οι «διάλυλοι» επιρροής για την επιλογή του αντικειμένου σπουδών (όταν είναι προσωπική επιλογή), απόκτηση γνώσης από τις σπουδές και η σειρά δήλωσης του τμήματος σπουδών στο μηχανογραφικό δελτίο των εισαγωγικών εξετάσεων.

Τα παραπάνω ευρήματα είναι συμβατά με ότι προκύπτει και από άλλες σχετικές μελέτες, όπου δίνονται και οι ανάλογες ερμηνείες και αιτιάσεις. Αυτό ισχύει για τα αποτελέσματα που προκύπτουν στο σύνολο του δείγματος (Murtaugh κ.α: 1999, DesJardins κ.α:1998, Finocchietti: 2004, Astin, A.W.:1965) αλλά και για εκείνα που αφορούν στις υποομάδες των ανδρών (Belloc κ.α:2010, Urquhart, και Pooley: 2007, Bayer: 1968) και των γυναικών (Colagno. κ.α: 2006, Jones: 2010).

4. ΣΥΜΠΕΡΑΣΜΑΤΑ

Σε αυτή την εργασία εξετάζονται οι παράγοντες που συνδέονται με την κατανομή της διάρκειας σπουδών φοιτητών κοινωνικών επιστημών. Η προσέγγιση (που αποτελεί επέκταση του προτύπου του Tinto) προϋποθέτει ότι η η διάρκεια σπουδών επηρεάζεται από πέντε κατηγορίες παραγόντων που αφορούν σε ατομικά χαρακτηριστικά των φοιτητών, σε ακαδημαϊκά χαρακτηριστικά καθώς και σε χαρακτηριστικά κινή-

τρων και προσδοκιών για και από τις σπουδές -συστήματα ενσωμάτωσης στο ακαδημαϊκό περιβάλλον-, σε χαρακτηριστικά προερχόμενα από το περιβάλλον εκτός πανεπιστημίου και σε στάσεις/ πεποιθήσεις των φοιτητών ως προς το θέμα της χωρίς όριο φοίτησης. Η ανάλυση βασίζεται σε ένα μοντέλο διατάξιμης λογιστικής παλινδρόμησης. Τα σημαντικότερα ευρήματα στηρίζουν τις υποθέσεις ότι στη συντόμευση των σπουδών συμβάλλουν κυρίως οι υψηλές επιδόσεις στην αρχή των σπουδών και τα θετικά κίνητρα και οι προσδοκίες για και από τις σπουδές. Αντίθετα στην πολύ εκτεταμένη φοίτηση ή εγκατάλειψη των σπουδών συμβάλει κυρίως η παρουσία εκτός ακαδημαϊκού πλαισίου χαρακτηριστικών -εργασία κατά τη διάρκεια των σπουδών, δυσάρεστα συμβάντα-, καθώς και μερικά ατομικά χαρακτηριστικά.

Στο πλαίσιο του παρόντος κειμένου δεν είναι δυνατός ο περαιτέρω σχολιασμός των αποτελεσμάτων καθώς και η αναφορά στις ειδικότερες ερμηνείες που μπορεί να τους αποδοθούν. Ωστόσο, σημειώνουμε ότι λαμβάνοντας υπόψη το φύλο των ερωτώμενων προκύπτουν ενδιαφέρουσες διαφορές ως προς τους εν λόγω παράγοντες, οι οποίες συνιστούν περαιτέρω αναλύσεις με σκοπό να αποτυπωθούν οι διαδρομές που καθορίζουν τις εξαρτήσεις των χαρακτηριστικών μεταξύ τους και ερμηνεύουν ή αιτιολογούν την σύνδεσή τους με τις διάφορες φάσεις της κατανομής της διάρκειας σπουδών. Στο θέμα αυτό προτιθέμεθα να ανταποκριθούμε σε μελλοντική εργασία.

ABSTRACT

Issues related to the duration of university studies, have attracted the interest of many researchers from different scientific fields, almost from the middle of the 20th century. In this context, a survival analysis methodology has been proposed for the study of the distribution of the duration of study for students of social sciences. In this paper, results are given with respect to the factors associated with this distribution. The approach followed here (which is an extension of Tinto's one) assumes that the duration of studies is influenced by factors related to students' individual characteristics, academic achievements/features, motivations and expectations concerning studies, characteristics derived from the environment outside the university and to students' attitudes/beliefs on the issue of unlimited duration of studies. The analysis involves an ordinal logistic regression model. The major findings suggest that factors contributing to shortening of the duration of studies concern high academic achievements at early stages, positive motivation, expectations, attitudes and beliefs about studies where as, those concerning external situations, negative academic experiences and some individual characteristics contribute to student attrition.

ΑΝΑΦΟΡΕΣ

- Aina, C. (2005), "Parental Background and College drop out. Evidence from Italy", EPUNet Conference, Working paper, 1-22.
- Ampaw, F., Jaeger, A. (2012), "Completing the Three Stages of Doctoral Education: An Event History Analysis", *Research in Higher Education*, 53, 640–660.
- Astin A. W. (1965), "Effect of different college environments on the vocational choices of high aptitude students", *Journal of Counseling Psychology*, 12, 28-34.
- Astin, A. (1975), *"Preventing students from dropping out"*, San Francisco Jossey-Bass.
- Bayer, A. (1968), "The College Drop-out: Factors Affecting Senior College Completion", *Sociology of Education*, 41, 3, 305-316.
- Bean, J. (1980), "Dropouts And Turnover: The Synthesis and Test of a Causal Model of Student Attrition", *Research in Higher Education*, 12, 2, 155-177.
- Belloc F., Maruotti A., Petrella L. (2011), "How individual characteristics affect university students drop-out: a semi parametric mixed-effects model for an Italian case study", *Journal of Applied Statistics*, 38, 10, 9-12.
- Berger, J. B., & Lyon, S. C. (2005), "Past to present: A historical look at retention" in A. Seidman (Ed.), *"College student retention: Formula for student success"*, Westport, CT: Praeger Publishers, 1-30.
- Cabrera, A. Castaneda, M. Nora, A. Hengstler, D. (1990), "The Convergence Between Two Theories of College Persistence", *The Journal of Higher Education*, 63, 2, 143-164.
- Cabrera, A. Nora, A. Castaneda, M. (1993), "College Persistence: Structural Equations Modeling Test of an Integrated Model of Student Retention", *The Journal of Higher Education*, 64, 2, 123-139.
- Cingano F., Cipollone P. (2007), "University drop-out – The case of Italy", *Banca D' Italia – Eurosistema*, 626, 1-25.
- Colagno J.C., Crosta P., Bailey T., Jenkins D. (2006), "Stepping Stones To A Degree: The Impact Of Enrollement Pathways And Milestones On Community College Student Outcomes", *Community College Research Center Working Paper*, 4, 1-31.
- De Francesco, P. and Corrado, T. (1977), "Drop-Outs from Italian Universities: 1960-1975", *Paedagogica Europaea*, 12, 3, 81-105.
- DesJardins S.L., Ahlburg D.A., McCall B.P. (1999), "An event history model of student departure", *Economics of Education Review*, 18, 376-377.

- Eckland, B. K. (1964), "College dropouts who came back", *Harvard Educational Review*, 34, 402-420.
- Finocchietti G. (2004), "Students and Universities in Italy in an Age of Reform", *European Journal of Education*, 39, 4, 464-465.
- Jones W. (2010), "The Impact of Social Integration on Subsequent Institutional Commitment Conditional on Gender", *Research in Higher Education*, 51, 687-700.
- Καλαμάρας Δ., Καλαματιανού Α. (2006), "Πίνακες Επιβίωσης για την Εκτίμηση της Κατανομής Διάρκειας Σπουδών: Περαιτέρω Αποτελέσματα", *Ελληνικό Στατιστικό Ινστιτούτο Πρακτικά 19ου Πανελληνίου Συνεδρίου Στατιστικής*, 221-229.
- Kalamatianou. A. G. (1999), Using Life Table Method on a Censored Data Set for Estimating Duration of Studies in University Departments, In *Bulletin of the International Statistical Institute*, 52nd Session, Contributed papers. Tome LVIII, 2, 101-102.
- Kalamatianou, A. G., McClean, S. (2003), "The Perpetual Student: Modeling Duration of Undergraduate Studies Based on Lifetime-Type Educational Data", *Lifetime Data Analysis*, 9, 311-330.
- Morgan, B. J. T., Smith D. M. (1992), "A note on Wadley's problem with over dispersion", *Applied Statistics*, 41, 349-354.
- Murtaugh P., Burns L., Schuster J. (1999), "Predicting the Retention of University Students", *Research in Higher Education*, 40, 3, 355-371.
- Samuel, S., (1978), "Variables Involved in Withdrawal during the First Two Years of College: Preliminary Findings from the National Longitudinal Study of the High School Class of 1972", *American Educational Research Journal*, 15, 3, 361-372.
- Spady, W. G. (1970), "Dropouts from Higher Education: An Interdisciplinary Review and Synthesis", *Interchange*, 1, 1, 64-85.
- Tinto, V. (1975), "Dropout from Higher Education: A Theoretical Synthesis of Recent Research", *Review of Educational Research*, 45, 1, 89-125.
- Urquhart B., Pooley J.A., (2007), "The Transition Experience of Australian Students to University: The Importance of Social Support", *The Australian Community Psychologist*, 19, 2, 79-80.
- Webster, A., Showers, V. (2011), "Measuring Predictors of Student Retention Rates", *American Journal of Economics and Business Administration*, 3, 2, 296-306.



ΜΗ ΓΡΑΜΜΙΚΗ ΑΝΑΛΥΣΗ ΧΡΟΝΟΣΕΙΡΩΝ ΤΙΜΩΝ ΑΓΡΟΤΙΚΩΝ ΠΡΟΪΟΝΤΩΝ: ΕΦΑΡΜΟΓΗ ΣΤΗΝ ΚΕΝΤΡΙΚΗ ΑΓΟΡΑ ΘΕΣΣΑΛΟΝΙΚΗΣ Α.Ε.

Καρακασίδου Σοφία

Κεντρική Αγορά Θεσσαλονίκης Α.Ε.
sofkarak@otenet.gr

ΠΕΡΙΛΗΨΗ

Στην παρούσα εργασία διερευνάται η χρήση μη γραμμικών μεθόδων ανάλυσης χρονοσειρών σε χρονοσειρές τιμών πώλησης επιλεγμένων προϊόντων λαχανικών στην Κεντρική Αγορά Θεσσαλονίκης Α.Ε. Πέρα από την περιγραφική στατιστική ανάλυση χρησιμοποιείται η μέθοδος της μέσης αμοιβαίας πληροφορίας για την εξαγωγή χαρακτηριστικών χρόνων καθώς και η μέθοδος των διαγραμμάτων επαναφοράς (Recurrence Plot Analysis). Η τελευταία έχει χρησιμοποιηθεί σε πολλές περιπτώσεις για τη μελέτη οικονομικών αλλά και πολλών φυσικών συστημάτων. Από τις παραπάνω αναλύσεις σε συνδυασμό και με τη χρήση ομαδοποίησης εξάγονται χαρακτηριστικά της δυναμικής των αντίστοιχων χρονοσειρών των τιμών των προϊόντων.

Λέξεις Κλειδιά: χρονοσειρές, μέση αμοιβαία πληροφορία, διαγράμματα επαναφοράς (recurrence plots), ποσοτική ανάλυση επαναφοράς (recurrence quantification analysis), ομαδοποίηση.

1. ΕΙΣΑΓΩΓΗ

Η ανάλυση χρονοσειρών διαδραματίζει σημαντικό ρόλο στην κατανόηση της δυναμικής των ποικίλων συστημάτων (φυσικών, οικονομικών, βιολογικών κ.λ.π.) και στην πρόβλεψη της συμπεριφοράς τους. Η πλειονότητα των μεθόδων ανάλυσης προϋποθέτει τη γραμμικότητα της συμπεριφοράς του δυναμικού συστήματος καθώς και την παρουσία στοχαστικού θορύβου. Γνωστές και ευρύτατα διαδεδομένες μέθοδοι όπως η συνάρτηση αυτοσυσχέτισης και το φάσμα ισχύος έχουν αποδειχθεί ιδιαίτερα επιτυχείς στην ανάλυση χρονοσειρών. Όμως οι μέθοδοι αυτοί συναντούν περιορισμούς σε ότι αφορά την ανάλυση μη γραμμικών ντετερμινιστικών συστημάτων καθώς αδυνατούν να ανιχνεύσουν μη γραμμικές συσχετίσεις.

Για την αναγνώριση χαοτικής ντετερμινιστικής συμπεριφοράς και διάκρισής της από την στοχαστική συμπεριφορά έχουν προταθεί μη γραμμικές μέθοδοι οι οποίες βασίζονται στην έννοια του ανακατασκευασμένου χώρου των φάσεων. Οι μη γραμμικές μέθοδοι περιλαμβάνουν τον υπολογισμό της διάστασης συσχέτισης, τη

μέση αμοιβαία πληροφορία (Kantz & Schreiber 2003), τα γραφήματα επαναφοράς (Recurrence Plots - RPs) (Eckmann et al. 1987) και την Ποσοτική Ανάλυση Επαναφοράς (Recurrence Quantification Analysis - RQA) (Zbilut and Webber, 1992, Trulla et al. 1996). Στην παρούσα μελέτη πέρα από την περιγραφική στατιστική ανάλυση επικεντρωνόμαστε στη μέση αμοιβαία πληροφορία καθώς και στα RPs και την RQA. Οι τελευταίες μέθοδοι έχουν χρησιμοποιηθεί στην ανάλυση χρονοσειρών και την ταυτοποίηση καταστάσεων σε μια πληθώρα συστημάτων όπως πρωτεΐνες (Zbilut et al., 2004; Giuliani et al., 2002), οικονομικά συστήματα (Strozzi et al., 2002; Fabretti & Ausloos, 2005), μοριακά συστήματα (Karakasidis et al., 2007).

Επίσης η ομαδοποίηση (clustering) είναι μια χρήσιμη μεθοδολογία για την κατηγοριοποίηση στοιχείων. Στην περίπτωση μας χρησιμοποιήσαμε την ιεραρχική ομαδοποίηση (hierarchical clustering) η οποία επιτρέπει το διαχωρισμό των στοιχείων σε κατηγορίες χωρίς να θέσουμε a priori περιορισμούς στον αριθμό των ομάδων στις οποίες διαχωρίζονται τα δεδομένα μας.

Στην παρούσα εργασία χρησιμοποιούμε τις παραπάνω μεθόδους στην ανάλυση χρονοσειρών τιμών λαχανικών οι οποίες προέρχονται από την Κεντρική Αγορά Θεσσαλονίκης Α.Ε. Σε πρώτη φάση επιλέξαμε να εργαστούμε με μερικές ομάδες προϊόντων προκειμένου να διερευνήσουμε τη μελλοντική χρήση για αυτόματα διαχωρισμό σε μεγάλο αριθμό προϊόντων.

Στην ενότητα 2 παρουσιάζονται συνοπτικά τα δεδομένα που χρησιμοποιήθηκαν στην παρούσα εργασία, στην ενότητα 3 η μεθοδολογία ενώ στην ενότητα 4 παρουσιάζονται τα αποτελέσματα και τέλος στην ενότητα 5 τα συμπεράσματα.

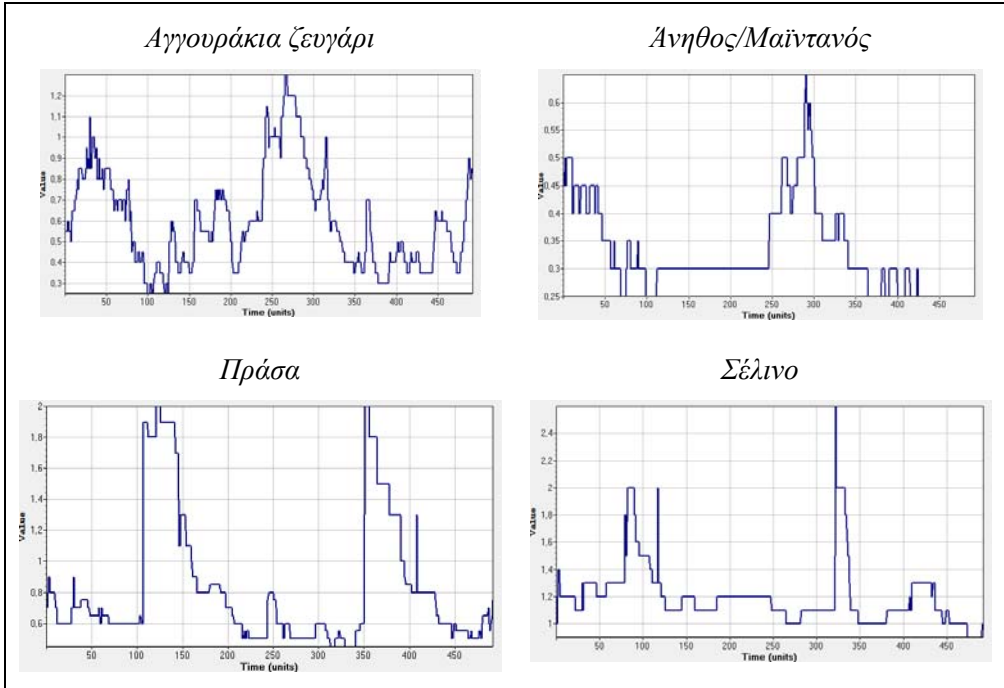
2. ΔΕΔΟΜΕΝΑ

Τα δεδομένα προέρχονται από το δελτίο τιμών της Κεντρικής Αγοράς Θεσσαλονίκης Α.Ε. και αφορούν ημερήσιες τιμές σε ορισμένα αγροτικά προϊόντα για δύο έτη 2008-2009. Τα προϊόντα ευρείας κατανάλωσης που επιλέγησαν προκειμένου να πραγματοποιηθεί μία πρώτη ανάλυση είναι:

1) Αγγουράκια το ζευγάρι 2) Άνηθος/Μαϊντανός δεμάτι μικρό 3) Πράσα 4) Ρόκα δεμάτι μικρό 5) Σαλάτες 6) Σέλινο 7) Σκόρδα 8) Σπανάκι 9) Ντομάτες.

Στην εικόνα 1 παρουσιάζονται οι σχετικές χρονοσειρές για μερικά από τα παραπάνω είδη.

Εικόνα 1. Αντιπροσωπευτικές χρονοσειρές ορισμένων προϊόντων.



3. ΜΕΘΟΔΟΛΟΓΙΑ ΑΝΑΛΥΣΗΣ

3.1 Συνάρτηση Αμοιβαίας Πληροφορίας

Η συνάρτηση αμοιβαίας πληροφορίας προτάθηκε από τους Fraser & Swinney (1986) ως ένα εργαλείο προσδιορισμού χρονικής υστέρησης για τη βέλτιστη εμφύθιση (embedding) κατά την ανακατασκευή του χώρου των φάσεων και δίνεται από τη σχέση

$$I = - \sum_{ij} p_{ij}(\tau) \ln \frac{p_{ij}(\tau)}{p_i p_j},$$

όπου p_i είναι η πιθανότητα να ευρεθεί μία τιμή της χρονοσειράς στο διάστημα i και $p_{ij}(\tau)$ είναι η δεσμευμένη πιθανότητα μία παρατήρηση που βρίσκεται στο διάστημα i να βρεθεί στο διάστημα j έπειτα από χρόνο τ για $i=1,2,\dots,n$ και $j=1,2,\dots,n$. Βασικό πλεονέκτημα της συνάρτησης αμοιβαίας πληροφορίας σε σχέση με τη συνάρτηση αυτοσυσχέτισης είναι ότι λαμβάνει υπόψη μη γραμμικές συσχετίσεις. Το πρώτο ελάχιστο που προκύπτει από τη συνάρτηση αμοιβαίας πληροφορίας αποτελεί και την καλύτερη επιλογή για χρόνο υστέρησης.

3.2 Γραφήματα Επαναφοράς

Τα γραφήματα επαναφοράς (RPs) είναι γραφικά εργαλεία που προτάθηκαν από τους Eckmann et al. (1987) για την εξαγωγή ποιοτικών χαρακτηριστικών από μια χρονοσειρά και βασίζονται στην ανακατασκευή του χώρου των φάσεων του συστήματος.

Κατά τη διαδικασία αυτή αρχικά κάνουμε εμφύθιση της χρονοσειράς $x(t)$, σε ένα m -διάστατο χώρο δημιουργώντας μια σειρά από διανύσματα $s_i = [x(t_i - (m-1)\tau_d), x(t_i - (m-2)\tau_d), \dots, x(t_i)]$, όπου το τ_d δηλώνει την επιλεγμένη χρονική υστέρηση. Στη συνέχεια υπολογίζεται η απόσταση $d_{ij} = \|s_i - s_j\|$ μεταξύ των σημείων i, j στον m -διάστατο αναδομημένο χώρο των φάσεων. Ως αποτέλεσμα εξάγεται ένας πίνακας $N \times N$ αποστάσεων d_{ij} , όπου N ο αριθμός των ανακατασκευασμένων διανυσμάτων κατάστασης s_i . Με βάση τον πίνακα των αποστάσεων κατασκευάζεται ένα διδιάστατο γράφημα αναπαριστώντας με μία μαύρη κουκκίδα στο (i, j) εάν η απόσταση μεταξύ των σημείων i και j είναι μικρότερη από μία δεδομένη τιμή r_i που ορίζεται σαν ακτίνα αποκοπής. Θεωρώντας ότι $r_i = r_j$ το εξαγόμενο γράφημα είναι συμμετρικό ως προς την κύρια διαγώνιο (Zbilut and Webber, 1992, Trulla et al. 1996). Το σκούρο σημείο καταδεικνύει το βαθμό επαναληπτικότητας της δυναμικής διαδικασίας και ένα RP δείχνει τη δομή των αυτοσυσχετίσεων του συστήματος. Υπάρχουν δυο γενικές κατηγορίες διαγραμμάτων επαναφοράς: Χωρίς κατώφλι (πλήρως ανεπτυγμένες) και με κατώφλι. Σε αυτά με κατώφλι, η κουκκίδα στο σημείο (i, j) είναι μαύρη εάν η απόσταση είναι μικρότερη από τη δεδομένη ακτίνα και λευκή εάν είναι μεγαλύτερη. Σε ένα πλήρως ανεπτυγμένο RP ως ακτίνα επιλέγεται η μέγιστη απόσταση μεταξύ των αναδομημένων σημείων με αποτέλεσμα όλα τα σημεία να θεωρούνται επαναληπτικά, οι κουκκίδες όμως είναι χρωματισμένες ανάλογα με το πόσο μεγάλη ή μικρή είναι η απόσταση d_{ij} . Στην παρούσα μελέτη χρησιμοποιούμε ανεπτυγμένα RPs. Η εξαγωγή ποσοτικών χαρακτηριστικών από τα RPs μας επιτρέπει να ποσοτικοποιήσουμε χρήσιμες πληροφορίες. Στην παρούσα εργασία χρησιμοποιήσαμε την ποσότητα trapping time η οποία υπολογίζεται από τη σχέση

$$TT = \frac{\sum_{v=v_{\min}}^N v P^{\varepsilon}(v)}{\sum_{v=v_{\min}}^N P^{\varepsilon}(v)},$$

όπου $P^{\varepsilon}(v)$ είναι η πιθανότητα παρουσίας κατακόρυφων γραμμών μήκους v (ε είναι η ακτίνα αποκοπής), $v_{\min}$ το ελάχιστο μήκος σημείων που θεωρούμε ως γραμμή. Μια συνήθης πρακτική είναι να λαμβάνεται ίσο με το πρώτο ελάχιστο τη μέσης αμοιβαίας πληροφορίας (Marwan et al. 2007).

Για την ανάλυση της μέσης αμοιβαίας πληροφορίας και των διαγραμμάτων επαναφοράς χρησιμοποιήθηκε το πρόγραμμα Visual Recurrence Analysis 9.0 (Kononov 2010) και για την στατιστική ανάλυση και ομαδοποίηση το Matlab έκδοση R2008b.

4. ΑΠΟΤΕΛΕΣΜΑΤΑ

4.1 Περιγραφική στατιστική

Σε πρώτη φάση υπολογίσαμε τη μέση τιμή, τη διασπορά, την τυπική απόκλιση, το συντελεστή ασυμμετρίας και το συντελεστή κύρτωσης. Τα αποτελέσματα παρουσιάζονται στον Πίνακα 1.

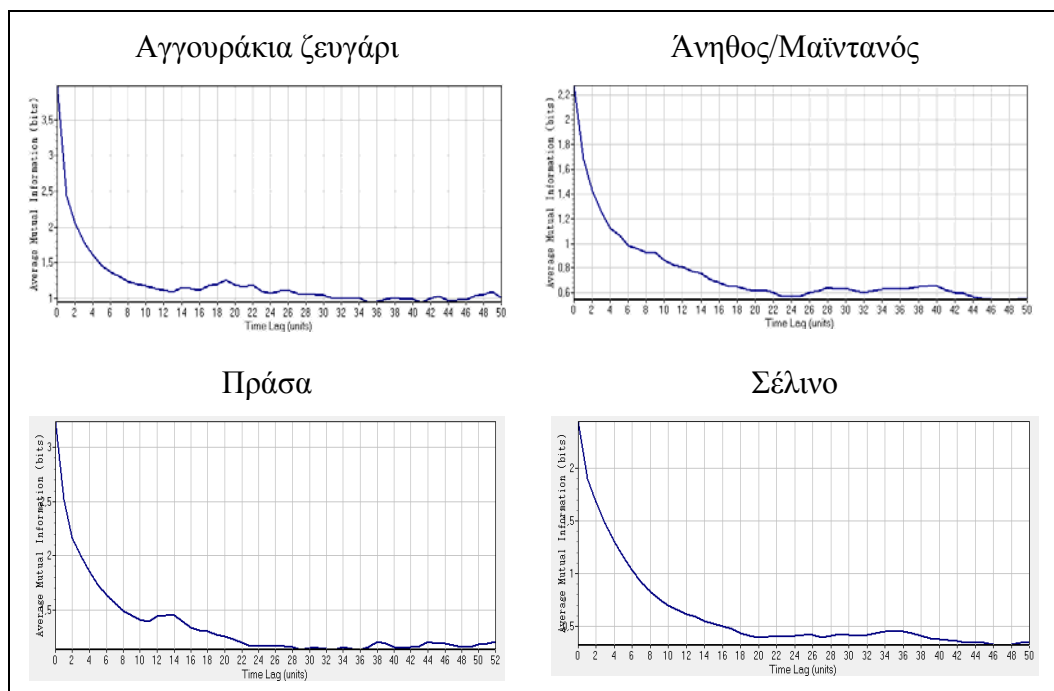
Πίνακας 1. Περιγραφικά μέτρα.

Είδος	Μέση τιμή	Διασπορά	Συντελεστής ασυμμετρίας	Συντελεστής κύρτωσης
Αγγουράκια	0,6025	0,0568	0,7801	-0,0936
Άνηθος/ Μαϊντανός	0,3277	0,0063	1,3013	1,2257
Πράσα	0,8268	0,1825	1,6160	1,3526
Ρόκα δεμάτι	0,3733	0,0068	2,2705	6,2772
Σαλάτες	0,9757	0,1812	0,8593	-0,4317
Σέλινο	1,1974	0,0523	2,3195	6,8383
Σκόρδα	2,2349	0,1697	1,3654	1,2076
Σπανάκι	0,8521	0,1551	1,0244	-0,0845
Ντομάτες	0,8804	0,1044	0,8470	0,4232

4.2 Αμοιβαία Πληροφορία

Στην Εικόνα 2 παρουσιάζονται ενδεικτικά διαγράμματα της μέσης αμοιβαίας πληροφορίας για επιλεγμένα είδη. Τα εντοπισμένα ελάχιστα για όλα τα είδη παρουσιάζονται συνοπτικά στον Πίνακα 2. Όπως μπορούμε να παρατηρήσουμε σε ορισμένα είδη το πρώτο ελάχιστο της αμοιβαίας πληροφορίας παρουσιάζεται για χρόνο $\tau=10$, ενώ σε μερικά άλλα όταν $\tau=20$ και σε κάποιο μάλιστα όταν $\tau=29$. Υπενθυμίζουμε ότι ο χρόνος (τ) μετράται σε εργάσιμες ημέρες.

Εικόνα 2. Ενδεικτικά διαγράμματα μέσης αμοιβαίας πληροφορίας.



Πίνακας 2. 1^ο ελάχιστο μέσης αμοιβαίας πληροφορίας.

Είδος	τ_d (εργάσιμες ημέρες)
Σαλάτες	10
Σπανάκι	10
Πράσα	11
Ντομάτες	11
Αγγουράκια	13
Άνηθος/Μαϊντανός	20
Σέλινο	20
Ρόκα δεμάτι	21
Σκόρδα	29

4.3 Διαγράμματα Επαναφοράς

Σε όλες τις περιπτώσεις χρησιμοποιήθηκε ως διάσταση εμβύθισης $m=5$ (δεδομένου ότι τα αποτελέσματα συγκλίνουν) και ως χρόνοι τα πρώτα ελάχιστα που πρόεκυψαν από την ανάλυση της μέσης αμοιβαίας πληροφορίας (Πίνακας 2). Στην Εικόνα 3 παρουσιάζονται ενδεικτικά διαγράμματα επαναφοράς. Σε γενικές γραμμές τα μπλε σημεία αντιστοιχούν σε ισχυρές συσχετίσεις, τα κόκκινα σε μικρότερες, ενώ τα κίτρινα σε πιο ασθενείς συσχετίσεις.

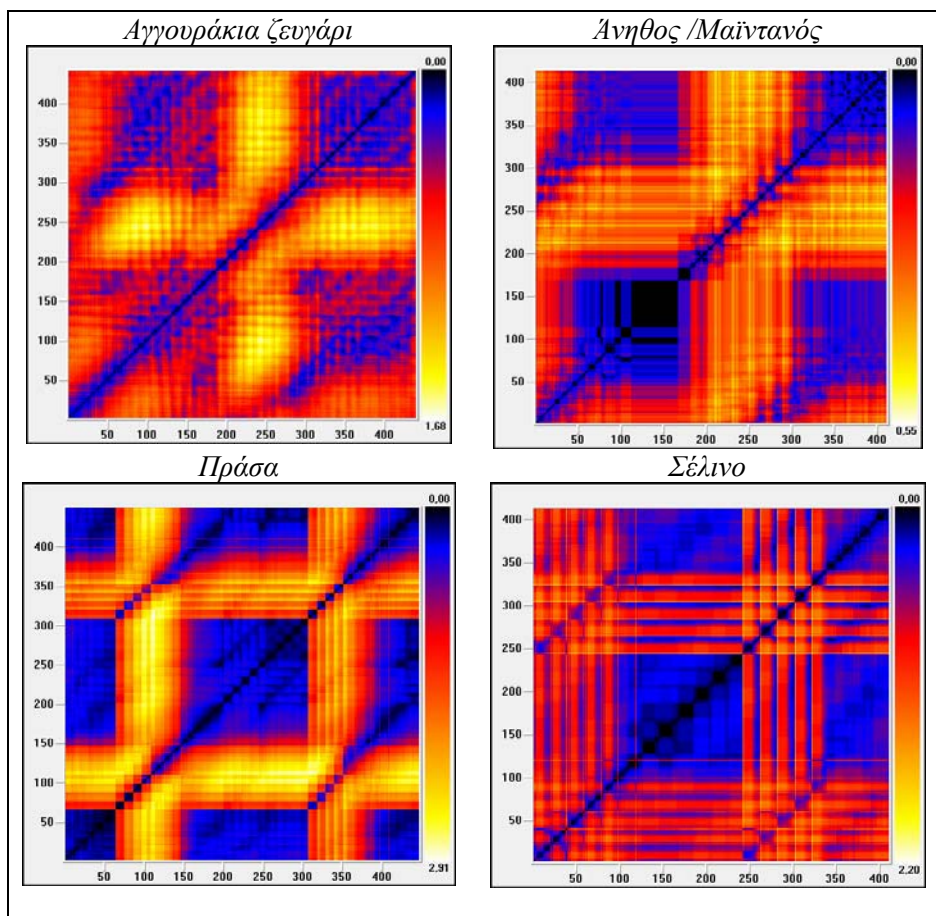
Χαρακτηριστικά στην περίπτωση των αγγουριών παρατηρούμε μικρές διαγώνιες γραμμές που σημαίνει ότι συχνά υπάρχουν διαταραχές (όπως μπορεί να δει κανείς στο αντίστοιχο γράφημα επαναφοράς της χρονοσειράς στην Εικόνα 3). Επίσης οι μικρές κατακόρυφες (ή οριζόντιες) μπλε γραμμές είναι ενδεικτικές παγίδευσης του συστήματος στην ίδια τιμή για μικρά χρονικά διαστήματα. Επίσης παρατηρούνται μεγάλες παράλληλες προς την κύρια διαγώνιο γραμμές που αντιστοιχούν σε μια απόσταση μεταξύ τους περίπου 250 εργάσιμων ημερών (και είναι ενδεικτικό της ετήσιας περιοδικότητας). Η μορφή φτερών αεροπλάνου γύρω από την κεντρική διαγώνιο είναι ενδεικτική της μικρής τάσης. Στην περίπτωση του άνηθου-μαϊντανού παρατηρούμε μεγαλύτερες διαγώνιες γραμμές που σημαίνει ότι υπάρχουν πιο αραιές διαταραχές, μεγάλες κατακόρυφες (οριζόντιες) μπλε γραμμές που δείχνουν παγίδευση του συστήματος στην ίδια τιμή για σχετικά μεγάλα διαστήματα. Διαφορετικές συμπεριφορές παρουσιάζουν τα πράσα και το σέλινο όπως προκύπτει από τα διάγραμμα επαναφοράς τους που εμφανίζονται στην Εικόνα 3.

Από τα διαγράμματα χρησιμοποιήσαμε την ποσότητα Trapping Time που αναφέρεται στον χρόνο παγίδευσης σε μια κατάσταση και την τάση (Trend). Τα αποτελέσματα συνοψίζονται στον πίνακα 3.

Πίνακας 3. *Trapping time και Trend.*

	Trapping Time	Trend
Σαλάτες	12,857	-9,302
Σπανάκι	13,854	2,481
Πράσα	12,765	-12,230
Ντομάτες	16,261	-10,062
Αγγουράκια	13,000	-8,368
Άνηθος/Μαϊντανός	40,163	- 27,423
Σέλινο	23,571	-37,590
Ρόκα δεμάτι	29,647	-14,997
Σκόρδα	32,818	-65,064

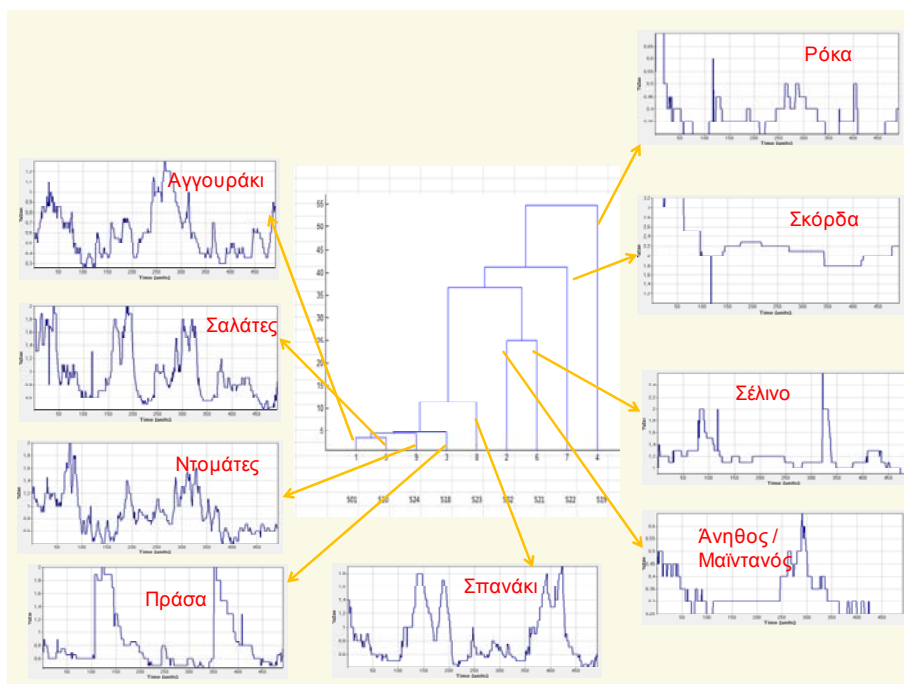
Εικόνα 3.Ενδεικτικά διαγράμματα επαναφοράς.



4.4 Ομαδοποίηση

Για την ομαδοποίηση χρησιμοποιήθηκαν από τα στατιστικά μεγέθη η μέση τιμή ο συντελεστής συμμετρίας και κύρτωσης αφού διαιρέθηκαν με την τυπική απόκλιση (προκειμένου να κανονικοποιηθούν), το trapping time και το trend. Τα αποτελέσματα παρουσιάζονται σχηματικά στην Εικόνα 4. Μπορούμε να δούμε ότι σχηματίζονται διάφορες ομάδες. Η πρώτη ομάδα που ξεχωρίζει είναι η χρονοσειρά με τη ρόκα. Ακολουθεί η χρονοσειρά με τα σκόρδα, ενώ μια ομάδα αποτελούν το σέλινο ο άνηθος και ο μαϊντανός. Συνεχίζοντας την ομαδοποίηση οι επόμενες ομάδες που ξεχωρίζουν είναι το σπανάκι, τα πράσα, και οι ντομάτες. Στην τελευταία ομάδα που προέκυψε συγκαταλέγονται οι σαλάτες και τα αγγουράκια που οι χρονοσειρές τους έχουν παρόμοιες συμπεριφορές όπως προκύπτει από τη διαγραμματική απεικόνιση των χρονοσειρών τους.

Εικόνα 4. Αποτελέσματα ομαδοποίησης.



5. ΣΥΜΠΕΡΑΣΜΑΤΑ

Η χρήση μη γραμμικών εργαλείων όπως RPs είναι ιδιαίτερα χρήσιμη για την εξαγωγή ποσοτικών και ποιοτικών συμπερασμάτων σχετικών με τη μεταβολή της τιμής των προϊόντων. Η ομαδοποίηση καταδεικνύει χαρακτηριστικά για τη συμπεριφορά των προϊόντων και τη δυναμική τους. Η μεθοδολογία μπορεί να επεκταθεί και στη χρήση άλλων προϊόντων όχι μόνο αγροτικών. Προκειμένου να επιβεβαιωθούν περαιτέρω τα αποτελέσματα της παρούσας έρευνας προτείνεται η μελέτη χρονοσειρών με μεγαλύτερη χρονική διάρκεια καθώς και η εφαρμογή σε μεγαλύτερο αριθμό προϊόντων.

ABSTRACT

In the present paper we investigate the use of nonlinear analysis methods of prices of selected agricultural product from the Central Market of Thessaloniki, Greece. Besides the descriptive statistical analysis we employ in our analysis the method of the average mutual information as well as Recurrence Plots and Recurrence Quantification analysis. The results in combination with clustering analysis reveals interesting characteristics of the dynamics of the products' process. The procedure could be envisaged to be applied in the classification of large number of products.

ΑΝΑΦΟΡΕΣ

- Eckmann J.-P., Kamphorst S. O. and Ruelle D. (1987). Recurrence plots of dynamical systems. *Europhysics Letters*, **4**, 973-977.
- Fabretti, A. and Ausloos, M. (2005). Recurrence plot and recurrence quantification analysis techniques for detecting a critical regime. Examples from financial market indices. *Int. J. Mod. Phys. C*, **16**, 671-706.
- Fraser A.M. and Swinney H.L. (1986). Independent coordinates for strange attractors from mutual information. *Physical Review A*, **33**, 1134-1140.
- Giuliani, A., Benigni, R., Zbilut, J. P., Webber Jr., C. L., Sirabella, P. and Colosimo, A. (2002). Nonlinear signal analysis methods in the elucidation of protein sequence-structure relationships. *Chem. Rev.*, **102**, 1471-1491.
- Kantz H. and Schreiber T. (2003). *Nonlinear Time Series Analysis*, 2nd edition, Cambridge.
- Karakasidis T. E., Fragkou A, Liakopoulos A. (2007). System dynamics revealed by recurrence quantification analysis: Application to molecular dynamics simulations. *Physical Review E*, **76**, 021120.
- Kononov E. (2010). <http://www.myjavaserver.com/~nonlinear/vra/download.html>.
- Marwan N., M. Romano C., Thiel, M. and Kurths J. (2007). Recurrence plots for the analysis of complex systems. *Physics Reports*, **438**, 237-329.
- Strozzi, F., Zaldiovar, J.-M. and Zbilut, J. P. (2002). Application of nonlinear time series analysis techniques to high-frequency currency exchange data. *Physica A*, **312**, 520-538.
- Trulla L.L., Giuliani A., Zbilut J. and Webber, C.L. Jr. (1996). Recurrence quantification analysis of the logistic equation with transients. *Physics Letters A*, **223**, 255-260.
- Zbilut, J.P., Webber, C.L., Jr. (1992). Embeddings and delays as derived from quantification of recurrence plots, *Physics Letters A*, **171**, 199-203.
- Zbilut, J. P., Giuliani, A., Colosimo, A., Mitchell, J. C., Colafranceschi, M., Marwan, N., Webber Jr., C.L. and Uversky, V.N. (2004). Charge and hydrophobicity patterning along the sequence predicts the folding mechanism and aggregation of proteins: *A computational approach. J. Proteome Res.*, **3**, 1243-1253.

ΕΝΑ ΓΕΝΙΚΟ ΜΟΝΤΕΛΟ ΓΙΑ ΤΕΣΤ ΕΚΚΙΝΗΣΗΣ

Μ.Β. Κούτρας¹, Ν. Balakrishnan², Φ. Σ. Μηλιένος³

¹Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης, Πανεπιστήμιο Πειραιά
mkoutras@unipi.gr

²Department of Mathematics and Statistics, McMaster University
bala@mcmaster.ca

³Τμήμα Μαθηματικών και Στατιστικής, Πανεπιστήμιο Κύπρου
milienos@yahoo.com

ΠΕΡΙΛΗΨΗ

Τα τεστ εκκίνησης αποτελούν ένα τρόπο για την αξιολόγηση της αξιοπιστίας ενός εξοπλισμού, μέσω διαδοχικών προσπαθειών εκκίνησης του εξοπλισμού. Στη σχετική βιβλιογραφία έχουν παρουσιαστεί και μελετηθεί διάφορα κριτήρια για το πότε ένας εξοπλισμός πρέπει να θεωρείται αξιόπιστος ή όχι, πολλά από τα οποία συνδέονται άμεσα με τη θεωρία σχηματισμών σε ακολουθίες διακριτών τυχαίων μεταβλητών. Σημαντικό ρόλο στη μελέτη των τεστ εκκίνησης παίζει η τυχαία μεταβλητή που εκφράζει τον αριθμό των προσπαθειών, μέχρι τον τερματισμό του τεστ (δηλαδή μέχρι τη λήψη απόφασης) η οποία ονομάζεται μήκος του τεστ. Στην παρούσα εργασία μελετάμε τεστ εκκίνησης στα οποία κάθε προσπάθεια εκκίνησης καταλήγει σε δυο δυνατά αποτελέσματα (επιτυχία ή αποτυχία) και εισάγουμε ένα γενικότερο μοντέλο και τρόπο μελέτης τους. Υποθέτοντας ότι τα αποτελέσματα των προσπαθειών, σχηματίζουν μια ακολουθία ανταλλάξιμων δίτιμων τυχαίων μεταβλητών (exchangeable random variables), δίνονται ακριβή και προσεγγιστικά αποτελέσματα για την κατανομή του μήκους του τεστ.

Λέξεις Κλειδιά: Ανταλλάξιμες τυχαίες μεταβλητές; τεστ εκκίνησης; συστήματα αξιοπιστίας.

1. ΕΙΣΑΓΩΓΗ

Τις τελευταίες δεκαετίες, λόγω και της αυξανόμενης ανάγκης δημιουργίας συστημάτων ή γενικότερα προϊόντων με υψηλή αξιοπιστία, έχει παρατηρηθεί μεγάλο ερευνητικό ενδιαφέρον για τη μελέτη των τεστ εκκίνησης (start-up demonstration tests). Τα τεστ εκκίνησης αποτελούν ένα μέσο για την αξιολόγηση της ποιότητας ενός εξοπλισμού (συστήματα πυρασφάλειας, ηλεκτρονικές συσκευές, ηλεκτρογεννήτριες, αντλίες νερού, μπαταρίες κ.α.), μέσω διαδοχικών προσπαθειών εκκίνησης του εξοπλισμού.

Οι Hahn and Gage (1983) μελέτησαν ένα τεστ εκκίνησης με το οποίο ο εξοπλισμός γίνεται αποδεκτός εάν παρατηρηθεί ένα πλήθος από συνεχόμενες επιτυχείς εκκινήσεις/δοκιμές (μοντέλο CS). Ο αριθμός των αποτυχημένων εκκινήσεων δεν έπαιζε κανένα ρόλο στη διαδικασία ελέγχου που πρότειναν και το μοντέλο μελετήθηκε κάτω από την υπόθεση ότι οι δοκιμές είναι ανεξάρτητες και ισόνομες δίτιμες τυχαίες

μεταβλητές. Οι Viveros and Balakrishnan (1993), Balakrishnan *et al.* (1995) και Balakrishnan *et al.* (1997) μελέτησαν και αυτοί το μοντέλο CS υποθέτοντας ανεξάρτητες και ισόνομες ή μαρκοβιανά εξαρτημένες δοκιμές. Οι χρόνοι αναμονής που σχετίζονται με το μοντέλο αυτό έχουν μελετηθεί και στο πλαίσιο της γεωμετρικής κατανομής τάξεως k από πολλούς συγγραφείς (βλ. π.χ. Aki *et al.* 1984, ή το βιβλίο των Balakrishnan and Koutras 2002, Κεφ. 2).

Οι Balakrishnan and Chan (2000) πρότειναν ένα μοντέλο στο οποίο ο εξοπλισμός γίνεται αποδεκτός εάν παρατηρηθεί ένα πλήθος από συνεχόμενες επιτυχείς δοκιμές πριν από ένα σύνολο αποτυχημένων δοκιμών (όχι απαραίτητα συνεχόμενων), διαφορετικά, ο εξοπλισμός απορρίπτεται (μοντέλο CSTF). Οι Smith and Griffith (2005), Martin (2004, 2008), Chan *et al.* (2008), Scollnik (2010, 2011) και Eryilmaz and Chakraborti (2008), μελέτησαν επίσης το μοντέλο CSTF κάτω από διάφορες υποθέσεις για την εξάρτηση των δοκιμών.

Στις εργασίες των Smith and Griffith (2008), Martin (2008), Eryilmaz (2010) και Yalcin and Eryilmaz (2012) εξετάζονται τα μοντέλα TSTF, CSCF και TSF. Βάσει του μοντέλου π.χ. TSTF, ο εξοπλισμός γίνεται αποδεκτός εάν παρατηρηθεί ένα πλήθος από επιτυχείς δοκιμές (όχι απαραίτητα συνεχόμενες) πριν από ένα σύνολο αποτυχημένων δοκιμών (όχι απαραίτητα συνεχόμενων). Επίσης, διάφορα τεστ εκκίνησης με ποικίλα κριτήρια αποδοχής και απόρριψης, αποτελούν αντικείμενο μελέτης των εργασιών Koutras and Balakrishnan (1999), Antzoulakos *et al.* (2009), Gera (2010, 2011), Smith and Griffith (2011) κ.α.

Στην παρούσα εργασία, αφού αναφερθούμε αρχικά σε κάποιους χρήσιμους ορισμούς και συμβολισμούς, εισάγουμε ένα γενικό μοντέλο και έναν ενιαίο τρόπο μελέτης των τεστ εκκίνησης με δυο δυνατά αποτελέσματα. Ως ειδικές περιπτώσεις του μοντέλου αυτού, προκύπτουν τα περισσότερα από τα μέχρι τώρα προταθέντα μοντέλα. Τέλος, υποθέτοντας ότι τα αποτελέσματα των διαδοχικών δοκιμών, σχηματίζουν μια ακολουθία ανταλλάξιμων δίτιμων τυχαίων μεταβλητών (exchangeable random variables) δίνουμε ακριβή και προσεγγιστικά αποτελέσματα για την κατανομή του μήκους του τεστ, δηλαδή του πλήθους των δοκιμών που απαιτούνται μέχρι τον τερματισμό του τεστ εκκίνησης.

2. ΚΥΡΙΑ ΑΠΟΤΕΛΕΣΜΑΤΑ

Εστω $X_i, i = 1, 2, \dots$, μια ακολουθία ανταλλάξιμων δίτιμων τυχαίων μεταβλητών (exchangeable random variables). Ένα από τα σημαντικότερα θεωρήματα στη στατιστική κατά Bayes που αφορά ακολουθίες ανταλλάξιμων τυχαίων μεταβλητών, το θεώρημα αναπαράστασης de Finetti (de Finetti's Representation Theorem, βλ. π.χ. Jackman 2009, Κεφ. 1), θα παίξει σημαντικό ρόλο στην απόδειξη των αποτελεσμάτων μας. Πριν προχωρήσουμε στη διατύπωση του συγκεκριμένου θεωρήματος, ας αναφέρουμε αρχικά τον ορισμό των ανταλλάξιμων δίτιμων τυχαίων μεταβλητών.

Ορισμός 1. Η πεπερασμένη ακολουθία των δίτιμων τυχαίων μεταβλητών $X_1, X_2, \dots, X_n$, καλείται ανταλλάξιμη (finitely exchangeable) εάν η από κοινού συνάρτηση

πιθανότητας $p(x_1, x_2, \dots, x_n)$ των $X_1, X_2, \dots, X_n$, ικανοποιεί τη σχέση

$$p(x_1, x_2, \dots, x_n) = P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) = p(x_{\pi(1)}, x_{\pi(2)}, \dots, x_{\pi(n)})$$

για κάθε μετάθεση π των δεικτών $\{1, 2, \dots, n\}$ και για κάθε διάνυσμα $(x_1, x_2, \dots, x_n) \in \{0, 1\}^n$. Επιπλέον, η άπειρη ακολουθία των τυχαίων μεταβλητών $X_1, X_2, \dots$ καλείται ανταλλάξιμη (*infinitely exchangeable*) εάν κάθε πεπερασμένη υπακολουθία της είναι ανταλλάξιμη.

Από τον παραπάνω ορισμό, εύκολα κανείς μπορεί να διαπιστώσει ότι:

α) ανταλλάξιμες τυχαίες μεταβλητές έχουν την ίδια κατανομή, και

β) ανεξάρτητες και ισόνομες τυχαίες μεταβλητές είναι ανταλλάξιμες.

Δίνεται στη συνέχεια το θεώρημα αναπαράστασης de Finetti (de Finetti 1931), στην περίπτωση των δίτιμων τυχαίων μεταβλητών.

Θεώρημα 1. Εάν $X_1, X_2, \dots$ είναι μια άπειρη ακολουθία ανταλλάξιμων δίτιμων τυχαίων μεταβλητών με $P(X_i = 1) = p, i = 1, 2, \dots$, τότε

$$p(x_1, x_2, \dots, x_n) = \int_0^1 \prod_{i=1}^n p^{x_i} (1-p)^{1-x_i} dF(p)$$

για κάθε $x_i \in \{0, 1\}, i = 1, 2, \dots$ και $n \in \{1, 2, \dots\}$, όπου $F(p)$ είναι η οριακή κατανομή του p , δηλαδή

$$F(p) = \lim_{n \rightarrow \infty} P\left(\frac{X_1 + X_2 + \dots + X_n}{n} \leq p\right). \quad (1)$$

Βάσει του προηγούμενου θεωρήματος παρατηρούμε ότι οι ανταλλάξιμες δίτιμες τυχαίες μεταβλητές είναι ανεξάρτητες και ισόνομες τυχαίες μεταβλητές, δοθείσας της τιμής πιθανότητας p . Αυτό σημαίνει ότι η έννοια των ανταλλάξιμων τυχαίων μεταβλητών είναι ισοδύναμη με αυτή της υπό συνθήκη ανεξαρτησίας με κοινή κατανομή.

Ας υποθέσουμε ότι το αποτέλεσμα της i δοκιμής του τεστ εκκίνησης περιγράφεται από την τυχαία μεταβλητή $X_i, i = 1, 2, \dots$, όπου

$$X_i = \begin{cases} 1, & \text{εάν το αποτέλεσμα της } i \text{ δοκιμής είναι επιτυχία} \\ 0, & \text{εάν το αποτέλεσμα της } i \text{ δοκιμής είναι αποτυχία.} \end{cases}$$

Μετά την πραγματοποίηση των n πρώτων δοκιμών, η διαδικασία ελέγχου θα έχει τερματιστεί εάν έχουμε απορρίψει ή αποδεχτεί τον εξοπλισμό, διαφορετικά, οι δοκιμές θα πρέπει να συνεχιστούν. Η παρούσα εργασία επικεντρώνεται στη μελέτη του απαιτούμενου αριθμού δοκιμών μέχρι τον τερματισμό του τεστ εκκίνησης, τον οποίο συμβολίζουμε με X . Επομένως, το μήκος του τεστ θα είναι μεγαλύτερο από n εάν και μόνο εάν μετά τις πρώτες δοκιμές ούτε έχουμε απορρίψει ούτε έχουμε αποδεχτεί τον εξοπλισμό.

Ας συμβολίσουμε με $P = \{P_j : j = 1, 2, \dots, M_n\}$ μια οικογένεια συνόλων με τις ακόλουθες ιδιότητες:

- $P_j \subseteq \{1, 2, \dots, n\}$ για όλα τα $j \in \{1, 2, \dots, M_n\}$,
- εάν $X_i = 1$ για κάθε $i \in P_j$, και για τουλάχιστον ένα $j \in \{1, 2, \dots, M_n\}$, τότε ο εξοπλισμός θα έχει γίνει αποδεκτός πριν ή κατά την n -οστή δοκιμή,
- δεν υπάρχει υποσύνολο του P_j με την παραπάνω ιδιότητα (για κάθε $j \in \{1, 2, \dots, M_n\}$).

Επιπλέον, εισάγουμε και την οικογένεια συνόλων $C = \{C_j : j = 1, 2, \dots, N_n\}$ η οποία ικανοποιεί τις παρακάτω ιδιότητες, οι οποίες είναι παρόμοιες με αυτές της οικογένειας P :

- $C_j \subseteq \{1, 2, \dots, n\}$ για όλα τα $j \in \{1, 2, \dots, N_n\}$,
- εάν $X_i = 0$ για κάθε $i \in C_j$, και για τουλάχιστον ένα $j \in \{1, 2, \dots, N_n\}$, τότε ο εξοπλισμός θα έχει απορριφθεί πριν ή κατά την n -οστή δοκιμή,
- δεν υπάρχει υποσύνολο του C_j με την παραπάνω ιδιότητα (για κάθε $j \in \{1, 2, \dots, N_n\}$).

Σύμφωνα με τους παραπάνω συμβολισμούς ο εξοπλισμός θα έχει γίνει αποδεκτός πριν ή κατά την n -οστή δοκιμή, εάν $X_i = 1$ για $i \in P_j$ (για τουλάχιστον ένα $j \in \{1, 2, \dots, M\}$) και ταυτόχρονα δεν έχουμε αποτυχίες σε όλες τις δοκιμές που περιλαμβάνονται στο σύνολο C_j (για κάθε $j \in \{1, 2, \dots, N\}$).

Είναι φανερό ότι οι δυο παραπάνω οικογένειες συνόλων ικανοποιούν ιδιότητες παρόμοιες με αυτές των ελάχιστων συνόλων διακοπής και λειτουργίας, που συναντάμε στη θεωρία αξιοπιστίας (βλ. π.χ. Barlow and Proschan 1981, Κεφ. 1, σελ. 9). Συγκεκριμένα, η οικογένεια $P = \{P_j : j = 1, 2, \dots, M_n\}$ μπορούμε να υποθέσουμε ότι αντιστοιχεί στα ελάχιστα σύνολα λειτουργίας ενός συστήματος με n μονάδες, και επομένως η πιθανότητα ολόκληρο το σύστημα να λειτουργεί είναι ίση με την πιθανότητα να έχουμε αποδεχθεί τον εξοπλισμό πριν ή κατά την n -οστή δοκιμή. Μια παρόμοια αναπαράσταση υπάρχει και για την οικογένεια $C = \{C_j : j = 1, 2, \dots, N_n\}$. Στη συνέχεια και για λόγους απλοποίησης των συμβολισμών τα M_n και N_n , θα γράφονται ως M και N , αντίστοιχα.

Με βάση τις οικογένειες συνόλων P και C , ορίζουμε τις δίτιμες (0-1) συναρτήσεις $\phi_0(\mathbf{x}_n)$, $\phi_1(\mathbf{x}_n)$ και $\phi(\mathbf{x}_n)$ με τον εξής τρόπο:

$$\phi_0(\mathbf{x}_n) = \prod_{j=1}^N \left(1 - \prod_{i \in C_j} (1 - x_i) \right), \quad \phi_1(\mathbf{x}_n) = \prod_{j=1}^M \left(1 - \prod_{i \in P_j} x_i \right),$$

και

$$\phi(\mathbf{x}_n) = \phi_0(\mathbf{x}_n)\phi_1(\mathbf{x}_n),$$

όπου $\mathbf{x}_n = (x_1, x_2, \dots, x_n) \in \{0, 1\}^n$. Είναι φανερό ότι $\phi(\mathbf{X}_n) = 1$ εάν και μόνο εάν $X > n$ και επομένως

$$P(X > n) = E(\phi(\mathbf{X}_n)).$$

Το θεώρημα που ακολουθεί μας προσφέρει ένα γενικό τύπο για την κατανομή του μήκους του τεστ, για κάθε τεστ εκκίνησης με ανταλλάξιμες δοκιμές, για το οποίο είναι γνωστές οι οικογένειες P και C .

Θεώρημα 2. Ας υποθέσουμε ότι οι δοκιμές $X_i, i = 1, 2, \dots$ ενός τεστ εκκίνησης, σχηματίζουν μια άπειρη ακολουθία ανταλλάξιμων δίτιμων τυχαίων μεταβλητών. Τότε ισχύει

$$P(X > n) = \sum_{k=0}^n c_{nk} \int_0^1 p^k (1-p)^{n-k} dF(p), \quad (2)$$

όπου

$$c_{nk} = \sum_{\mathbf{x}_n \in \{0,1\}^n: \mathbf{x}_n \mathbf{x}_n' = k} \phi(\mathbf{x}_n), k = 0, 1, \dots, n,$$

και η $F(p)$ να δίδεται από (1).

Απόδειξη. Αρκεί να παρατηρήσουμε ότι

$$P(X > n) = E(\phi(\mathbf{X}_n)) = \sum_{\mathbf{x}_n \in \{0,1\}^n} \phi(\mathbf{x}_n) P(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n),$$

($\mathbf{x}_n = (x_1, x_2, \dots, x_n)$) και χρησιμοποιώντας το Θεώρημα 1 να πάρουμε

$$\begin{aligned} E(\phi(\mathbf{X}_n)) &= \sum_{k=0}^n \int_0^1 p^k (1-p)^{n-k} dF(p) \left(\sum_{\mathbf{x}_n \in \{0,1\}^n: \mathbf{x}_n \mathbf{x}_n' = k} \phi(\mathbf{x}_n) \right) \\ &= \sum_{k=0}^n c_{nk} \int_0^1 p^k (1-p)^{n-k} dF(p) \end{aligned}$$

με $c_{nk} = \sum_{\mathbf{x}_n \in \{0,1\}^n: \mathbf{x}_n \mathbf{x}_n' = k} \phi(\mathbf{x}_n), k = 0, 1, \dots, n.$ □

Ουσιαστικά οι ποσότητες c_{nk} εκφράζουν τους τρόπους με τους οποίους k επιτυχίες τοποθετούνται σε n διαφορετικές θέσεις έτσι ώστε $\phi(\mathbf{X}_n) = 1$ (άρα, $c_{nk} \in \{0, 1, \dots, \binom{n}{k}\}$). Για παράδειγμα, στο μοντέλο CSTF όπου ένας εξοπλισμός γίνεται αποδεκτός όταν ένας αριθμός από συνεχόμενες επιτυχίες, έστω m , παρατηρηθεί πριν από ένα συνολικό αριθμό αποτυχιών, έστω f (διαφορετικά απορρίπτεται), έχουμε

$$c_{nk} = 0, \text{ εάν } k \leq n - f \text{ και } c_{nk} = N(n, k, m - 1), \text{ εάν } k > n - f$$

για $n \geq \max\{m, f\}$, όπου η ποσότητα (βλ. π.χ. Balakrishnan and Koutras 2002, Κεφ. 2, σελ. 55)

$$N(n, k, m - 1) = \sum_{j=0}^{\lfloor k/m \rfloor} (-1)^j \binom{n - k + 1}{j} \binom{n - j(m)}{n - k}, \quad m - 1 \leq k \leq n$$

εκφράζει τους τρόπους με τους οποίους k επιτυχίες τοποθετούνται σε n διαφορετικές θέσεις, με το πολύ $m - 1$ να είναι συνεχόμενες. Όταν $n < \max\{m, f\}$ παίρνουμε

πολύ απλούστερες εκφράσεις. Επομένως, αντιστοιχώντας μια κατανομή στην τυχαία μεταβλητή p , το Θεώρημα 2 μετατρέπει το πρόβλημα υπολογισμού της κατανομής του μήκους του τεστ, σε ένα πρόβλημα συνδυαστικής.

Σε μερικά τεστ εκκίνησης, με περισσότερα και πιο πολύπλοκα κριτήρια αποδοχής και απόρριψης, ο υπολογισμός των ποσοτήτων c_{nk} μπορεί να αποτελεί μια υπολογιστική διαδικασία, άκρως απαιτητική. Τότε θα ήταν ενδιαφέρον να αναπτυχθούν κάποιες προσεγγίσεις για την κατανομή του μήκους του τεστ. Για το σκοπό αυτό ως συμβολίσουμε με c_{nk}^1 και c_{nk}^0 τις ποσότητες

$$c_{nk}^i = \sum_{\mathbf{x}_n \in \{0,1\}^n : \mathbf{x}_n \mathbf{x}_n' = k} \phi_i(\mathbf{x}_n), i = 0, 1,$$

για $k = 0, 1, \dots, n$. Η ποσότητα c_{nk}^1 (αντίστοιχα η c_{nk}^0) εκφράζει τους τρόπους με τους οποίους k επιτυχίες τοποθετούνται σε n διαφορετικές θέσεις, έτσι ώστε ο εξοπλισμός να μην έχει γίνει αποδεκτός (αντίστοιχα, να μην έχει απορριφθεί) μετά τις n πρώτες δοκιμές. Για παράδειγμα, για το CSTF μοντέλο με $n \geq m$, θα έχουμε

$$c_{nk}^1 = \binom{n}{k}, \text{ εάν } k < m \text{ και } c_{nk}^1 = N(n, k, m - 1), \text{ εάν } k \geq m$$

ενώ εάν $m > n$ τότε $c_{nk}^1 = \binom{n}{k}$ για κάθε $k = 0, 1, \dots, n$. Παράλληλα, εάν $n \geq f$ έχουμε

$$c_{nk}^0 = \binom{n}{k}, \text{ εάν } k > n - f \text{ και } c_{nk}^0 = 0, \text{ εάν } k \leq n - f,$$

ενώ αν $f > n$ τότε $c_{nk}^0 = \binom{n}{k}$ για κάθε $k = 0, 1, \dots, n$. Πρέπει να σημειωθεί ότι, στον ορισμό της ποσότητας c_{nk}^1 (c_{nk}^0), τα κριτήρια απόρριψης (αποδοχής) δεν παίζουν κανένα ρόλο. Επομένως, ο υπολογισμός των παραπάνω ποσοτήτων πραγματοποιείται ευκολότερα από εκείνον των c_{nk} . Στο επόμενο θεώρημα, δίνεται ένα άνω φράγμα για την πιθανότητα $P(X > n)$ το οποίο βασίζεται στις ποσότητες c_{nk}^1 και c_{nk}^0 . Στην απόδειξή του γίνεται χρήση των ιδιοτήτων των συναφών τυχαίων μεταβλητών (associated random variables). Υπενθυμίζεται ότι οι τυχαίες μεταβλητές $X_i, i = 1, 2, \dots$, καλούνται συναφείς εάν για οποιεσδήποτε αύξουσες κατά συντεταγμένες συναρτήσεις $f, g : \mathbb{R}^n \rightarrow \mathbb{R}$ και κάθε n , ισχύει

$$\text{cov}(f(X_1, X_2, \dots, X_n), g(X_1, X_2, \dots, X_n)) \geq 0.$$

Δυο πολύ σημαντικές ιδιότητες των συναφών τυχαίων μεταβλητών είναι οι εξής (βλ. π.χ. Barlow and Proschan 1981, Κεφ. 2, σελ. 30):

- α) οι ανεξάρτητες τυχαίες μεταβλητές είναι συναφείς, και
- β) αύξουσες συναρτήσεις συναφών τυχαίων μεταβλητών, είναι συναφείς.

Θεώρημα 3. Ας υποθέσουμε ότι οι δοκιμές ενός τεστ εκκίνησης $X_i, i = 1, 2, \dots$, σχηματίζουν μια άπειρη ακολουθία ανταλλάξιμων δίτιμων τυχαίων μεταβλητών. Τότε ισχύει

$$P(X > n) \leq \sum_{k=0}^n \sum_{l=0}^n c_{nk}^0 c_{nl}^1 \int_0^1 p^{k+l} (1-p)^{2n-k-l} dF(p) = UB.$$

Απόδειξη. Δοθείσης της πιθανότητας p , οι δίτιμες συναρτήσεις $\phi_0(\mathbf{X}_n)$ και $1 - \phi_1(\mathbf{X}_n)$ είναι αύξουσες συναρτήσεις ανεξάρτητων και ισόνομων τυχαίων μεταβλητών. Επομένως σύμφωνα με τις ιδιότητες που αναφέραμε παραπάνω, δοθείσης της πιθανότητας p , οι $\phi_0(\mathbf{X}_n)$ και $1 - \phi_1(\mathbf{X}_n)$ είναι συναφείς τυχαίες μεταβλητές. Άρα

$$\begin{aligned} \text{cov}(\phi_0(\mathbf{X}_n), 1 - \phi_1(\mathbf{X}_n)|p) &\geq 0 \Rightarrow \\ E(\phi_0(\mathbf{X}_n)\phi_1(\mathbf{X}_n)|p) &\leq E(\phi_0(\mathbf{X}_n)|p)E(\phi_1(\mathbf{X}_n)|p). \end{aligned} \quad (3)$$

Επιπλέον,

$$E(\phi_0(\mathbf{X}_n)|p) = \sum_{k=0}^n c_{nk}^0 p^k (1-p)^{n-k}, \quad E(\phi_1(\mathbf{X}_n)|p) = \sum_{k=0}^n c_{nk}^1 p^k (1-p)^{n-k},$$

και

$$E(\phi_0(\mathbf{X}_n)|p)E(\phi_1(\mathbf{X}_n)|p) = \sum_{k=0}^n \sum_{l=0}^n c_{nk}^0 c_{nl}^1 p^{k+l} (1-p)^{2n-k-l}.$$

Άρα,

$$E[E(\phi_0(\mathbf{X}_n)|p)E(\phi_1(\mathbf{X}_n)|p)] = \sum_{k=0}^n \sum_{l=0}^n c_{nk}^0 c_{nl}^1 \int_0^1 p^{k+l} (1-p)^{2n-k-l} dF(p),$$

και η απόδειξη ολοκληρώνεται λαμβάνοντας υπόψη ότι

$$P(X > n) = E[E(\phi_0(\mathbf{X}_n)\phi_1(\mathbf{X}_n)|p)].$$

□

Με παρόμοιο τρόπο και ως άμεση συνέπεια της ανισότητας

$$1 - \phi_1(\mathbf{X}_n) \geq \phi_0(\mathbf{X}_n)(1 - \phi_1(\mathbf{X}_n)) \quad (4)$$

παίρνουμε το παρακάτω κάτω φράγμα

$$\begin{aligned} P(X > n) &\geq \sum_{k=0}^n c_{nk}^1 \int_0^1 p^k (1-p)^{n-k} dF(p) \\ &\quad + \sum_{k=0}^n c_{nk}^0 \int_0^1 p^k (1-p)^{n-k} dF(p) - 1 = LB. \end{aligned}$$

Στον Πίνακα 1 δίνονται ορισμένα αριθμητικά αποτελέσματα για την αξιολόγηση των παραπάνω δυο φραγμάτων, για την περίπτωση του μοντέλου CSTF όταν η p ακολουθεί μια κατανομή Βήτα με παραμέτρους $\alpha > 0, \beta > 0$, δηλαδή

$$dF(p) = \frac{1}{B(\alpha, \beta)} p^{\alpha-1} (1-p)^{\beta-1} dp, \text{ με } B(\alpha, \beta) = \int_0^1 p^{\alpha-1} (1-p)^{\beta-1} dp.$$

Είναι φανερό ότι και για μικρές τιμές του n οι προσεγγίσεις είναι πολύ ικανοποιητικές.

Αξίζει επίσης να αναφέρουμε ότι, με βάση τις οικογένειες P και C , μπορούμε να ορίσουμε δυο διαφορετικά συστήματα αξιοπιστίας που το καθένα θα έχει συνάρτηση δομής την $\phi_0(\mathbf{X}_n)$ και την $1 - \phi_1(\mathbf{X}_n)$, αντιστοίχως. Συμβολίζοντας με R_0 και R_1 τις αντίστοιχες πιθανότητες λειτουργίας των συστημάτων δοθείσης της πιθανότητας p (δηλαδή, τις αξιοπιστίες των συστημάτων - βλ. π.χ. Barlow and Proschan 1981, Κεφ. 2, σελ. 20), έχουμε

$$R_0 = E[\phi_0(\mathbf{X}_n)|p] \text{ και } R_1 = E[1 - \phi_1(\mathbf{X}_n)|p].$$

Κάνοντας χρήση των ανισοτήτων (3) και (4) καταλήγουμε στις επόμενες ανισότητες

$$R_0 - R_1 \leq E[\phi_0(\mathbf{X}_n)\phi_1(\mathbf{X}_n)|p] \leq R_0(1 - R_1).$$

Αξίζει να σημειώσουμε ότι, εάν οι ποσότητες c_{nk}^1 και c_{nk}^0 δεν μπορούν να υπολογιστούν εύκολα, θα μπορούσαμε να χρησιμοποιήσουμε τα παραπάνω φράγματα σε συνδυασμό με γνωστά φράγματα από τη θεωρία αξιοπιστίας, για συστήματα με ανεξάρτητες και ισόνομες μονάδες (όπως τα Esary-Proschan και Min-Max, βλ. π.χ. Barlow and Proschan 1981, Κεφ. 2, σελ. 32). Τότε θα παίρναμε ένα πιο απλό σύνολο φραγμάτων τα οποία όμως θα παρείχαν μικρότερη ακρίβεια. Για παράδειγμα, φράγματα από τη θεωρία αξιοπιστίας κατάλληλα να χρησιμοποιηθούν για αυτό το σκοπό είναι τα εξής

$$\begin{aligned} L(p) &= p^n, U(p) = 1 - (1-p)^n, \\ L(p) &= p^{\min_j |P_j|}, U(p) = 1 - (1-p)^{\min_j |C_j|}, \\ L(p) &= \prod_{j=1}^N (1 - (1-p)^{|C_j|}), U(p) = 1 - \prod_{j=1}^M (1 - p^{|P_j|}). \end{aligned}$$

ABSTRACT

In this paper, we study the class of binary start-up demonstration tests under a general framework. Assuming that the outcomes of the start-up tests are described by a sequence of exchangeable random variables, we develop a general form for the exact waiting time distribution associated with the length of the test. Several approximations for the tail probabilities of this distribution are also derived.

α	β	m	f	n	LB	$P(X > n)$	UB	$UB - LB$
9.5	1.5	10	3	15	0.197201	0.210283	0.274012	0.076811
				20	0	0.049337	0.126171	0.126171
				25	0	0.007673	0.060454	0.060454
5.5	5.5	20	5	25	0.028292	0.028295	0.029044	0.000752
				30	0.012699	0.012894	0.014121	0.001422
				35	0.005093	0.005914	0.007273	0.002180

Πίνακας 1: Φράγματα για την $P(X > n)$ στο μοντέλο $CSTF$.

ΑΝΑΦΟΡΕΣ

- Aki, S., Kuboki, H. and Hirano, K. (1984). On discrete distributions of order k . *Annals of the Institute of Statistical Mathematics*, **36**, 431-440.
- 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**, 1-12.
- 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. and Koutras, M.V. (2002). *Runs and Scans with Applications*. John Wiley & Sons, New York.
- 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., 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.
- Barlow, R. and Proschan, F. (1981). *Statistical Theory of Reliability and Life Testing*. Holt, Reinhart and Winston, New York.
- Chan, P.S., Ng, H.K.T. 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.
- de Finetti, B. (1931). Funcione caratteristica di un fenomeno aleatorio. *Atti della Reale Accademia Nazionale dei Lincii*, **4**, 251-299.
- Eryilmaz, S.(2010). Start-up demonstration tests under Markov dependence. *Pakistan Journal of Statistics*, **26**, 637-647.
- Eryilmaz, S. and Chakraborti, S. (2008). On start-up demonstration tests under exchangeability. *IEEE Transactions on Reliability*, **57**, 627-632.
- Gera, A.E. (2010). A new start-up demonstration test. *IEEE Transactions on Reliability*, **59**, 128-131.
- Gera, A.E. (2011). A general model for start-up demonstration tests. *IEEE*

- Transactions on Reliability*, **60**, 295-304.
- Hahn, G.J. and Gage, J.B. (1983). Evaluation of a start-up demonstration test. *Journal of Quality Technology*, **15**, 103-106.
- Jackman, S. (2009). *Bayesian Analysis for the Social Sciences*. John Wiley & Sons, Hoboken, New Jersey.
- Koutras, M.V. and Balakrishnan, N. (1999). A start-up demonstration test using a simple scan-based statistic. In *Scan Statistics and Applications* (J. Glaz and N. Balakrishnan, eds), pp. 251-267, Birkhauser, Boston.
- 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.
- Martin, D.E.K. (2008). Application of auxiliary Markov chains to start-up demonstration tests. *European Journal of Operational Research*, **184**, 574-582.
- 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. and Griffith, W.S. (2005). Start-up demonstration tests based on consecutive successes and total failures. *Journal of Quality Technology*, **37**, 186-198.
- Smith, M.L. and Griffith, W.S. (2008). The analysis and comparison of start-up demonstration tests. *European Journal of Operational Research*, **186**, 1029-1045.
- Smith, M.L. and Griffith, W.S. (2011). Multistate start-up demonstration tests. *International Journal of Reliability, Quality and Safety Engineering*, **18**, 99-117.
- Viveros, R. and Balakrishnan, N. (1993). Statistical inference from start-up demonstration test data. *Journal of Quality Technology*, **25**, 119-130.
- Yalcin, F. and Eryilmaz, S. (2012). Start-up demonstration test based on total successes and total failures with dependent start-ups. *IEEE Transactions on Reliability*, **61**, 227-230.

ΠΡΟΣΑΡΜΟΓΗ ΠΡΟΤΥΠΩΝ POISSON-ΓΑΜΜΑ ΜΕ ΤΗ ΜΕΘΟΔΟ ΜΙΚΤΩΝ ΡΟΠΩΝ

Ι.Α. Κουτρουβέλης

Γενικό Τμήμα, Πανεπιστήμιο Πατρών

koutrouv@upatras.gr

ΠΕΡΙΛΗΨΗ

Η σύνθετη κατανομή Poisson-γάμμα (*PGAM*) έχει εφαρμοστεί σε διάφορους επιστημονικούς τομείς όπως στη στοχαστική υδρολογία, τη μετεωρολογία, τη θεωρία ρίσκου και τη σεισμολογία. Στην εργασία αυτή παρουσιάζεται μια μέθοδος μικτών ροπών για την εκτίμηση των τριών παραμέτρων της κατανομής *PGAM*. Η μέθοδος βασίζεται στην εμπειρική συνάρτηση ημιαναλλοιώτων και στις δύο πρώτες δειγματικές ροπές. Βρίσκεται η ασυμπτωτική κατανομή των προκυπτουσών εκτιμητριών και η προτεινόμενη μέθοδος συγκρίνεται με τη μέθοδο των ροπών.

Λέξεις Κλειδιά: σύνθετη Poisson-γάμμα κατανομή, εμπειρική συνάρτηση ημιαναλλοιώτων, μέθοδος μικτών ροπών

1. ΕΙΣΑΓΩΓΗ

Η σύνθετη Poisson-γάμμα κατανομή, $PGAM(\lambda, \alpha, \beta)$, ορίζεται ως η κατανομή της τυχαίας μεταβλητής (τ.μ.) X όταν η X μπορεί να παρασταθεί ως

$$X = \begin{cases} \sum_{m=1}^N Y_m, & \text{αν } N > 0 \\ 0, & \text{αν } N = 0 \end{cases},$$

όπου N είναι τ.μ. με κατανομή Poisson και μέση τιμή λ , Y_m ($m=1, 2, \dots$) είναι τ.μ.'ς που έχουν κατανομή γάμμα με παράμετρο μορφής α και παράμετρο κλίμακας β και $N, Y_1, Y_2, \dots$ είναι ανεξάρτητες. Τρεις διαφορετικοί χαρακτηρισμοί (τρόποι περιγραφής) του προτύπου είναι οι παρακάτω:

- Κατανομή με Poisson τερματισμό αθροίσματος και γάμμα προσθετέους.
- Γενικευμένη Poisson με γενικεύουσα κατανομή γάμμα.
- Poisson μίξη μιας οριακής κατανομής με μάζα στο 0 και κατανομών γάμμα.

Η $PGAM$ είναι κατανομή μικτού τύπου με αθροιστική συνάρτηση κατανομής

$$F(x) = \begin{cases} e^{-\lambda} + \sum_{n=1}^{\infty} \frac{\lambda^n e^{-\lambda}}{n!} \int_0^x \frac{1}{\beta^{n\alpha} \Gamma(n\alpha)} y^{n\alpha-1} e^{-y/\beta} dy, & x > 0 \\ e^{-\lambda}, & x = 0 \\ 0, & x < 0. \end{cases}$$

Η σ.π.π. της X δεν ορίζεται με τη συνήθη μορφή, αλλά στη μορφή

$$f(x) = e^{-\lambda} \delta(x) + (1 - e^{-\lambda}) f^+(x),$$

όπου $\delta(x)$ η συνάρτηση δέλτα του Dirac (βλ. π.χ. Arfken & Weber, 2000) και $f^+(x)$ η συνάρτηση πυκνότητας πιθανότητας της θετικής συνιστώσας $X_+ = X(X > 0)$. Η τελευταία συνάρτηση είναι

$$f^+(x) = d[P(X \leq x | X > 0)] / dx = (e^\lambda - 1)^{-1} \exp(-x/\beta) x^{-1} \sum_{n=1}^{\infty} \frac{\lambda^n x^{n\alpha}}{n! \beta^{n\alpha} \Gamma(n\alpha)}.$$

Το λ ($\lambda > 0$) είναι παράμετρος μορφής που καθορίζει την μάζα πιθανότητας στο 0. Το α ($\alpha > 0$) είναι επίσης παράμετρος μορφής ενώ το β είναι παράμετρος κλίμακας. Έτσι, αν $X \sim PGAM(\lambda, \alpha, \beta)$, τότε $X/\beta \sim PGAM(\lambda, \alpha, 1)$.

Όταν είναι γνωστό ότι $\alpha = 1$, έχουμε την ειδική περίπτωση σύνθετης Poisson-εκθετικής κατανομής, (PE). Το πρότυπο αυτό συνδέεται με τη μη κεντρική κατανομή χι-τετράγωνο με 0 βαθμούς ελευθερίας, $\chi_0^2(\lambda)$, μέσω της σχέσης

$$PE(\lambda, \beta) = (\beta/2) \chi_0^2(2\lambda).$$

Ο Siegel (1985) απέδειξε την ύπαρξη και μοναδικότητα εκτιμητών μέγιστης πιθανοφάνειας (ε.μ.π.) $\hat{c}, \hat{\lambda}$ της $c\chi_0^2(\lambda)$ και ανέπτυξε μια υπολογιστική μέθοδο προσδιορισμού τους που κάνει χρήση της τροποποιημένης συνάρτησης Bessel I_0^u είδους. Οι Koutrouvelis & Meintanis (2002) εφήρμοσαν αυτήν τη μέθοδο και την συνέκριναν με δύο μεθόδους μικτών ροπών για την εκτίμηση των παραμέτρων της $PE(\lambda, \beta)$ που είναι υπολογιστικά απλούστερες. Οι τελευταίες μέθοδοι βασίζονται σε ένα πρότυπο παλινδρόμησης που εμπλέκει μια τυποποιημένη μορφή της συνάρτησης ημιαναλλοιώτων σε συνδυασμό με την εξίσωση πρώτης ροπής και παράγουν εκτιμήτριες με τουλάχιστον 80% αποδοτικότητα σε σχέση με τις ε.μ.π. Στη γενικότερη περίπτωση της $PGAM(\lambda, \alpha, \beta)$ με (οποιαδήποτε) γνωστή τιμή της α , οι Withers & Nadarajah (2011) απέδειξαν την ύπαρξη ε.μ.π. των παραμέτρων λ, β και πρότειναν αριθμητική μέθοδο προσδιορισμού τους.

Παρότι μπορούν να οριστούν ε.μ.π. στην γενική περίπτωση που και οι τρεις παράμετροι της $PGAM(\lambda, \alpha, \beta)$ είναι άγνωστες, η εύρεσή τους είναι περίπλοκη [akward κατά τους Withers & Nadarajah (2011)] και από ό,τι γνωρίζουμε δεν έχουν προταθεί υπολογιστικές μέθοδοι προσδιορισμού τους. Επομένως, πέραν από τη

μέθοδο των ροπών (MOM) που μπορεί να εφαρμοστεί στη γενική περίπτωση, δεν έχει αναπτυχθεί άλλη μέθοδος προσαρμογής της $PGAM(\lambda, \alpha, \beta)$. Στην εργασία αυτή παρουσιάζεται μια μέθοδος μικτών ροπών (MXM) για την προσαρμογή της κατανομής $PGAM(\lambda, \alpha, \beta)$. Η προτεινόμενη μέθοδος συνδυάζει την εξίσωση παλινδρόμησης μιας τυποποιημένης μορφής της συνάρτησης ημιαναλλοιώτων με τις εξισώσεις των δύο πρώτων ακέραιων ροπών.

Η παρουσίαση του υπόλοιπου μέρους της εργασίας έχει οργανωθεί ως ακολούθως: Στην Ενότητα 2 γίνεται μια σύντομη αναφορά της εφαρμογής της $PGAM(\lambda, \alpha, \beta)$ στην περιγραφή της συνολικής βροχόπτωσης/χιονόπτωσης καθώς και άλλων εφαρμογών του προτύπου. Η Ενότητα 3 περιέχει πληροφορίες για τις γεννήτριες συναρτήσεις που χρησιμοποιούνται στη εργασία και τις ροπές της $PGAM(\lambda, \alpha, \beta)$. Στην Ενότητα 4 παρουσιάζονται η μέθοδος MOM και η προτεινόμενη μέθοδος MXM για την εκτίμηση των παραμέτρων λ , α και β . Στην Ενότητα 5 δίνεται η ασυμπτωτική κατανομή των εκτιμητριών μικτών ροπών. Τέλος, η Ενότητα 6 περιέχει τα αποτελέσματα μιας μελέτης Monte Carlo τα οποία δείχνουν ότι με κατάλληλη επιλογή σημείων στα οποία υπολογίζεται η εμπειρική ροπογεννήτρια συνάρτηση η μέθοδος MXM παράγει σημαντικά υπέρτερες εκτιμήτριες ως προς τη μεροληπτικότητα και το μέσο τετραγωνικό σφάλμα από ό,τι η μέθοδος MOM για χρήσιμες, στις εφαρμογές, περιοχές του παραμετρικού χώρου.

2. ΕΦΑΡΜΟΓΕΣ

Η περισσότερο διαδεδομένη εφαρμογή της $PGAM$ είναι για την περιγραφή της συνολικής βροχόπτωσης ή χιονόπτωσης σε συγκεκριμένο χρονικό διάστημα (μήνας/ημέρα). Αρχικά προτάθηκε ως πρότυπο για τη συνολική βροχόπτωση από τους Fisher & Cornish (1960). Ο Revfeim (1984), εφήρμοσε την ειδική περίπτωση του προτύπου της κατανομής PE υποθέτοντας μια ομογενή διαδικασία Poisson για την εμφάνιση των βροχοπτώσεων και την εκθετική για την ποσότητα Y_m κάθε γεγονότος βροχόπτωσης σε καθορισμένο χρονικό διάστημα. Το πρότυπο της PE έχει εφαρμοστεί και στην περίπτωση μη ομογενούς διαδικασίας Poisson για την περιγραφή της συνολικής βροχόπτωσης σε έναν χρόνο, βλ. Alexanderson (1986).

Στη συνέχεια αναφέρονται μεταβλητές για την περιγραφή των οποίων βρίσκει εφαρμογή το πρότυπο της $PGAM$:

- Συνολικές απαιτήσεις ασφαλιστικών αποζημιώσεων.
- Συνολικές εισπράξεις καταστήματος.
- Συνολική έκκληση ενέργειας κύριων σεισμών στα Βαλκάνια.
- Συνολικός χρόνος εξυπηρέτησης σε ουρά αναμονής με Poisson αφίξεις, γάμμα χρόνους εξυπηρέτησης και ∞ εξυπηρετητές.

3. ΓΕΝΝΗΤΡΙΕΣ ΣΥΝΑΡΤΗΣΕΙΣ ΚΑΙ ΡΟΠΕΣ

3.1 Θεωρητικές συναρτήσεις και ροπές

Η ροπογεννήτρια συνάρτηση μιας τ.μ. X με κατανομή $PGAM(\lambda, \alpha, \beta)$ είναι

$$\psi(u) = E(e^{uX}) = \exp\left\{\lambda \left[(1 - \beta u)^{-\alpha} - 1\right]\right\}, \quad u < \frac{1}{\beta}.$$

Επομένως, η γεννήτρια συνάρτηση ημιαναλλοιώτων είναι

$$K(u) = \lambda \left[(1 - \beta u)^{-\alpha} - 1\right] = \lambda \sum_{j=1}^{\infty} \frac{\alpha(\alpha+1)\cdots(\alpha+j-1)}{j!} (\beta u)^j$$

και η ημιαναλλοιώτος j -οστής τάξης δίνεται από τη σχέση

$$k_j = \lambda \alpha(\alpha+1)\cdots(\alpha+j-1)\beta^j \quad (1)$$

Οι κεντρικές ροπές τάξεων 2-4 δίνονται από τις σχέσεις

$$\mu_2 = \lambda \alpha(\alpha+1)\beta^2, \quad \mu_3 = \lambda \alpha(\alpha+1)(\alpha+2)\beta^3,$$

$$\mu_4 = \lambda \alpha(\alpha+1)[(\alpha+2)(\alpha+3) + 3\lambda(\alpha+1)]\beta^4,$$

ενώ οι τρεις πρώτες ροπές περί την αρχή είναι

$$\mu = \lambda \alpha \beta, \quad E(X^2) = \lambda [\alpha(\alpha+1) + \lambda \alpha^2] \beta^2,$$

$$E(X^3) = \lambda \alpha [(\alpha+1)(\alpha+2) + 3\lambda \alpha(\alpha+1) + \lambda^2 \alpha^2] \beta^3.$$

Γενικοί τύποι ροπών έχουν δοθεί από τους Withers & Nadarajah (2011) με χρήση πολυωνύμων Bell.

3.2 Εμπειρικές γεννήτριες συναρτήσεις

Έστω $x_1, x_2, \dots, x_n$ τυχαίο δείγμα από πληθυσμό με κατανομή $PGAM(\lambda, \alpha, \beta)$.

Ορίζουμε την εμπειρική ροπογεννήτρια συνάρτηση ($EP\Sigma$), $\psi_n(u)$, ως

$$\psi_n(u) = \frac{1}{n} \sum_{i=1}^n e^{ux_i}$$

και την εμπειρική συνάρτηση ημιαναλλοιώτων ($E\Sigma H$), $K_n(u)$, ως

$$K_n(u) = \ln \psi_n(u).$$

Επειδή η συνάρτηση $K(u/\beta) = \lambda [(1-u)^{-\alpha} - 1]$ δεν εξαρτάται από την παράμετρο β , η μέθοδος μικτών ροπών που προτείνεται εδώ χρησιμοποιεί τη στατιστική συνάρτηση

$$T_n(u) = K_n(u/\hat{\beta}) = \ln \psi_n(u/\hat{\beta}), \quad u < 1,$$

με $\hat{\beta}$ μια συνεπή εκτιμήτρια της β .

4. ΕΚΤΙΜΗΣΗ ΠΑΡΑΜΕΤΡΩΝ

4.1 Μέθοδος ροπών (MOM)

Έστω

$$\kappa_1 = \bar{x}, \quad \kappa_2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}, \quad \kappa_3 = \frac{\sum_{i=1}^n (x_i - \bar{x})^3}{n}$$

οι τρεις πρώτες δειγματικές ημιαναλλοιώτες. Με χρήση του τύπου (1) και επίλυση του συστήματος των εξισώσεων των τριών πρώτων ημιαναλλοιώτων με τις αντίστοιχες δειγματικές ημιαναλλοιώτες ως προς τις παραμέτρους λ , α , β προκύπτουν οι ακόλουθοι τύποι των εκτιμητών ροπών

$$\hat{\lambda}_0 = \kappa_1^2 \kappa_2^{-1} (2 - \kappa_1 \kappa_2^{-2} \kappa_3)^{-1}, \quad \hat{\alpha}_0 = (\kappa_1 \kappa_2^{-2} \kappa_3 - 1)^{-1} - 1, \quad \hat{\beta}_0 = \kappa_2^{-1} \kappa_3 - \kappa_2 \kappa_1^{-1}. \quad (2)$$

4.2 Μέθοδος μικτών ροπών (MXM)

Η μέθοδος MXM είναι μια επαναληπτική διαδικασία που χρησιμοποιεί το απλό πρότυπο παλινδρόμησης

$$K_n(u / \hat{\beta}) = \lambda [(1-u)^{-\hat{\alpha}} - 1] + \varepsilon_n(u)$$

και τις δύο πρώτες εξισώσεις ροπών $\mu = \bar{x}$, $E(X^2) = m'_2$, όπου $m'_2 = \sum x_i^2 / n$. Τα βασικά βήματα της διαδικασίας είναι τα ακόλουθα:

Βήμα 0: Εύρεση των εκτιμητών ροπών $\hat{\lambda}_0, \hat{\alpha}_0, \hat{\beta}_0$

Βήμα j (j ≥ 1): Εύρεση του εκτιμητή $\hat{\lambda}_j$ με εφαρμογή συνήθων ελαχίστων τετραγώνων στο απλό πρότυπο παλινδρόμησης

$$\ln \psi_n(u_k / \hat{\beta}_{j-1}) = \lambda [(1-u_k)^{-\hat{\alpha}_{j-1}} - 1] + \varepsilon_n(u_k) \quad [u_k = -k\tau, k = 1, 2, \dots, r],$$

όπου τ ($\tau > 0$), r (r ακέραιος) κατάλληλα προεπιλεγμένες παράμετροι χρήσης,

και των εκτιμητών $\hat{\alpha}_j, \hat{\beta}_j$ από τις σχέσεις

$$\hat{\alpha}_j = \bar{x}^2 / (\hat{\lambda}_j m_2 - \bar{x}^2), \quad \hat{\beta}_j = \bar{x} / (\hat{\lambda}_j \hat{\alpha}_j).$$

Η διαδικασία σταματά στο βήμα J αν $|(\hat{\beta}_J - \hat{\beta}_{J-1}) / \hat{\beta}_{J-1}| < \varepsilon$, όπου ε ($\varepsilon > 0$) η προεπιλεγμένη σταθερά σύγκλισης. Λόγω της αλληλεξάρτησης των $\hat{\lambda}_j, \hat{\alpha}_j, \hat{\beta}_j$, το κριτήριο σύγκλισης που επιλέχθηκε βασίζεται μόνο στον εκτιμητή $\hat{\beta}_j$ που υπολογίζεται μετά τους εκτιμητές $\hat{\lambda}_j, \hat{\alpha}_j$.

Οι προκύπτοντες εκτιμητές MXM των παραμέτρων λ και α είναι αμετάβλητοι κατά κλίμακα ενώ ο εκτιμητής MXM της παραμέτρου β ισομεταβάλλεται με την αλλαγή κλίμακας. Το ίδιο βέβαια ισχύει για τους εκτιμητές ροπών.

5. ΑΣΥΜΠΤΩΤΙΚΗ ΚΑΤΑΝΟΜΗ ΕΚΤΙΜΗΤΡΙΩΝ ΜΙΚΤΩΝ ΡΟΠΩΝ

Έστω $\hat{\boldsymbol{\theta}} = (\hat{\lambda}_J, \hat{\alpha}_J, \hat{\beta}_J)'$ το διάνυσμα των τελικών εκτιμητριών που αντιστοιχούν στη διαδικασία MXM και $\boldsymbol{\theta} = (\lambda, \alpha, \beta)'$. Ισχύει το επόμενο θεώρημα.

Θεώρημα: Κάτω από το πρότυπο της $PGAM(\lambda, \alpha, \beta)$,

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{L} N_3(\mathbf{0}, \mathbf{V}),$$

όπου

$$\mathbf{V} = \mathbf{A}^{-1} \begin{bmatrix} \mathbf{s}'\boldsymbol{\Sigma}\mathbf{s} & \mathbf{s}'\mathbf{a} & \mathbf{s}'\mathbf{b} \\ \mathbf{s}'\mathbf{a} & k_2 & c \\ \mathbf{s}'\mathbf{b} & c & d \end{bmatrix} (\mathbf{A}^{-1})',$$

$\mathbf{s}$, $\mathbf{a}$, $\mathbf{b}$ είναι διανύσματα στήλης τάξης r με k -στές συνιστώσες

$$s_k \equiv s_k(\alpha) = (1 - u_k)^{-\alpha} - 1, \quad a_k = \lambda\alpha\beta[(1 - u_k)^{-(\alpha+1)} - 1],$$

$$b_k = \lambda\alpha\beta^2\{(\alpha+1)[(1 - u_k)^{-(\alpha+2)} - 1] + \lambda\alpha[(1 - u_k)^{-2(\alpha+1)} - 1]\},$$

$\boldsymbol{\Sigma} = (\sigma_{kl})$ είναι πίνακας διαστάσεων $r \times r$ με στοιχεία

$$\sigma_{kl} = \exp\{\lambda\{[1 - (u_k + u_l)]^{-\alpha} - (1 - u_k)^{-\alpha} - (1 - u_l)^{-\alpha} + 1\}\} - 1,$$

$$\mathbf{A} = \begin{bmatrix} \sum s_k^2 & -\lambda s_k(1 - u_k)^{-\alpha} \ln(1 - u_k) & \lambda\alpha\beta^{-1} \sum u_k s_k (1 - u_k)^{-(\alpha+1)} \\ \alpha\beta & \lambda\beta & \lambda\alpha \\ \alpha\beta^2(2\lambda\alpha + \alpha + 1) & \lambda\beta^2(2\lambda\alpha + 2\alpha + 1) & 2\lambda\beta[\alpha(\alpha + 1) + \lambda\alpha^2] \end{bmatrix},$$

$$c = \lambda\beta^3\alpha(\alpha + 1)(2\lambda\alpha + \alpha + 2)$$

και

$$d = \lambda\alpha(\alpha + 1)\beta^4[4\lambda^2\alpha + 2\lambda\alpha(\alpha + 1) + 4\lambda\alpha(\alpha + 2) + (\alpha + 2)(\alpha + 3)].$$

Απόδειξη:

Με την υπόθεση ότι αρχίζουμε την επαναληπτική διαδικασία από μια συνεπή εκτιμήτρια $\hat{\boldsymbol{\theta}}_0$, όπως είναι η εκτιμήτρια ροπών, η τελική εκτιμήτρια μικτών ροπών είναι μια συνεπής λύση του συστήματος:

$$L_1 \equiv \sum_{k=1}^r s_k(\alpha) \left[\ln \psi_n \left(\frac{u_k}{\beta} \right) - \lambda s_k(\alpha) \right] = 0,$$

$$L_2 \equiv \bar{x} - \lambda\alpha\beta = 0,$$

$$L_3 \equiv m'_2 - \lambda[\alpha(\alpha + 1) + \lambda\alpha^2]\beta^2.$$

Η 1^η ισότητα βρίσκεται εξισώνοντας με 0 την μερική παράγωγο ως προς λ του αθροίσματος τετραγώνων των σφαλμάτων στο πρότυπο

$$\ln \psi_n(u_k / \beta) = \lambda s_k(\alpha) + \varepsilon_{nk}.$$

Η 2^η και 3^η ισότητα βρίσκονται εξισώνοντας μ και $E(X^2)$ με $\bar{x}$ και m'_2 , αντίστοιχα.

Αν συμβολίσουμε με $\mathbf{L}$ το διάνυσμα στήλης με συνιστώσες L_1, L_2, L_3 , από το θεώρημα της μέσης τιμής και το θεώρημα του Slutsky μπορεί να αποδειχθεί ότι

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \approx \mathbf{A}^{-1} \sqrt{n} \mathbf{L}(\boldsymbol{\theta}),$$

όπου $\approx$ συμβολίζει ισότητα ασυμπτωτικών κατανομών. Επίσης,

$$\sqrt{n} \mathbf{L}(\boldsymbol{\theta}) = \begin{bmatrix} \mathbf{s}' & 0 & 0 \\ \mathbf{0}' & 1 & 0 \\ \mathbf{0}' & 0 & 1 \end{bmatrix} \sqrt{n} \begin{bmatrix} \xi_n(\beta) - \xi(\lambda, \alpha) \\ \bar{x} - \lambda\alpha\beta \\ m'_2 - \lambda[\alpha(\alpha+1) + \lambda\alpha^2]\beta^2 \end{bmatrix},$$

όπου $\xi_n(\beta)$ και $\xi(\lambda, \alpha)$ είναι διανύσματα στήλης με k -στές συνιστώσες

$$\xi_{nk}(\beta) = \ln \psi_n(u_k / \beta) \quad \text{και} \quad \xi_k(\lambda, \alpha) = \ln \psi(u_k / \beta) = \lambda s_k(\alpha).$$

Από το Κ.Ο.Θ. προκύπτει το συμπέρασμα του θεωρήματος αφού ισχύουν οι σχέσεις

$$\begin{aligned} \sqrt{n} [\xi_n(\beta) - \xi(\lambda, \alpha)] &\xrightarrow{L} N_r(\mathbf{0}, \boldsymbol{\Sigma}), \\ \lim_{n \rightarrow \infty} n \text{cov}[\xi_{nk}(\beta), \bar{x}] &= a_k, \quad \lim_{n \rightarrow \infty} n \text{cov}[\xi_{nk}(\beta), m'_2] = b_k, \\ \lim_{n \rightarrow \infty} n \text{var}(\bar{x}) &= k_2, \quad \lim_{n \rightarrow \infty} n \text{cov}[\bar{x}, m'_2] = c, \quad \lim_{n \rightarrow \infty} n \text{var}(m'_2) = d. \end{aligned}$$

6. ΕΠΙΛΟΓΗ ΠΑΡΑΜΕΤΡΩΝ ΧΡΗΣΗΣ ΚΑΙ ΑΠΟΤΕΛΕΣΜΑΤΑ ΠΕΙΡΑΜΑΤΩΝ ΠΡΟΣΟΜΟΙΩΣΗΣ

Οι βασικές παράμετροι χρήσης της μεθόδου MXM είναι οι σταθερές r (πλήθος σημείων $u_k = -k\tau$ για την προσαρμογή ευθείας παλινδρόμησης) και τ (απόσταση μεταξύ των σημείων u_k). Η ασυμπτωτική αποτελεσματικότητα των εκτιμητριών μικτών ροπών εξαρτάται κυρίως από το διάστημα του αρνητικού ημιάξονα που καλύπτουν τα σημεία $\{u_k; k=1, 2, \dots, r\}$ και όχι από την ακριβή θέση τους. Η αναζήτηση κατάλληλων u_k απλουστεύεται με επιλογή ισαπεχόντων σημείων, προσδιορισμό μιας τιμής του r που δε δημιουργεί υπολογιστικά προβλήματα σε μικρά δείγματα και εύρεση κατάλληλης απόστασης τ μεταξύ των σημείων.

Από αρχικά πειράματα προσομοίωσης βρέθηκε ότι με την επιλογή $r=40$, $\tau=16$ και δείγματα μεγέθους $n \geq 50$ προκύπτουν εκτιμήτριες μικτών ροπών με πολύ μικρότερες τιμές μεροληψίας και μέσου τετραγωνικού σφάλματος από τις αντίστοιχες τιμές των εκτιμητριών ροπών όταν οι πραγματικές τιμές των λ και α δεν είναι ούτε πολύ μικρές ούτε πολύ μεγάλες, δηλαδή για τιμές των παραμέτρων που αντιστοιχούν στις περισσότερες εφαρμογές. Για μεγάλες πραγματικές τιμές των παραμέτρων λ και α η μέθοδος των ροπών υπερτερεί της μεθόδου των μικτών ροπών. Για δείγματα μικρού μεγέθους δεν μπορεί να γίνει ικανοποιητική προσαρμογή της κατανομής PGAM αφού και οι δύο μέθοδοι παράγουν εκτιμήτριες με μεγάλες τιμές μεροληψίας και μέσου τετραγωνικού σφάλματος.

Στον Πίνακα 1 δίνονται οι μέσες τιμές και τα πολλαπλασιασμένα με το μέγεθος του δείγματος μέσα τετραγωνικά σφάλματα (NMSE) των εκτιμητριών των τριών

παραμέτρων που προέκυψαν με εφαρμογή της μεθόδου ροπών (MOM) και της μεθόδου μικτών ροπών (MXM) σε 1000 δείγματα μεγέθους 50 και 100 από κατανομές $PGAM(\lambda, \alpha, \beta)$ με $\lambda = 0.5, 1, 2$, $\alpha = 0.5, 1, 2$ και $\beta = 1$. Μπορεί να διαπιστωθεί ότι η μέθοδος MXM υπερτερεί κατά πολύ της μεθόδου MOM στην εκτίμηση καθεμιάς παραμέτρου εκτός από την εκτίμηση της λ για $n = 50$, $\lambda = 0.5$ και $\alpha = 0.5$. Πρέπει να σημειωθεί ότι η μέθοδος MOM αποτυγχάνει όταν το δείγμα περιέχει μόνο μηδενικές τιμές. Επίσης, αν ένα δείγμα παρουσιάζει αρνητική λοξότητα, οι εκτιμητές ροπών $\hat{\alpha}_0$ και $\hat{\beta}_0$ δεν είναι αποδεκτοί αφού λόγω της (2) παίρνουν αρνητικές τιμές. Σε αυτές τις περιπτώσεις τα δείγματα απορρίφθηκαν και δεν ελήφθησαν υπόψη στα αποτελέσματα των πειραμάτων προσομοίωσης.

Στον Πίνακα 2 δίνονται ο μέσοι και τα NMSE που βρέθηκαν με εφαρμογή της μεθόδου MXM σε 1000 δείγματα μεγέθους $n = 1000$ από την $PGAM(\lambda, \alpha, \beta)$, καθώς και οι ασυμπτωτικές διασπορές των εκτιμητριών MXM στις ίδιες περιπτώσεις πραγματικών τιμών των παραμέτρων όπως αυτές του Πίνακα 1. Βλέπουμε ότι οι εμπειρικές μέσες τιμές είναι πολύ κοντά στις πραγματικές τιμές των παραμέτρων και τα NMSE πλησιάζουν τις αντίστοιχες ασυμπτωτικές τιμές διασπορών. Τα αποτελέσματα αυτά επιβεβαιώνουν την ασυμπτωτική κατανομή της μεθόδου MXM που δόθηκε στο Θεώρημα της Ενότητας 5.

Πίνακας 1. Εμπειρικοί μέσοι (ΜΕΣΟΣ) και κανονικοποιημένα μέσα τετραγωνικά σφάλματα (NMSE) για δείγματα μεγέθους $n = 50, 100$ ($\beta = 1, r = 40, \tau = 16$)^{*}

(λ, α)	n	Μέτρο	Εκτιμήτριες λ		Εκτιμήτριες α		Εκτιμήτριες β	
			MOM	MXM	MOM	MXM	MOM	MXM
(0.5, 0.5)	50	ΜΕΣΟΣ	0.35	1.58	12.2	0.73	0.49	0.89
			0.41	0.52	2.53	0.64	0.60	0.90
	100	NMSE	17.1	1.4E4	2.0E6	9.19	22.4	16.8
			53.1	2.42	4.1E3	9.23	32.6	16.3
	50	ΜΕΣΟΣ	0.42	0.51	31.1	1.28	0.62	0.95
			0.49	0.51	4.23	1.18	0.72	0.94
(0.5, 1)	100	NMSE	3.69	0.82	2.0E7	21.0	24.3	14.7
			49.6	0.69	2.6E4	19.2	34.8	12.8
	50	ΜΕΣΟΣ	0.98	0.50	21.9	2.33	0.76	1.05
			0.55	0.50	8.08	2.27	0.85	1.02
	100	NMSE	1.3E4	0.68	1.2E6	67.4	29.4	14.4
			1.6E2	0.67	3.5E4	81.8	50.2	16.1
(1, 0.5)	50	ΜΕΣΟΣ	0.91	1.10	7.90	0.64	0.57	0.93
			0.98	1.02	2.27	0.59	0.66	0.93
	100	NMSE	6.1E2	64.2	2.1E5	5.12	20.5	10.7
			4.6E2	2.75	1.5E3	5.04	29.2	10.6
	50	ΜΕΣΟΣ	1.09	1.03	14.5	1.15	0.70	1.01
			1.44	1.01	12.0	1.13	0.74	0.96
(1, 1)	100	NMSE	2.9E2	2.18	1.7E6	13.3	23.0	10.0
			5.0E3	1.86	4.6E6	13.3	33.4	9.33

(1, 2)	50	ΜΕΣΟΣ	1.20	1.01	19.3	2.21	0.87	1.11
	100		1.24	1.01	24.8	2.15	0.95	1.04
	50	NMSE	4.2E2	1.90	5.0E5	2.6E2.	31.6	12.7
	100		7.8E2	1.90	1.7E7	66.1	56.5	13.7
(2, .5)	50	ΜΕΣΟΣ	6.58	4.65	6.50	0.59	0.61	0.96
	100		2.49	2.06	3.22	0.56	0.72	0.96
	50	NMSE	8.1E5	2.8 ^E 5	3.9E4	4.14	18.3	7.36
	100		6.7E3	12.7	2.8E4	3.92	26.0	7.55
(2, 1)	50	ΜΕΣΟΣ	21.0	2.11	8.47	1.09	0.81	1.05
	100		2.54	2.05	8.84	1.07	0.80	1.00
	50	NMSE	1.6E7	10.3	5.4E4	18.0	22.4	8.00
	100		3.3E3	8.32	1.1E6	11.8	33.5	8.36
(2, 2)	50	ΜΕΣΟΣ	3.92	2.07	11.0	2.20	1.07	1.16
	100		2.67	2.05	12.7	2.07	1.03	1.07
	50	NMSE	5.5E4	9.34	1.1E5	1.6E2	37.2	14.2
	100		3.5E2	7.92	4.0E5	70.5	60.0	13.6

* xEy συμβολίζει $x \times 10^y$.

Πίνακας 2. Εμπειρικοί μέσοι (ΜΕΣΟΣ) και NMSE για δείγματα μεγέθους $n = 1000$ και ασυμπτωτικές διασπορές (AVAR) με τη μέθοδο MXM ($\beta = 1$, $r = 40$, $\tau = 16$)

(λ, α)	Μέτρο	$\hat{\lambda}$	$\hat{\alpha}$	$\hat{\beta}$
(.5, .5)	ΜΕΣΟΣ	0.50	0.52	0.98
	NMSE	0.84	6.64	22.2
	AVAR	0.86	6.73	23.2
(.5, 1)	ΜΕΣΟΣ	0.50	1.02	0.99
	NMSE	0.64	16.5	17.4
	AVAR	0.65	15.6	17.2
(.5, 2)	ΜΕΣΟΣ	0.50	2.04	1.00
	NMSE	0.63	73.0	18.0
	AVAR	0.65	69.7	19.4
(1, .5)	ΜΕΣΟΣ	1.00	0.51	0.99
	NMSE	2.31	4.29	13.4
	AVAR	2.37	4.36	14.1
(1, 1)	ΜΕΣΟΣ	1.00	1.02	1.00
	NMSE	1.67	12.1	12.0
	AVAR	1.75	11.8	12.2
(1, 2)	ΜΕΣΟΣ	1.00	2.05	0.99

	NMSE	1.63	72.4	16.5
	AVAR	1.72	62.1	16.4
(2, .5)	MEΣΟΣ	2.00	0.51	0.99
	NMSE	9.07	3.66	8.97
	AVAR	9.17	3.42	9.62
(2, 1)	MEΣΟΣ	2.00	1.02	0.99
	NMSE	6.50	12.3	9.87
	AVAR	6.61	11.3	9.99
(2, 2)	MEΣΟΣ	2.00	2.04	1.00
	NMSE	6.30	77.3	15.7
	AVAR	6.40	69.8	16.2

ABSTRACT

The compound Poisson-gamma (*PGAM*) distribution has been applied in various scientific fields, such as stochastic hydrology, meteorology, risk theory and seismology. This paper presents a method of mixed moments for the estimation of the three parameters of *PGAM*. The method is based on the empirical cumulant generating function and the first two sample moments. The asymptotic distribution of the resulting estimators is found and the proposed method is compared with the method of moments.

ΑΝΑΦΟΡΕΣ

- Alexaderson, H. (1986). A homogeneity test applied to precipitation data, *J. Climatology* **6**: 661-675.
- Arfken, G. B. & Weber, H. J. (2000). *Mathematical Methods for Physicists* (5th ed.), Boston, MA: Academic Press,
- Fisher, R.A & Cornish, E.A. (1960). The percentile points of distributions having known cumulants, *Technometrics* **2**: 209–225.
- Koutrouvelis, I.A. & Meintanis, S. (2002). Estimating the parameters of Poisson-exponential models, *Aust. N. Z. J. Stat.* **44**: 233-245.
- Revfeim, K.J.A. (1984). An initial model of the relationship between rainfall events and daily amounts, *J. Hydrol.* **75**: 357-364.
- Siegel, (1985). Modeling data containing exact zeros using zero degrees of freedom, *J. R. Stat. Soc. Ser. B Stat. Methodol.* **47**: 267-271.
- Withers, S.A. & Nadarajah, S. (2011). On the compound Poisson-gamma distribution, *Kybernetika*, **47**: 15-37.



ΑΝΑΛΥΣΗ ΟΔΟΝΤΙΑΤΡΙΚΩΝ ΔΕΔΟΜΕΝΩΝ ΕΠΙΔΗΜΙΟΛΟΓΙΚΗΣ ΕΡΕΥΝΑΣ ΣΕ ΣΧΟΛΕΙΑ ΤΟΥ ΔΗΜΟΥ ΠΕΙΡΑΙΑ

*Γεώργιος Μπάρτζης¹, Ιωάννης Μπερτόλης¹, Βαρβάρα Λαγού¹, Ειρήνη
Σταματοπούλου¹, Αγγελική Σεληνίδου², Ανδρεάνα Τρικαλιώτου²*

¹Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης, Πανεπιστήμιο Πειραιώς
mpartzis_g@hotmail.com, {bertolis.j, varvara.lagou, estamatop}@gmail.com

²Οδοντίατροι Τμήματος Πρόληψης και Προαγωγής Υγείας
{angelselgr, nadiatrik}@yahoo.gr

ΠΕΡΙΛΗΨΗ

Στην εργασία αυτή αναλύονται οδοντιατρικά δεδομένα που προέκυψαν από την πρωτότυπη και εκτενή επιδημιολογική έρευνα που διενεργήθηκε από την Διεύθυνση Δημόσιας Υγείας και Κοινωνικών Υπηρεσιών Δήμου Πειραιά σε συνεργασία με το Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς στα πλαίσια του προγράμματος πρόληψης και προαγωγής της στοματικής υγείας σε παιδικούς σταθμούς, νηπιαγωγεία και δημοτικά σχολεία του δήμου Πειραιά. Ο στόχος του προγράμματος ήταν η προσφορά υπηρεσιών πρωτογενούς και δευτερογενούς οδοντιατρικής πρόληψης και η προαγωγή της συνεργασίας μεταξύ διαφόρων φορέων και επιστημόνων. Η εργασία κλείνει με την παρουσίαση κάποιων βασικών ευρημάτων μέσω της χρήσης βασικών στατιστικών μεθόδων. Οι έλεγχοι αλλά και οι μέθοδοι που χρησιμοποιήθηκαν στην ανάλυση βασίστηκαν κατά κύριο λόγο σε μη παραμετρική στατιστική (έλεγχοι Mann-Whitney, Kruskal-Wallis, λογιστική παλινδρόμηση ως μοντέλο πρόβλεψης, χ^2 τεστ ανεξαρτησίας) λόγω της φύσης των δεδομένων. Τα βασικότερα αποτελέσματα που προέκυψαν, σχέτισαν τα δημογραφικά χαρακτηριστικά των εξετασθέντων μαθητών με τα οδοντιατρικά τους προβλήματα όπως είναι οι δείκτες τερηδόνας, η προσβολή από αυτή και γενικότερα με το επίπεδο στοματικής υγιεινής.

Λέξεις Κλειδιά: Στοματική Υγεία, Αποτελεσματικότητα, Συνεργασία, Δημόσιοι Φορείς, Κοινωνική Προσφορά, Εκπαίδευση, Έρευνα, Μη Παραμετρική Στατιστική

1. ΕΙΣΑΓΩΓΗ

Η Διεύθυνση Δημόσιας Υγείας και Κοινωνικών Υπηρεσιών του Δήμου Πειραιά διεξήγαγε για πρώτη φορά το σχολικό έτος 2011-2012 πρόγραμμα πρόληψης και προαγωγής της στοματικής υγείας σχεδόν ολόκληρου του σχολικού πληθυσμού του Δήμου Πειραιά. Τα δεδομένα που προέκυψαν επεξεργάστηκαν και αναλύθηκαν από το Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιά, το οποίο για πρώτη φορά ασχολήθηκε επίσημα με τον τομέα της έρευνας και ήρθε σε επαφή με πραγματικά δεδομένα. Η ευκαιρία για την διεξαγωγή της εν λόγω έρευνας και της παραγωγικής συνεργασίας μεταξύ δημόσιων φορέων ανέκυψε κατόπιν αιτήσεως αρωγής των οδοντιάτρων της εν λόγω Διεύθυνσης (κ. Σεληνίδου Α., κ. Τρικαλιώτου Α.) προς τον Πρόεδρο, κ. Τσίμπο και τους καθηγητές του Τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιά για την επεξεργασία και ανάλυση του τεράστιου όγκου συλλεχθέντων δεδομένων παράλληλα με την νύξη των φοιτητών του Τμήματος για την εκπόνηση μελέτης με θέμα κοινωνικού και οικονομικού ενδιαφέροντος με πραγματικά δεδομένα. Με την ομόφωνη συμφωνία των Καθηγητών του Τμήματος συστάθηκε Τριμελής Επιτροπή (κ. Τσίμπος Κλέων, κ. Τζαβελάς Γεώργιος, κ. Μπερσίμης Σωτήριος) η οποία με την συνδρομή 25 εθελοντών φοιτητών διεξήγαγε την έρευνα των συλλεχθέντων δεδομένων.

Η έρευνα αφορούσε στα αποτελέσματα της υλοποίησης του προγράμματος πρόληψης και προαγωγής της στοματικής υγείας σε παιδικούς σταθμούς, νηπιαγωγεία και δημοτικά σχολεία. Σύμφωνα με το πρόγραμμα πραγματοποιήθηκαν εικοσάλεπτες ενημερώσεις ανά τάξη αναφορικά με την σημασία και τη φροντίδα των δοντιών και την εξοικείωση με τον οδοντίατρο. Κατόπιν ακολουθούσε οδοντιατρική εξέταση στους μαθητές με τη συγκατάθεση των γονέων τους και τέλος εξατομικευμένες συμβουλές. Η ενημέρωση των γονέων για την κατάσταση της στοματικής υγιεινής των τέκνων τους έγινε με αποστολή επιστολών και με τη χορήγηση πιστοποιητικού οδοντοστοματολογικής εξέτασης, απαραίτητου για την εγγραφή των μαθητών στο νηπιαγωγείο ή το δημοτικό.

Το γεγονός της διεξαγωγής τέτοιας έρευνας για πρώτη φορά στον Δήμο Πειραιά και το ότι διενεργήθηκε μεταξύ σχεδόν ολόκληρου του σχολικού πληθυσμού σε μεγάλη έκταση αναφορικά με τα στοιχεία προς καταγραφή, την καθιστούν πρωτότυπη. Σημειωτέον ότι για πρώτη φορά αναλαμβάνεται έρευνα τέτοιου τύπου από το Τμήμα Στατιστικής και Ασφαλιστικής επιστήμης του Πανεπιστημίου Πειραιά.

Πολλαπλά οφέλη αποκόμισαν από αυτό και οι εθελοντές φοιτητές, αφού με την ενασχόληση στον τομέα της έρευνας ήρθαν σε επαφή με μεγάλο όγκο πραγματικών δεδομένων (3.800 εγγραφές και 180 μεταβλητές), διαχειρίστηκαν ελλείπουσες τιμές και έλλειψη κανονικότητας. Έτσι οδηγήθηκαν από την θεωρία στην πράξη, ανέπτυξαν κριτική σκέψη και συνεργατικότητα. Τα πρακτικά οφέλη των φοιτητών συνίστανται στην πρακτική άσκησή τους και στο πρωταρχικό κομμάτι κάθε έρευνας, δηλαδή το στάδιο της καταχώρησης δεδομένων, το οποίο δεν μπορεί να γνωρίσει σε βάθος ο φοιτητής αποκλειστικά από την θεωρία. Πρακτικά μαθαίνει την

μεταφορά της πληροφορίας σε διαχειρίσιμη μορφή μέσω της κωδικοποίησης, κάτι που διευκολύνει την καταχώριση, την προστασία των προσωπικών δεδομένων και εξυπηρετεί στον υπολογισμό των οδοντιατρικών δεικτών. Βέβαια αυτό αναπτύσσει την ευθύνη της ορθής αντιγραφής από τα ερωτηματολόγια όσον αφορά στην αποφυγή λαθών και τον έλεγχο σφαλμάτων. Αξιοσημείωτο είναι πως και οι φοιτητές αξιολόγησαν πολύ θετικά την έρευνα όσον αφορά στη διαχείριση δεδομένων με H/Y, την επίλυση πραγματικών προβλημάτων με χρήση στατιστικής, την αφομοίωση των προπτυχιακών μαθημάτων, την κατανόηση του στατιστικού τρόπου σκέψης και την διαδικασία των υπολογισμών.

Η εν λόγω έρευνα κρίνεται παραγωγική λόγω της συνεργασίας που αναπτύχθηκε μεταξύ δημόσιων φορέων, τη μεταφορά τεχνογνωσίας, την ανάδειξη της ανάγκης επιμόρφωσης από όλες τις πλευρές, την εμβάθυνση σε πραγματικά προβλήματα και τη διαμόρφωση προτάσεων σχετικών με την υγεία.

2. ΑΝΑΛΥΣΗ

Οι διαθέσιμες μεταβλητές που καταγράφηκαν για την παρούσα έρευνα ήταν: i) η ημ/νία εξέτασης, ii) ο οδοντίατρος εξεταστής, iii) το δημοτικό διαμέρισμα που ανήκει το σχολείο, iv) το ίδιο το σχολείο και v) η τάξη που ανήκει ο κάθε μαθητής.

Η ημερομηνία εξέτασης κωδικοποιήθηκε ως (1) για το σχολικό έτος 2011-2012 καθώς ήταν η πρώτη χρονιά συνεργασίας και ανάλυσης και ενδεχομένως θα ακολουθήσει συγκριτική ανάλυση νέων σχολικών ετών σε σχέση με την υπάρχουσα χρονιά. Επίσης το σχολείο και η τάξη κωδικοποιήθηκαν για να γίνουν μελλοντικές συγκρίσεις.

Όσον αφορά στα δημογραφικά στοιχεία, καταγράφηκαν: i) το φύλο, ii) η ηλικία και iii) η εθνικότητα του παιδιού, iv) το εκπαιδευτικό επίπεδο των γονέων και v) το επάγγελμα των γονέων του. Οι διαθέσιμες μεταβλητές σχετικά με τα στοιχεία κατάστασης δοντιών που καταγράφηκαν ήταν: i) η οδοντική τερηδόνα, ii) οι ανάγκες θεραπείας τερηδόνας και iii) οι ανάγκες ορθοδοντικής θεραπείας.

Από τα στοιχεία που καταγράφηκαν στο πεδίο της οδοντικής τερηδόνας, οδηγηθήκαμε στον υπολογισμό του δείκτη DMFT για τα μόνιμα αλλά και του dmft για τα νεογιά δόντια. Ο δείκτης DMFT/dmft αποτελεί έναν αθροιστικό δείκτη οδοντικών προβλημάτων καθώς αθροίζει το αριθμό των τερηδονισμένων (Decayed), ελλειπόντων (Missing) και σφραγισμένων (Filled) δοντιών. Στο πεδίο των αναγκών ορθοδοντικής θεραπείας, καταγράφηκαν οι περιπτώσεις ύπαρξης, μη ύπαρξης αλλά και οι υπό θεραπεία περιπτώσεις. Πρέπει να αναφερθεί πως οι δείκτες που γράφονται με κεφαλαία γράμματα αντιστοιχούν στα μόνιμα δόντια ενώ σε αντίθετη περίπτωση αναφέρονται σε νεογιά δόντια.

Τέλος οι διαθέσιμες μεταβλητές σχετικά με τα στοιχεία κατάστασης στόματος ήταν i) ο φραγμός, ii) οι μαλακές εναποθέσεις DI-S, iii) οι χρωστικές, iv) η τρυγία, v) πιθανό οδοντικό τραύμα, vi) οδοντικό κάταγμα και vii) το απόστημα ή/και συρίγγιο.

Οι μαλακές εναποθέσεις καταγράφηκαν για να μας βοηθήσουν στον υπολογισμό του απλουστευμένου δείκτη στοματικής υγιεινής (DI-S). Για τον υπολογισμό του δείκτη χρησιμοποιούνται 6 δόντια τα οποία βαθμολογούνται με κλίμακα 0-3 με κριτήριο το ποσοστό των οργανικών εναποθέσεων στην επιφάνεια του δοντιού. Τελικώς, ο δείκτης DI-S του ατόμου προσδιορίζεται προσθέτοντας τις ενδείξεις των εξετασθέντων δοντιών και διαιρώντας με το πλήθος τους.

3. ΜΕΘΟΔΟΛΟΓΙΑ

Η ανάλυση της παρούσας έρευνας έγινε με το στατιστικό πακέτο SPSS 20. Όλα τα συμπεράσματα βασίστηκαν σε επίπεδο στατιστικής σημαντικότητας 5%. Για τον έλεγχο ανεξαρτησίας μεταξύ ποιοτικών μεταβλητών χρησιμοποιήθηκε ο έλεγχος χ^2 του Pearson. Για τη διαφορά μεταξύ δύο πληθυσμιακών ομάδων χρησιμοποιήθηκε το μη παραμετρικό τεστ των Mann and Whitney ενώ για διαφορά μεταξύ περισσοτέρων ομάδων χρησιμοποιήθηκε το μη παραμετρικό τεστ των Kruskal and Wallis. Ως μοντέλα πρόβλεψης χρησιμοποιήθηκαν η λογιστική αλλά και η γραμμική παλινδρόμηση. Για να διαχωριστούν πληθυσμιακές ομάδες χρησιμοποιήθηκε η ανάλυση κατά συστάδες και τέλος χρησιμοποιήθηκε ο έλεγχος t όταν είχαμε κανονικούς πληθυσμούς.

4. ΑΠΟΤΕΛΕΣΜΑΤΑ

Σαν πρώτο αντικείμενο μελέτης της συγκεκριμένης έρευνας ήθελε να μελετηθεί η σχέση μεταξύ της εθνικότητας και του επιπέδου της στοματικής υγιεινής στα παιδιά δημοτικών σχολείων. Έτσι μέσω του ελέγχου χ^2 του Pearson βρέθηκε στατιστικώς σημαντική σχέση μεταξύ των δύο αυτών ποιοτικών μεταβλητών καθώς η p-value ήταν 0.02. Παρατηρήθηκε πως το 44.2% των Ελλήνων μαθητών είχε καλή στοματική υγιεινή, το 50.4% είχε μέτρια στοματική υγιεινή ενώ το υπόλοιπο 5.4% είχε κακή στοματική υγιεινή. Αντίθετα, το 34.6% των αλλοδαπών μαθητών είχε καλή στοματική υγιεινή δηλαδή αρκετά μικρότερο ποσοστό σε σχέση με τους Έλληνες μαθητές. Το 58.4% είχε μέτρια στοματική υγιεινή και τέλος το υπόλοιπο 7% είχε κακή στοματική υγιεινή. Τα αποτελέσματα παρουσιάζονται στον Πίνακα 1:

Πίνακας 1. Στοματική υγιεινή έναντι εθνικότητας σε 1768 Έλληνες και 399 αλλοδαπούς μαθητές

Στοματική Υγιεινή	Έλληνες	Αλλοδαποί	Σύνολο
Καλή	44.2%	34.6%	42.4%
Μέτρια	50.4%	58.4%	51.9%
Κακή	5.4%	7.0%	5.7%

Το επόμενο που μελετήθηκε είναι η σχέση μεταξύ εθνικότητας και προσβολής από τερηδόνα σε μαθητές νηπιαγωγείων και παιδικών σταθμών. Παρατηρήθηκε πως οι αλλοδαποί μαθητές έχουν διπλάσιους και τριπλάσιους κάποιες

φορές δείκτες dmft (νεογιλά δόντια) σε σχέση με τους Έλληνες μαθητές. Φάνηκε πως ο μέσος Έλληνας μαθητής έχει σταθερά λιγότερα από ένα προβλήματα στο στόμα του ενώ ένας αλλοδαπός έχει περισσότερα από ένα προβλήματα που μερικές φορές είναι δύο και τρία. Τα αποτελέσματα συνοψίζονται στον Πίνακα 2:

Πίνακας 2. Μέσος αριθμός προσβεβλημένων δοντιών από τερηδόνα σε 2680 Έλληνες και 579 αλλοδαπούς μαθητές

dmft	Έλληνες	Αλλοδαποί
Νηπιαγωγεία	0.72	2.4
Α' Τάξη	0.88	3.06
Δ' Τάξη	0.94	1.96
Στ' Τάξη	0.35	0.6

Ο έλεγχος χ^2 του Pearson έδειξε στατιστικώς σημαντική σχέση μεταξύ της εθνικότητας και της προσβολής από τερηδόνα ($p\text{-value}<0.01$).

Ένα ερώτημα που είχε ενδιαφέρον να μελετηθεί ήταν το αν υπάρχει στατιστικώς σημαντική διαφορά στους δείκτες στοματικής υγιεινής και τερηδογονικής προσβολής ανάμεσα στα σχολεία των πέντε Δημοτικών διαμερισμάτων του Δήμου Πειραιά ή μεταξύ των νηπιαγωγείων Πειραιά και Περάματος. Αυτό είχε ενδιαφέρον καθώς οι οδοντίατροι που εξέταζαν του μαθητές έβλεπαν τη διαφορά και έμενε να το διαπιστώσουν και στατιστικά. Κάτι τέτοιο δεν επιβεβαιώθηκε και δεν φαίνεται τα Δημοτικά διαμερίσματα να διαφέρουν.

Μεταξύ των τάξεων του Δημοτικού (Α', Δ', Στ') και του αριθμού τραυμάτων και καταγμάτων των δοντιών βρέθηκε στατιστικώς σημαντική γραμμική εξάρτηση (linear-by-linear association) ($p\text{-value}<0,01$) συνεπώς υπάρχει όπως είναι αναμενόμενο αύξηση του ποσοστού των οδοντικών τραυμάτων και καταγμάτων όσο αυξάνεται η ηλικία.

Το επόμενο βήμα της έρευνας ήταν η δημιουργία ενός νέου συνόλου δεδομένων όπου για κάθε σχολείο που εξετάστηκε αντιστοιχήθηκε ένας μέσος δείκτης προβλημάτων των DMFT (μόνιμα δόντια) και dmft (νεογιλά δόντια) καθώς επίσης ο αριθμός των αγοριών και κοριτσιών σε κάθε σχολείο, ο αριθμός των Ελλήνων και Αλλοδαπών στα σχολεία, ένας μέσος απλουστευμένος δείκτης στοματικής υγιεινής (DI-S), οι μέσοι δείκτες τερηδονισμένων (DT/dt), απολεσθέντων (MT/mt) και εμφραχθέντων δοντιών (FT/ft). Τέλος αντιστοιχήθηκε και τα μέσα οδοντικά τραύματα και κατάγματα των παιδιών στο εκάστοτε σχολείο.

Με βάση το νέο σύνολο δεδομένων έγινε ανάλυση κατά συστάδες σε: i) δύο ομάδες με βάση το μέσο Δείκτη DMFT, ii) δύο ομάδες σχολείων σύμφωνα με τα ποσοστά ανά φύλο και iii) δύο ομάδες σύμφωνα με το ποσοστό αλλοδαπών που υπάρχει σε κάθε σχολείο.

Για τις δύο ομάδες που δημιουργήθηκαν με βάση το ποσοστό αλλοδαπών μελετήθηκε αν διαφέρουν στατιστικώς σημαντικά τα σχολεία με μεγάλο ποσοστό αλλοδαπών από τα σχολεία με χαμηλό ποσοστό αλλοδαπών ως προς το μέσο DMFT. Φάνηκε να υπάρχει σημαντική διαφοροποίηση του μέσου DMFT σύμφωνα με τη συγκεκριμένη συσταδοποίηση ($p\text{-value}<0.001$) και έτσι σχολεία με μικρό ποσοστό αλλοδαπών φαίνεται ότι έχουν μικρότερα οδοντιατρικά προβλήματα.

Ενδιαφέρον θα είχε με τα δεδομένα που έχουμε στα χέρια μας να μπορούσε να προβλέψει κάποιος και να κατηγοριοποιήσει κάποιο σχολείο σε προβληματικό ή όχι ακόμα και αν δεν είναι οδοντίατρος σύμφωνα με κάποια χαρακτηριστικά. Για το λόγο αυτόν, εφαρμόστηκε μοντέλο λογιστικής παλινδρόμησης με ανεξάρτητες μεταβλητές τον αριθμό των μαθητών, το ποσοστό των αλλοδαπών μαθητών, το δημοτικό διαμέρισμά που ανήκει το σχολείο και την κατηγορία του σχολείου ως προς το ποσοστό αγοριών μαθητών (σύμφωνα με την προηγούμενη ανάλυση συστάδων). Έτσι το μοντέλο που εφαρμόστηκε σύμφωνα με τον έλεγχο Hosmer & Lemeshow έδειξε καλή προσαρμογή καθώς το $p\text{-value}$ ήταν 0.467 και επίσης είχε ποσοστό ορθής ταξινόμησης 82.5%.

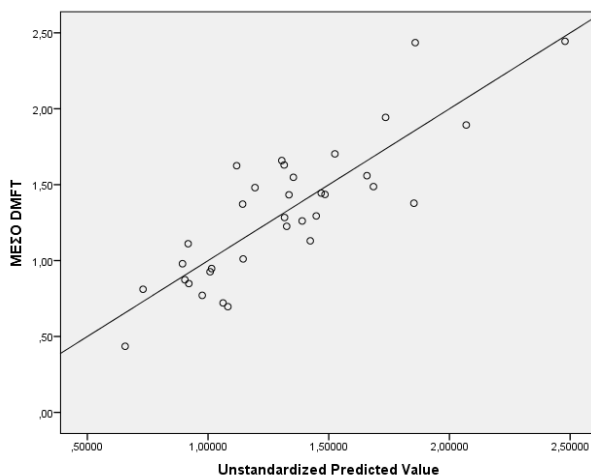
Το επόμενο που θα είχε ενδιαφέρον είναι να μπορούσε κάποιος να προβλέψει το μέσο DMFT για κάποιο σχολείο ακόμα και αν δεν είναι οδοντίατρος. Έτσι, εφαρμόστηκε μοντέλο γραμμικής παλινδρόμησης με ανεξάρτητες μεταβλητές το ποσοστό των ελλήνων μαθητών, τον αριθμό των μαθητών σε κάθε σχολείο και το μέσο DI-S. Στο μοντέλο που εξετάστηκε ισχύουν όλες οι προϋποθέσεις του γραμμικού μοντέλου και ο συντελεστής καλής προσαρμογής R^2 ήταν 72%. Το εκτιμηθέν μοντέλο έχει τη μορφή:

$$\text{μέσο\_DMFT} = 2.176 - 0.982 * \text{Ποσοστό\_Ελλήνων} + 0.805 * \text{Μέσο\_DI-S} - 0.009 * \text{Αριθμός μαθητών}.$$

(0.314) (0.353)
(0.159)
(0.002)

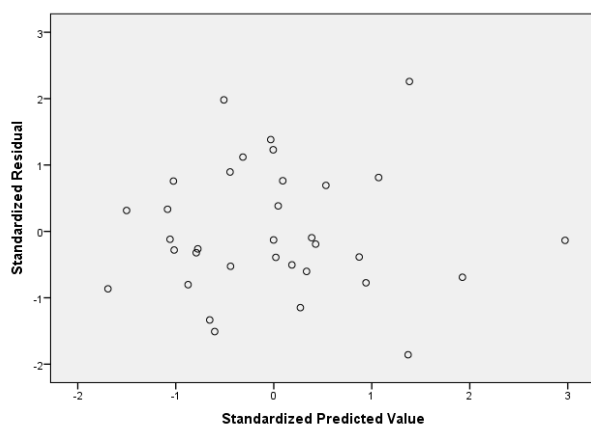
Το μοντέλο είναι απλό καθώς για να υπολογιστεί το μέσο DI-S χρειάζεται να εξεταστούν 6 δόντια απλώς στην επιφάνειά τους και τη δουλειά αυτή μπορεί να την κάνει και κάποιος που δεν είναι οδοντίατρος και να αποφανθεί αν οι οργανικές εναποθέσεις στην επιφάνεια των δοντιών καλύπτουν το 1/3, τα 2/3 ή περισσότερο από 2/3. Στην παρακάτω εικόνα παρουσιάζεται το διάγραμμα των εκτιμώμενων τιμών του μοντέλου έναντι των παρατηρηθεισών τιμών του μέσου DMFT.

Εικόνα 1. Εκτιμώμενες έναντι παρατηρηθεισών τιμών



Θα πρέπει να αναφερθεί πως ελέγχθηκε και η υπόθεση της ομοσκεδαστικότητας διαγραμματικά με τη δημιουργία διαγράμματος διασποράς των τυποποιημένων καταλοίπων έναντι των τυποποιημένων εκτιμώμενων τιμών. Το διάγραμμα φαίνεται στην παρακάτω εικόνα.

Εικόνα 2. Τυποποιημένα κατάλοιπα έναντι τυποποιημένων εκτιμώμενων τιμών



Η ίδια διαδικασία έγινε με εξαρτημένη μεταβλητή το μέσο DT με ανεξάρτητες μεταβλητές αυτή τη φορά το ποσοστό αλλοδαπών μαθητών, τον αριθμό των μαθητών και το μέσο DI-S. Αυτή τη φορά ο συντελεστής καλής προσαρμογής ήταν 73.6% και το μοντέλο έχει τη μορφή:

μέσο_DT=0.783-1.285*Ποσοστό Αλλοδαπών+0.505*Μέσο DI-S-0.007*Αριθμός μαθητών.

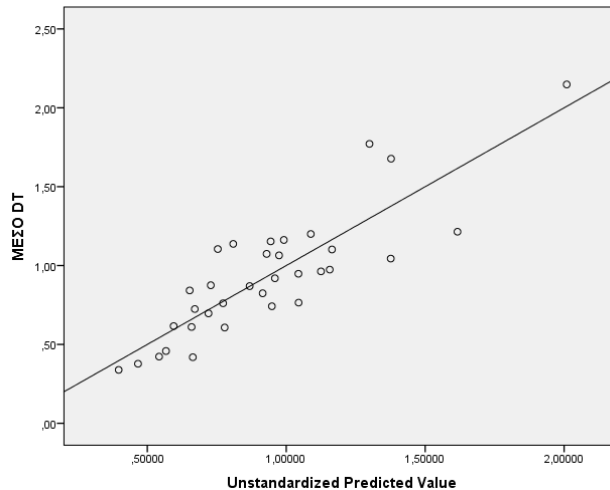
(0,115) (0,292)

(0,131)

(0,002)

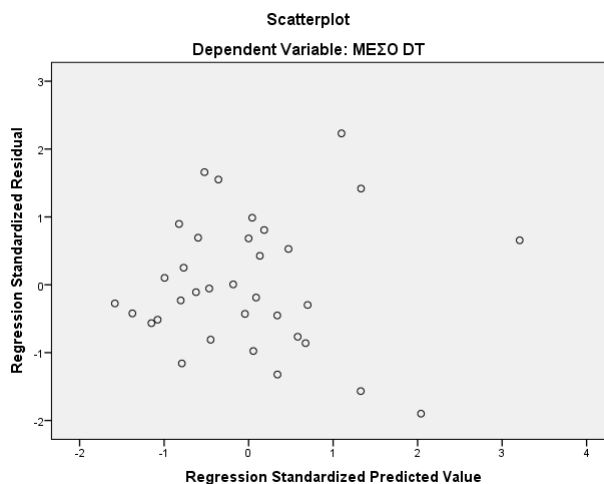
Το διάγραμμα των εκτιμώμενων τιμών του μοντέλου έναντι των παρατηρηθείσων τιμών του μέσου DT φαίνεται στην εικόνα 2.

Εικόνα 3. Εκτιμώμενες έναντι παρατηρηθείσων τιμών



Και σε αυτήν την περίπτωση ελέγχθηκε η υπόθεση της ομοσκεδαστικότητας διαγραμματικά με τη δημιουργία διαγράμματος διασποράς των τυποποιημένων καταλοίπων έναντι των τυποποιημένων εκτιμώμενων τιμών. Στην παρακάτω εικόνα φαίνεται το διάγραμμα.

Εικόνα 4. Τυποποιημένα κατάλοιπα έναντι τυποποιημένων εκτιμώμενων τιμών



Τέλος, εφαρμόσθηκε η μέθοδος ανάλυσης κατά συστάδες με τη μέθοδο K-Means (3 ομάδες) με σκοπό την κατηγοριοποίηση των σχολείων σε τρεις ομάδες με βάση το DMFT και το DI-S. Οι ομάδες που προέκυψαν είναι κλιμακωτά από λιγότερα στοματολογικά προβλήματα μέχρι περισσότερα. Στον Πίνακα 3 φαίνεται η κατηγοριοποίηση:

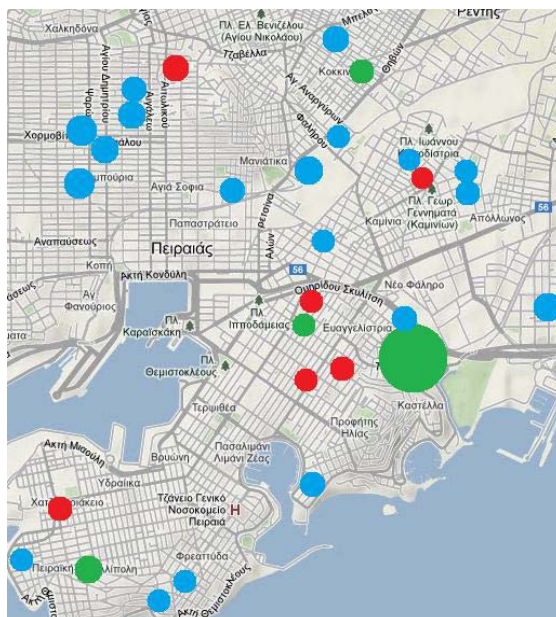
Πίνακας 3. Συστάδες

	Ομάδα 1	Ομάδα 2	Ομάδα 3
Κέντρο ομάδας(DMFT)	0,9961	1,2502	2,0128
Κέντρο ομάδας (DI-S)	0,1355	0,84	0,895
Αριθμός σχολείων	9	19	6

Ουσιαστικά οι ομάδες που δημιουργήθηκαν κατηγοριοποιούν τα σχολεία σε καλά (μικροί δείκτες DI-S και DMFT), μέτρια (μέτριοι δείκτες DI-S και DMFT) και κακά (μεγάλοι δείκτες DI-S και DMFT).

Έπειτα από την ανάλυση αυτή δημιουργήθηκε χάρτης για να φανεί οπτικά αν υπάρχουν γειτονιές με αρκετά προβλήματα, ή με λίγα προβλήματα. Στην Εικόνα 3 φαίνεται ο χάρτης αυτός:

Εικόνα 5. Χάρτης Δήμου Πειραιά με τα σχολεία βάσει των στοματολογικών προβλημάτων



Στον παραπάνω χάρτη χρωματίστηκαν με πράσινο σχολεία που έχουν μικρούς δείκτες DI-S και DMFT, με μπλε σχολεία που έχουν μέτριους δείκτες DI-S και DMFT και με κόκκινο σχολεία με μεγάλους δείκτες DI-S και DMFT.

Ευχαριστίες: Οι συγγραφείς ευχαριστούν όλους τους καθηγητές που βοήθησαν στην έρευνα αλλά και όσους συνέδραμαν στην ολοκλήρωσή της.

ABSTRACT

The Piraeus Directorate of Public Health and Social Services in collaboration with the Department of Statistics and Insurance Science of University of Piraeus conducted an epidemiological survey within the program of prevention and promotion of oral health among the students of Piraeus kindergartens and primary schools. The obtained dental data of this innovative and extensive survey are analyzed in this paper. The objective of this particular program was to provide primary, secondary dental services and promote the cooperation between public services and scientists. Certain results developed by basic statistical methods are presented. The tests and the methods used were based on nonparametric statistics (Mann-Whitney, Kruskal-Wallis, logistic regression, Chi-Square test of independence) because of the data's nature. As a result, demographic characteristics of the examined students are related to dental problems and generally oral hygiene.

Ευχαριστίες: Οι συγγραφείς ευχαριστούν όλους τους καθηγητές που βοήθησαν στην έρευνα αλλά και όσους συνέδραμαν στην ολοκλήρωσή της.

ΑΝΑΦΟΡΕΣ

- Κούτρας, Μ. και Ευαγγελάρας, Χ. (2010). *Ανάλυση Παλινδρόμησης: Θεωρία και Εφαρμογές*, Αθήνα: εκδόσεις Αθ. Σταμούλης.
- Καφφές, Δ. Ι. (1989). *Μαθήματα Αναλύσεως Διακυμάνσεως*, Πειραιάς: εκδόσεις Αθ. Σταμούλης.
- Paggano, M. and Gauvreau K. (2000). *Αρχές Βιοστατιστικής*, Αθήνα: εκδόσεις Έλλην.
- Σιώμοκς, Γ. Ι. και Βασιλικοπούλου, Α. Ι. (2005). *Εφαρμογή Μεθόδων Ανάλυσης στην Έρευνα Αγοράς*, Αθήνα: εκδόσεις Αθ. Σταμούλης.



ΠΡΟΣΕΓΓΙΣΗ ΜΕΣΩ ΤΗΣ ΚΑΤΑΝΟΜΗΣ ΓΑΜΜΑ ΑΝΤΙ ΤΗΣ ΚΑΝΟΝΙΚΗΣ ΚΑΤΑΝΟΜΗΣ ΣΤΟ ΚΕΝΤΡΙΚΟ ΟΡΙΑΚΟ ΘΕΩΡΗΜΑ

Μ. Β. Μπούτσικας

Τμ. Στατιστικής και Ασφ. Επιστ., Πανεπιστήμιο Πειραιώς

mbouts@unipi.gr

ΠΕΡΙΛΗΨΗ

Στην παρούσα εργασία μελετάται η (προ-ασυμπτωτική) προσέγγιση της κατανομής ενός αθροίσματος ανεξάρτητων και ισόνομων τυχαίων μεταβλητών μέσω της κατανομής γάμμα. Συγκεκριμένα, δίνονται ακριβή και ασυμπτωτικά φράγματα για την απόσταση ολικής κύμανσης μεταξύ της κατανομής του αθροίσματος n τυχαίων μεταβλητών με μη μηδενική λοξότητα και απόλυτα συνεχές μέρος, και κατάλληλης μετατοπισμένης γάμμα κατανομής. Τα φράγματα αυτά συγκλίνουν στο μηδέν με τάξη n^{-1} , ενώ αντίστοιχη η τάξη σύγκλισης στην τυπική κανονική κατανομή (σύμφωνα με το ΚΟΘ) είναι μικρότερη, $n^{-1/2}$. Το γεγονός αυτό υποδηλώνει ότι στην περίπτωση λοξών κατανομών, η προ-ασυμπτωτική προσέγγιση μέσω της κατανομής γάμμα είναι πολύ πιο ικανοποιητική από ό,τι η συνήθης κανονική προσέγγιση. Επίσης παρουσιάζονται και δυο παραδείγματα που αφορούν την προσέγγιση της κατανομής αθροισμάτων τυχαίων μεταβλητών που ακολουθούν την λογαριθμοκανονική και την κατανομή Pareto, στα οποία γίνεται φανερή η βελτίωση της προσέγγισης.

Λέξεις κλειδιά: Κεντρικό οριακό θεώρημα, προσέγγιση μέσω κατανομής γάμμα, απόσταση ολικής κύμανσης, τάξη σύγκλισης, άθροισμα λογαριθμοκανονικών τ.μ., άθροισμα τ.μ. Pareto.

1. ΕΙΣΑΓΩΓΗ

Το Κεντρικό Οριακό Θεώρημα (ΚΟΘ) διαδραματίζει πολύ σημαντικό ρόλο στη Θεωρία Πιθανοτήτων και τη Στατιστική προσφέροντας μια πολύ εύχρηστη και αποτελεσματική προσέγγιση για την άγνωστη ή δύσχρηστη κατανομή ενός αθροίσματος τυχαίων μεταβλητών (τ.μ.). Σύμφωνα με την πιο απλή μορφή του ΚΟΘ, αν $Y_1, Y_2, \dots$ είναι ανεξάρτητες και ισόνομες (αν.ισ.) τ.μ. με $E(Y_i) = \mu$, $V(Y_i) = \sigma^2 < \infty$ τότε ο τυποποιημένος μέσος των $Y_1, Y_2, \dots, Y_n$ συγκλίνει κατά κατανομή στην τυπική κανονική κατανομή, $N(0,1)$, για $n \rightarrow \infty$. Ισοδύναμα, για τυποποιημένες αν.ισ. τ.μ. $X_1, X_2, \dots \sim F$, (δηλ. $X_i = (Y_i - \mu)/\sigma$) ισχύει ότι

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n X_i \rightarrow_p N(0,1).$$

Συμβολίζουμε με F_n την συνάρτηση κατανομής (σ.κ.) της τ.μ. $n^{-1/2} \sum_{i=1}^n X_i$. Μεγάλο πρακτικό και θεωρητικό ενδιαφέρον παρουσιάζει η ταχύτητα της παραπάνω σύγκλισης, δηλαδή η ταχύτητα της σύγκλισης (κατά σημείο ή ομοιόμορφα) της ακολουθίας των συναρτήσεων F_n , στην σ.κ. Φ της $N(0,1)$, η οποία προφανώς εξαρτάται από τη μορφή της κατανομής F των X_i . Έχουν δημοσιευτεί αρκετά σχετικά αποτελέσματα:

(Α) Μέσω *φραγμάτων τύπου Berry-Esseen* (B-E) που αφορούν την απόσταση Kolmogorov, ρ , μεταξύ των F_n , Φ και είναι της μορφής

$$\rho(F_n, \Phi) = \sup_x |F_n(x) - \Phi(x)| \leq C \frac{E(|X|^3)}{\sqrt{n}},$$

όπου C είναι μια σταθερά που ανήκει στο διάστημα $(0.4097, 0.4748)$. Το παραπάνω φράγμα συγκλίνει στο 0 με τάξη $n^{-1/2}$, και ισχύει για όλες τις κατανομές F με $E(|X|^3) < \infty$, αλλά δεν είναι ασυμπτωτικά βέλτιστο, δηλαδή το πηλίκο της απόστασης προς το φράγμα δεν συγκλίνει στην μονάδα.

(Β) Μέσω *ασυμπτωτικών αναπτυγμάτων* (π.χ. Edgeworth expansions) της μορφής

$$F_n(x) - \Phi(x) = \frac{E(X^3)(1-x^2)e^{-x^2/2}}{3!\sqrt{2\pi n}} + o(n^{-1/2}).$$

Το παραπάνω προσεγγιστικό ανάπτυγμα της F_n είναι ακριβέστερο ασυμπτωτικά, αλλά δεν ισχύει για όλες τις F (μόνο για τις μη αριθμητικές) και δεν είναι κλειστού τύπου όπως το φράγμα Berry-Esseen. Η ποσότητα $o(n^{-1/2})$ είναι κάποια ακολουθία του n που συγκλίνει στο 0 πιο γρήγορα από ό,τι η ακολουθία $n^{-1/2}$.

Τα παραπάνω αποτελέσματα (Α), (Β) συνήθως εξάγονται μέσω ανάλυσης Fourier, αντιστρέφοντας χαρακτηριστικές συναρτήσεις (βλ. Petron, 1995). Στο παρελθόν έχουν αναπτυχθεί και άλλες τεχνικές μελέτης της τάξης σύγκλισης, π.χ. μέσω ιδιοτήτων μετρικών πιθανότητας (βλ. π.χ. Rachev, 1991), μέσω της μεθόδου του Stein (βλ. π.χ. Barbour and Hall, 1984, Chen and Shao, 2001). Αξίζει να αναφερθεί και η μέθοδος που βασίζεται σε πυρήνες συνδιακύμανσης (covariance kernels) (βλ. π.χ. Cacoullos, Papadatos and Papathanasiou, 1997).

Σε μια πρόσφατη εργασία (Boutsikas, 2011) αποδείχθηκε χρησιμοποιώντας τεχνικές μετρικών πιθανότητας και ανισότητες εξομάλυνσης ότι αν: (i) η F έχει *απόλυτα συνεχές μέρος* (π.χ. είναι συνεχής ή είναι μίξη συνεχούς με κάποια άλλη κατανομή) και (ii) έχει *τις ίδιες k πρώτες ροπές* με την Φ , τότε

$$\rho(F_n, \Phi) = \sup_x |F_n(x) - \Phi(x)| \leq \mathbf{R}_{k,n} \sim \frac{\|\phi^{(k)}\|_\infty |E(X^{k+1}) - \mu_{k+1}|}{(k+1)! n^{(k-1)/2}},$$

όπου το άνω φράγμα $\mathbf{R}_{k,n}$ έχει συγκεκριμένη κλειστή μορφή και υπό συγκεκριμένες συνθήκες είναι ασυμπτωτικά βέλτιστο, δηλαδή $\rho(F_n, \Phi) \sim \mathbf{R}_{k,n} (a_n \sim b_n \Leftrightarrow a_n / b_n \rightarrow 1)$. Υπενθυμίζεται ότι $\|\phi^{(k)}\|_\infty = \sup_x \{\phi^{(k)}(x)\}$ και $\phi^{(k)}$ είναι η παράγωγος k τάξης της σ.π.π. της $N(0,1)$, ενώ με μ_k συμβολίζεται η ροπή k τάξης της $N(0,1)$. Για να γίνει περισσότερο κατανοητό το παραπάνω, στη συνέχεια εφαρμόζεται για $k = 2, 3$.

(i) Αν $k = 2$ (δηλαδή, $E(X_i) = 0, E(X_i^2) = 1$) τότε η $\rho(F_n, \Phi)$ συγκλίνει στο 0 με τάξη $n^{-1/2}$. Για παράδειγμα, αν οι αν.ισ. $Y_1, Y_2, \dots \sim G(a, \lambda)$, (κατανομή γάμμα) θεωρούμε τις τυποποιημένες $X_i = (Y_i - a / \lambda)(a / \lambda^2)^{-1/2}$, $i = 1, 2, \dots$ για τις οποίες ισχύουν οι παραπάνω συνθήκες με $k = 2$ και επομένως

$$\rho(F_n, \Phi) \leq \mathbf{R}_{2,n} \sim \frac{\|\phi^{(2)}\|_\infty |E(X^3)|}{3! \sqrt{n}} = \frac{E(X^3)}{3! \sqrt{2\pi n}} = \frac{2/\sqrt{a}}{3! \sqrt{2\pi n}} \approx \frac{0.13298}{\sqrt{a n}},$$

όπου F_n είναι η κατανομή της τ.μ. $\sqrt{n}\bar{X}_n$. Συγκριτικά, το φράγμα B-E δίνει

$$\rho(F_n, \Phi) \leq C \frac{E(|X|^3)}{\sqrt{n}} = C \frac{2(6-e)/e}{\sqrt{a n}} \approx \frac{1.14643}{\sqrt{a n}},$$

το οποίο έχει την ίδια τάξη σύγκλισης αλλά είναι αρκετά μεγαλύτερο.

(ii) Αν $k = 3$ (δηλαδή $E(X_i) = 0, E(X_i^2) = 1, E(X_i^3) = 0$) τότε είναι ενδιαφέρον ότι η $\rho(F_n, \Phi)$ συγκλίνει στο 0 με τάξη n^{-1} . Για παράδειγμα, αν $Y_i \sim U(0,1)$ (Ομοιόμορφη) και $X_i = (Y_i - E(Y_i))V(Y_i)^{-1/2}$ τότε $k = 3$ και

$$\rho(F_n, \Phi) \leq \mathbf{R}_{3,n} \sim \frac{\|\phi^{(3)}\|_\infty |E(X^4) - \mu_4|}{4! n} = \frac{e^{\frac{\sqrt{6}-3}{2}} \sqrt{9-3\sqrt{6}}}{4! \sqrt{\pi}} \frac{3 - \frac{9}{5}}{n} \approx \frac{0.027529}{n}.$$

Συγκριτικά, το φράγμα B-E δίνει

$$\rho(F_n, \Phi) \leq C \frac{E(|X|^3)}{\sqrt{n}} = C \frac{3\sqrt{3}}{4\sqrt{n}} \approx \frac{0.6167}{\sqrt{n}},$$

το οποίο όχι μόνο δεν είναι βέλτιστο, αλλά δεν έχει και την σωστή τάξη σύγκλισης.

Παρόμοια αποτελέσματα μπορούν να θεμελιωθούν χρησιμοποιώντας την απόσταση ολικής κύμανσης μεταξύ των κατανομών δύο οποιωνδήποτε τ.μ. X, W :

$$\mathbf{d}(F_X, F_W) = \sup_A |P(X \in A) - P(W \in A)| = \frac{1}{2} \int_{-\infty}^{\infty} |f_X(x) - f_W(x)| dx.$$

Η δεύτερη ισότητα ισχύει όταν υπάρχουν οι σ.π.π. f_X, f_W των X, W . Για την απόσταση $\mathbf{d}$ ισχύει η εξής χρήσιμη ανισότητα

$$\mathbf{d}(F_{g(X)}, F_{g(W)}) \leq \mathbf{d}(F_X, F_W),$$

για κάθε μετρήσιμη συνάρτηση g . Επίσης, $\rho(F_X, F_W) \leq \mathbf{d}(F_X, F_W)$. Οπότε, αν έχουμε ένα φράγμα για την απόσταση $\mathbf{d}$ μεταξύ των κατανομών των X, W τότε το ίδιο φράγμα ισχύει και για την απόσταση $\mathbf{d}$ ή ρ μεταξύ των κατανομών τ.μ. της μορφής $g(X), g(W)$. Επίσης, στη συνέχεια θα χρησιμοποιήσουμε και την απόσταση

$$\zeta_k(F_X, F_W) = \int_{-\infty}^{\infty} |F_X^{[k-1]}(x) - F_W^{[k-1]}(x)| dx = \frac{E(X^k) - E(W^k)}{k!},$$

η οποία συνήθως καλείται ιδεώδης μετρική του Zolotarev τάξης k . Με $F^{[k]}$ συμβολίζεται η παράγουσα k -τάξης της F . Η τελευταία ισότητα $=^*$ ισχύει όταν $X \geq_{k-\text{cx}} W$ (βλ. Boutsikas and Vaggelatos, 2002), δηλαδή όταν η κατανομή της X είναι μεγαλύτερη από αυτήν της W κατά την k κυρτή διάταξη (δηλ. $E(\varphi(X)) \geq E(\varphi(W))$ για όλες τις k -κυρτές συναρτήσεις φ). Για περισσότερες λεπτομέρειες στα παραπάνω παραπέμπουμε στην εργασία Boutsikas (2011).

2. ΠΡΟΣΕΓΓΙΣΗ ΜΕΣΩ ΤΗΣ ΚΑΤΑΝΟΜΗΣ ΓΑΜΜΑ

Μια γόνιμη παρατήρηση από τα παραπάνω είναι ότι όταν οι αν.ισ. τ.μ. $X_i \sim F$ και οι αν.ισ. τ.μ. $Z_i \sim N(0,1)$, και οι X_i, Z_i έχουν περισσότερες από τις δύο πρώτες ροπές ίσες τότε η σύγκλιση κατά κατανομή

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n X_i \rightarrow_d \frac{1}{\sqrt{n}} \sum_{i=1}^n Z_i \sim N(0,1),$$

γίνεται με μεγαλύτερη ταχύτητα. Π.χ. αν $\gamma = E(X^3) = 0$ ($k = 3$) τότε η τάξη είναι n^{-1} αντί $n^{-1/2}$. Εάν όμως $\gamma \neq 0$ τότε η τάξη της σύγκλισης θα είναι αναγκαστικά $n^{-1/2}$, διότι δεν μπορούμε να αλλάξουμε τον μηδενικό συντελεστή λοξότητας των $Z_i \sim N(0,1)$. Η ιδέα που γεννιάται εδώ είναι η ακόλουθη. Αν μπορούσαμε να αλλάξουμε τον συντελεστή λοξότητας των Z_i ώστε να γίνει ίσος με αυτόν των X_i τότε θα μπορούσαμε να επιτύχουμε μεγαλύτερη ταχύτητα σύγκλισης. Φυσικά αυτό δεν μπορεί να γίνει θεωρώντας ότι $Z_i \sim N(0,1)$ αλλά θα πρέπει τα Z_i να ακολουθούν μια άλλη κατανομή με τρεις παραμέτρους (σχήματος, κλίμακας και θέσης). Η κατανομή που θα πρέπει να επιλεγεί θα πρέπει να είναι *μη συμμετρική με τρεις παραμέτρους, συνεχής και απείρως διαιρετή* (infinitely divisible, ώστε να μπορεί να εμφανίζεται ως κατανομή αθροίσματος n αν.ισ. τ.μ., για κάθε n). Η απλούστερη τέτοια κατανομή είναι η κατανομή Γάμμα $G(a, \lambda, c)$ (μετατοπισμένη κατά c), με σ.π.π.

$$f_{G(a, \lambda, c)}(x) = \frac{\lambda^a}{\Gamma(a)} (x - c)^{a-1} e^{-\lambda(x-c)}, x \geq c.$$

Έστω λοιπόν γενικά ότι έχουμε αν.ισ. τ.μ $X_1, X_2, \dots \sim F$ με $E(X_i) = 0, E(X_i^2) = 1, E(X_i^3) = 2/\lambda > 0, E(X_i^4) < \infty$, για κάποιο $\lambda > 0$ (αν $E(X_i^3) < 0$, θεωρούμε τα $-X_i$). Η $G(a, \lambda, c)$ που έχει τις τρεις πρώτες της ροπές ίσες με τις παραπάνω είναι η $G(\lambda^2, \lambda, -\lambda)$

$\equiv SG_\lambda$, την οποία θα καλούμε *τυποποιημένη κατανομή Γάμμα*. Θεωρούμε λοιπόν αν.ισ. τ.μ. $Z_1, Z_2, \dots \sim SG_\lambda$. Αρκεί τώρα να διερευνηθεί αν ισχύει η παραπάνω εικασία, ότι δηλαδή οι κατανομές των τ.μ.

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n X_i \sim F_n \quad \text{και} \quad \frac{1}{\sqrt{n}} \sum_{i=1}^n Z_i \sim SG_{\lambda\sqrt{n}},$$

συγκλίνουν μεταξύ τους με τάξη μεγαλύτερη του $n^{-1/2}$, δεδομένου ότι έχουν τις τρεις πρώτες ροπές ίσες. Πράγματι, αποδεικνύεται το ακόλουθο (βλ. Boutsikas, 2013).

Θεώρημα 1 (Ασυμπτωτικό φράγμα.) *Αν η σ.κ. F των X_i έχει απόλυτα συνεχές μέρος τότε*

$$\limsup n \cdot \mathbf{d}(F_n, SG_{\sqrt{n}\lambda}) \leq \frac{1}{2} \|\phi^{(4)}\| \zeta_4(F, SG_\lambda), \quad \lambda = 2/E(X^3),$$

$$\text{όπου } \|\phi^{(4)}\| = \int_{-\infty}^{\infty} |\phi^{(4)}(x)| dx = 4e^{\frac{-3-\sqrt{6}}{2}} \frac{e^{\sqrt{6}} \sqrt{3-\sqrt{6}+\sqrt{3+\sqrt{6}}}}{\sqrt{\pi/3}} \approx 2.8006.$$

Από το παραπάνω θεώρημα προκύπτει ότι η απόσταση $\mathbf{d}$ μεταξύ της κατανομής F_n του τυποποιημένου αθροίσματος $n^{-1/2} \sum_{i=1}^n X_i$ και της *τυποποιημένης γάμμα* $SG_{\lambda\sqrt{n}}$ φράσσεται άνω ασυμπτωτικά από την ποσότητα

$$\mathbf{D}_n^* = \frac{\|\phi^{(4)}\|}{2n} \zeta_4(F, SG_\lambda) =_{(4-cx)} \frac{\|\phi^{(4)}\|}{16n} \left| 1 + \frac{E(X^3)^2}{2} - \frac{E(X^4)}{3} \right|.$$

Η δεύτερη ισότητα ισχύει όταν οι F, SG_λ διατάσσονται σύμφωνα με την κυρτή διάταξη 4^{th} τάξης (γεγονός που ισχύει στις περισσότερες εφαρμογές). Άρα οι $F_n, SG_{\lambda\sqrt{n}}$ συγκλίνουν μεταξύ τους με τάξη n^{-1} , ενώ και οι δύο συγκλίνουν στην $N(0,1)$ πιο αργά, με τάξη $n^{-1/2}$, διότι (βλ. Boutsikas, 2011)

$$\mathbf{d}(F_n, N(0,1)) \sim \frac{\|\phi^{(3)}\|}{12\sqrt{n}} E(X^3), \quad (\|\phi^{(3)}\| = \frac{2+8e^{-3/2}}{\sqrt{2\pi}} \approx 1.51).$$

Επίσης αποδεικνύεται το ακόλουθο φράγμα για την ίδια απόσταση το οποίο είναι χειρότερο (μεγαλύτερο) από το αντίστοιχο ασυμπτωτικό του Θεωρήματος 1 αλλά έχει την σωστή τάξη σύγκλισης και έχει κλειστή μορφή.

Θεώρημα 2 (Ακριβές φράγμα για κάθε n .) *Αν $\mathbf{d}_1 = \mathbf{d}(F, SG_\lambda) < 1/8, \lambda \geq 3$ και*

$$\varepsilon = \frac{8\mathbf{d}_1}{1+6\mathbf{d}_1} + 27\beta < 1,$$

όπου $\beta = \zeta_4(F, SG_\lambda)$, τότε

$$\mathbf{d}(F_n, SG_{\sqrt{n}\lambda}) \leq \mathbf{D}_n = \frac{n}{(n-1)^2} \frac{3\beta}{(1-\varepsilon)} + \frac{(2\mathbf{d}_1)^n}{2}, \quad n \geq 2.$$

Παρακάτω δίνεται ένα σχέδιο απόδειξης του Θεωρήματος 1. Για τις ανστηρές αποδείξεις βλ. Boutsikas (2013).

Σχέδιο Απόδειξης. Η απόδειξη χοντρικά γίνεται σε τρία βήματα.

(*Βήμα 1*) Η απόσταση $\mathbf{d}(\sum_{i=1}^n X_i, \sum_{i=1}^n Z_i)$ όπου $X_i \sim F$, $Z_i \sim SG_\lambda$ ασυμπτωτικά διασπάται (Lindeberg's decomposition) σε ένα άθροισμα αποστάσεων της μορφής

$$\mathbf{d}(X_1 + W, Z_1 + W) + \mathbf{d}(X_2 + W, Z_2 + W) + \dots + \mathbf{d}(X_n + W, Z_n + W), \quad W \sim G((n-1)\lambda^2, \lambda).$$

(*Βήμα 2*) Χρησιμοποιώντας συγκεκριμένες ανισότητες εξομάλυνσης (βλ. Boutsikas, 2011) οι παραπάνω αποστάσεις φράσσονται άνω ως ακολούθως,

$$\mathbf{d}(X_i + W, Z_i + W) \leq \frac{1}{2} \|f_{G((n-1)\lambda^2, \lambda)}^{(4)}\| \zeta_4(F, SG_\lambda),$$

όπου $f_{G((n-1)\lambda^2, \lambda)}^{(4)}$ είναι η παράγωγος 4^{ης} τάξης της σ.π.π. της κατανομής γάμμα με παραμέτρους $(n-1)\lambda^2, \lambda$.

(*Βήμα 3*) Αποδεικνύεται επίσης ότι $n^2 \|f_{G((n-1)\lambda^2, \lambda)}^{(4)}\| \rightarrow \|\phi^{(4)}\|$, $n \rightarrow \infty$.

Από τα (1), (2), (3), αποδεικνύεται ότι, ασυμπτωτικά,

$$\mathbf{d}(\sum_{i=1}^n X_i, \sum_{i=1}^n Z_i) \leq \frac{\|\phi^{(4)}\|}{2n} \zeta_4(F, SG_\lambda).$$

3. ΠΑΡΑΔΕΙΓΜΑΤΑ

3.1. Η κατανομή του αθροίσματος λογαριθμοκανονικών τ.μ.

Η λογαριθμοκανονική κατανομή εμφανίζεται σε πολλές εφαρμογές σε διάφορες ερευνητικές περιοχές (π.χ. χρηματοοικονομικά, αναλογισμός, θεωρία αξιοπιστίας, τηλεπικοινωνίες, βιολογία, υδρολογία). Η κατανομή του αθροίσματος αν.ισ. λογαριθμοκανονικών τ.μ. δεν έχει κλειστή αναλυτική έκφραση και επομένως είναι πολύ αναγκαία η εύρεση μια ικανοποιητικής προσέγγισης. (βλ. π.χ. Mehta, Wu, Molisch and Zhang, 2007, Dufresne, 2004 και τις εκεί αναφορές). Η κατανομή αυτή είναι λοξή και επομένως η κανονική προσέγγιση δεν πρέπει να είναι τόσο ικανοποιητική, ιδιαίτερα για όχι μεγάλες τιμές του n . Έστω λοιπόν αν.ισ. τ.μ. $L_1, L_2, \dots \sim LN(\mu, \sigma)$. Θεωρούμε τις τυποποιημένες τ.μ.

$$X_i = \frac{L_i - E(L_i)}{\sqrt{V(L_i)}} = \frac{L_i - e^{\mu + \sigma^2/2}}{e^{\mu + \sigma^2/2} \sqrt{e^{\sigma^2} - 1}}, \quad i = 1, 2, \dots$$

Θα είναι $E(X_i) = 0$, $E(X_i^2) = 1$,

$$E(X_i^3) = (2 + e^{\sigma^2}) \sqrt{e^{\sigma^2} - 1} = \frac{2}{\lambda}, \quad E(X_i^4) = -3 + e^{2\sigma^2} (3 + e^{\sigma^2} (2 + e^{\sigma^2})) < \infty.$$

Από το Θεώρημα 1 προκύπτει ότι η απόσταση $\mathbf{d}(F_n, SG_{\sqrt{n}\lambda})$, $\lambda = 2/E(X_i^3)$, φράσσεται άνω (ασυμπτωτικά) από την ποσότητα

$$\mathbf{D}_n^* = \frac{\|\phi^{(4)}\|}{16n} \left| 1 + \frac{E(X^3)^2}{2} - \frac{E(X^4)}{3} \right| = \frac{\|\phi^{(4)}\|}{96} \frac{2e^{4\sigma^2} + e^{3\sigma^2} - 3e^{2\sigma^2}}{n},$$

όπου $\|\phi^{(4)}\|/96 \approx 0.029173$. Επίσης, ένα ακριβές άνω φράγμα (Θεώρ. 2) είναι το

$$\mathbf{D}_n = \frac{n}{(n-1)^2} \frac{3\beta}{(1-\varepsilon)} + \frac{(2\mathbf{d}_1)^n}{2} \approx_{\sigma=0.2} 0.0170544 \frac{n}{(n-1)^2}, n \geq 2.$$

Ενώ για την αντίστοιχη απόσταση από την κανονική ισχύει

$$\mathbf{d}(F_n, N(0,1)) \sim \frac{1.51}{12\sqrt{n}} (2 + e^{\sigma^2}) \sqrt{e^{\sigma^2} - 1}.$$

Για παράδειγμα, αν $\sigma = 0.2$, οι τιμές των παραπάνω ποσοτήτων υπολογίζονται στον Πίνακα 1 (οι τιμές των αποστάσεων $\mathbf{d}$ έχουν υπολογιστεί με σχετικά μεγάλη ακρίβεια χρησιμοποιώντας μεθόδους αριθμητικής ανάλυσης).

Πίνακας 1. Προσέγγιση κατανομής αθροίσματος λογαριθμοκανονικών τ.μ. ($\sigma = 0.2$)

n	$\mathbf{d}(F_n, N(0,1))$	$\mathbf{d}(F_n, SG_{\sqrt{n}\lambda})$ $\lambda \approx 3.25577$	$\mathbf{D}_n^*$ <i>Ασυμπτωτικό φράγμα</i>	$\mathbf{D}_n$ <i>Ακριβές φράγμα</i>
2	0.054265	0.003267	0.003276	0.0341953
4	0.038500	0.001633	0.001638	0.0075798
8	0.027274	0.000817	0.000819	0.0027844
16	0.019305	0.000409	0.000409	0.0012127
32	0.013657	0.000204	0.000204	0.0005678
64	0.009659	0.000102	0.000102	0.0002750
128	0.006831	0.000051	0.000051	0.0001353

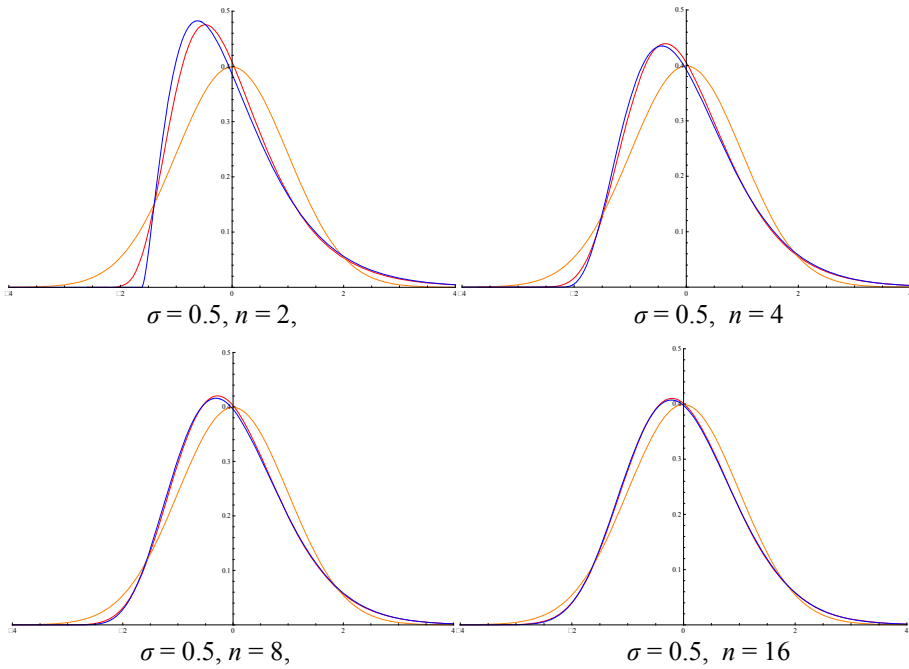
Επομένως συνοψίζοντας, αν $L_1, L_2, \dots \sim \text{LN}(\mu, \sigma)$, η κατανομή του $L_1 + L_2 + \dots + L_n$ προσεγγίζεται αρκετά ικανοποιητικά από την μετατοπισμένη κατανομή γάμμα,

$$G(n\lambda^2, \frac{\lambda}{\sigma_L}, n\mu_L - n\lambda\sigma_L),$$

όπου

$$\mu_L = E(L_i) = e^{\mu + \sigma^2/2}, \sigma_L = \sqrt{V(L_i)} = \sqrt{e^{\sigma^2} - 1} e^{\mu + \sigma^2/2}, \lambda = \frac{2}{(2 + e^{\sigma^2}) \sqrt{e^{\sigma^2} - 1}}.$$

Τα παρακάτω 4 γραφήματα απεικονίζουν την σ.π.π. του αθροίσματος των X_i , της $SG_{\sqrt{n}\lambda}$ (η καμπύλη που από αριστερά έχει τις μικρότερες τιμές), και της τυπικής κανονικής (συμμετρική καμπύλη, αριστερά έχει τις μεγαλύτερες τιμές) για διάφορες τιμές του n . Είναι φανερό ότι η σ.π.π. του αθροίσματος των X_i προσεγγίζεται πολύ ικανοποιητικότερα από την σ.π.π. της συγκεκριμένης γάμμα, αντί της $N(0,1)$.



3.2. Η κατανομή του αθροίσματος ανεξάρτητων τ.μ. Pareto

Έστω ότι $Z_1, Z_2, \dots \sim Pa(a)$ (Pareto) με σ.π.π. $f(x) = ax^{-a-1}, x > 1$. Η συγκεκριμένη κατανομή εμφανίζεται σε πολλές εφαρμογές (οικονομικά, αναλογισμός, υδρολογία, θεωρία αξιοπιστίας κ.λπ.). Επίσης χρησιμοποιείται στη θεωρία ακραίων τιμών για να εκφράσει τη βαριά ουρά κατανομών (Peaks Over Threshold method). Για παράδειγμα, οι υπερβάσεις τ.μ. με βαριά δεξιά ουρά (regularly varying ουρά με δείκτη $-a$) πάνω από ένα υψηλό κατώφλι, ακολουθούν ασυμπτωτικά (όταν το κατώφλι τείνει στο άπειρο) κατανομή $Pa(a)$. Επομένως, το σύνολο των υπερβάσεων πάνω από ένα υψηλό κατώφλι, αν.ισ. τ.μ. που έχουν βαριά δεξιά ουρά, μπορεί να θεωρηθεί ως ένα άθροισμα αν.ισ. τ.μ. που ακολουθούν κατανομή Pareto. Η ροπή k -τάξης της $Pa(a)$ είναι πεπερασμένη μόνο όταν $k < a$ και ως εκ τούτου υποθέτουμε $a > 4$ έτσι ώστε να έχουμε πεπερασμένες τις πρώτες τέσσερις ροπές. Σημειώνεται ότι, όταν $0 < a < 2$, η κατανομή του αθροίσματος των $Z_1, Z_2, \dots$ συνήθως προσεγγίζεται μέσω του γενικευμένου ΚΟΘ από μια a -stable κατανομή (π.χ. βλέπε Zaliapin, Kagan and Schoenberg, 2005 και τις εκεί αναφορές). Για ακέραιο a υπάρχουν προσεγγιστικές εκφράσεις μέσω της αναστροφής του μετασχηματισμού Laplace (βλ. Ramsay, 2006). Θα επιχειρήσουμε να προσεγγίσουμε την κατανομή του αθροίσματος των Z_i μέσω μιας μετατοπισμένης κατανομής Γάμμα. Θεωρούμε τις τυποποιημένες τ.μ.

$$X_i = \frac{Z_i - E(Z_i)}{\sqrt{V(Z_i)}} = \frac{Z_i - \frac{a}{a-1}}{\sqrt{\frac{a}{a-2} - \frac{a^2}{(a-1)^2}}}, i = 1, 2, \dots,$$

για τις οποίες ισχύει ότι $E(X_i) = 0, E(X_i^2) = 1$,

$$E(X_i^3) = \frac{2(a+1)\sqrt{a-2}}{\sqrt{a}(a-3)} = \frac{2}{\lambda}, \quad E(X_i^4) = \frac{3(a-2)(2+a+3a^2)}{a(a-3)(a-4)} < \infty.$$

Σύμφωνα με τα παραπάνω, ισχύει προσεγγιστικά ότι

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n X_i \sim SG_{\lambda\sqrt{n}}, \quad \lambda = \frac{\sqrt{a}(a-3)}{(a+1)\sqrt{a-2}},$$

ή ισοδύναμα,

$$\sum_{i=1}^n Z_i \sim G\left(\frac{na(a-3)^2}{(a+1)^2(a-2)}, \frac{(a-3)(a-1)}{(a+1)}, \frac{na(a-1)}{(a+1)(a-2)}\right).$$

Η απόσταση $\mathbf{d}(F_n, SG_{\sqrt{n}\lambda})$, $\lambda = 2/E(X_i^3)$, φράσσεται άνω (ασυμπτωτικά) από το

$$\mathbf{D}_n^* = \frac{\|\phi^{(4)}\|}{n} \frac{(a-1)^3}{4(a-4)(a-3)^2 a}, \quad \|\phi^{(4)}\| \approx 2.8.$$

Ενώ η αντίστοιχη απόσταση από την κανονική είναι

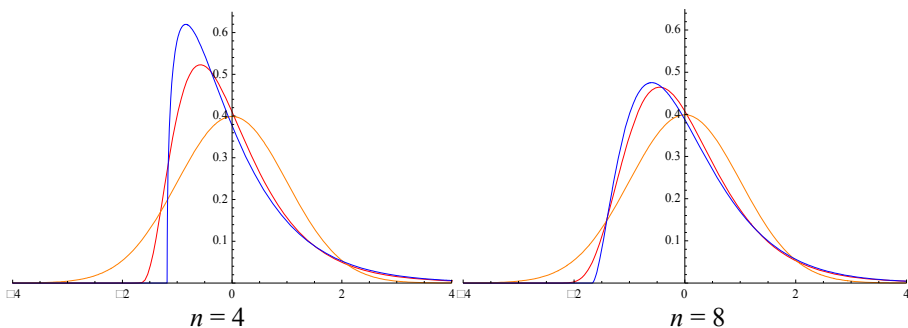
$$\mathbf{d}(F_n, N(0,1)) \sim \frac{\|\phi^{(3)}\|}{12\sqrt{n}} |E(X^3)| = \frac{\|\phi^{(3)}\|}{\sqrt{n}} \frac{(a+1)\sqrt{a-2}}{6(a-3)\sqrt{a}}, \quad \|\phi^{(3)}\| \approx 1.51.$$

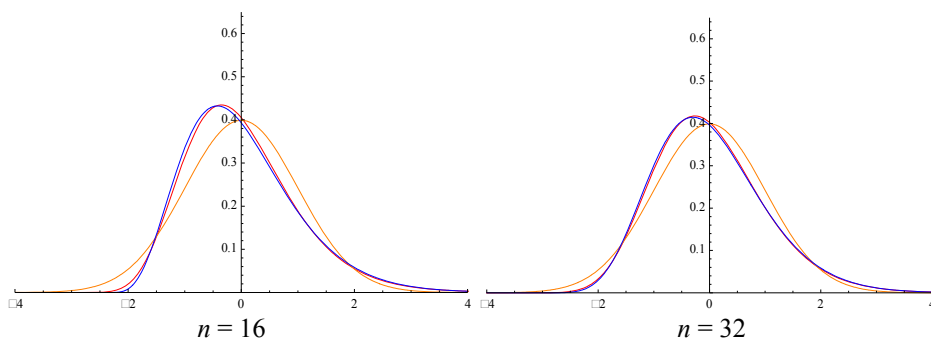
Π.χ. για $a = 7$ ($\lambda \approx 0.5916$), οι παραπάνω ποσότητες δίνονται στον ακόλουθο πίνακα.

Πίνακας 2. Προσέγγιση κατανομής αθροίσματος Pareto τ.μ. ($a = 7$)

n	$\mathbf{d}(F_n, N(0,1))$	$\mathbf{d}(F_n, SG_{\sqrt{n}\lambda})$	$\mathbf{D}_n^*$
2	0.2670	0.1966	0.2250
4	0.1889	0.0829	0.1125
8	0.1351	0.0369	0.0562
16	0.0978	0.0187	0.0281
32	0.0707	0.0099	0.0140
64	0.0509	0.0053	0.0070

Τα παρακάτω 4 γραφήματα απεικονίζουν όμοια με το προηγούμενο παράδειγμα την σ.π.π. του αθροίσματος των X_i , της $SG_{\sqrt{n}\lambda}$ (η καμπύλη που από αριστερά έχει τις μικρότερες τιμές), και της τυπικής κανονικής (συμμετρική καμπύλη, αριστερά έχει τις μεγαλύτερες τιμές) για διάφορες τιμές του n . Και εδώ, η σ.π.π. του αθροίσματος των X_i προσεγγίζεται πολύ ικανοποιητικότερα από την σ.π.π. της συγκεκριμένης γάμμα.





ABSTRACT

Sharp upper bounds are offered for the total variation distance between the distribution of a sum of n independent random variables following a skewed distribution with an absolutely continuous part, and an appropriate shifted gamma distribution. These bounds vanish at a rate n^{-1} as $n \rightarrow \infty$, while the corresponding distance to the normal distribution vanishes at a rate $n^{-1/2}$, implying that, for skewed summands, pre-asymptotic (penultimate) gamma approximation is much more accurate than the usual normal approximation. Two illustrative examples concerning lognormal and Pareto summands are presented along with numerical comparisons confirming the aforementioned ascertainment.

ΑΝΑΦΟΡΕΣ

- Barbour A.D. and Hall P. (1984) Stein's method and the Berry-Esseen theorem. *Australian J. Statist.* **26**, 8-15.
- Boutsikas, M.V. (2011) Asymptotically optimal Berry-Esseen-type bounds for distributions with an absolutely continuous part. *Journal of Statistical Planning and Inference*, **141**, 1250-1268.
- Boutsikas, M.V. (2013) Penultimate gamma approximation in the CLT for skewed distributions. (*submitted for publication*)
- Boutsikas, M.V. and Vaggelatou, E. (2002). On the distance between convex-ordered random variables. *Advances in Applied Probability* **34**, 349-374.
- Dufresne, D. (2004) The log-normal approximation in financial and other computations. *Advances in Applied Probability* **36**, 747-773.
- Mehta N., Wu J., Molisch A., Zhang J. (2007) Approximating a Sum of Random Variables with a Lognormal. *IEEE Transactions on Wireless Communications* **6**, 2690-2699.
- Petrov, V.V. (1995). *Limit Theorems of Probability Theory*. Clarendon Press, Oxford.
- Rachev S.T. (1991) *Probability Metrics and the Stability of Stochastic Models*. Wiley, Chichester, New York.
- Ramsay C.M. (2006) The Distribution of Sums of Certain I.I.D. Pareto Variates. *Communications in Statistics - Theory and Methods*, **35**:3, 395-405.
- Zaliapin I.V., Kagan Y.Y. and Schoenberg F.P. (2005) Approximating the Distribution of Pareto Sums. *Pure appl. geophys.* **162**, 1187--1228.



ΜΕΤΡΑ ΚΕΝΤΡΙΚΟΤΗΤΑΣ ΣΕ ΔΙΚΤΥΑ

Μουσιάδης Πολυχρόνης

Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης

cmoi@math.auth.gr

ΠΕΡΙΛΗΨΗ

Για τη μελέτη και κατανόηση της δομής των γραφημάτων χρειαζόμαστε διάφορες μετρικές και αποστάσεις. Πολλές από αυτές ορίστηκαν αρχικά σε κοινωνικά δίκτυα (social networks), και στο γεγονός αυτό οφείλουν τις ονομασίες τους, όμως σήμερα, έχει γενικευτεί η χρήση τους και σε άλλα δίκτυα, όπως βιολογικά, υπολογιστών, τυχαία, κλπ. Στην παρούσα εργασία γίνεται μια διδακτική προσέγγιση στα βασικά μέτρα κεντρικότητας (centrality measures), στον ορισμό, τη σημασία, τον υπολογισμό και τους περιορισμούς εφαρμογής των. Γίνεται επίσης σύγκριση των μέτρων αυτών σε απλά και τυχαία δίκτυα.

Λέξεις Κλειδιά: μέτρα (measures), κεντρικότητα (centrality), κοινωνικά δίκτυα (social networks), τυχαία δίκτυα (random networks).

1. ΜΕΤΡΑ ΚΕΝΤΡΙΚΟΤΗΤΑΣ

Η πρώτη αναζήτηση σε ένα δίκτυο, συνήθως, αφορά την εύρεση της σημαντικότερης κορυφής. Η σημαντικότερη κορυφή δεν ορίζεται μονοσήμαντα διότι εξαρτάται από το κριτήριο που χρησιμοποιήθηκε, δηλαδή από το μέτρο κεντρικότητας με το οποίο την βρήκαμε. Έχουν οριστεί δεκάδες τέτοια μέτρα, μερικά από τα οποία περιγράφονται στη συνέχεια:

1.1 Βαθμική Κεντρικότητα (Degree Centrality)

Το απλούστερο μέτρο που ορίζεται στα δίκτυα είναι ο βαθμός (degree) που ορίζεται ως το πλήθος ακμών που συνδέονται με τη συγκεκριμένη κορυφή και προφανώς είναι ακέραιος αριθμός μεταξύ 0 και $n-1$, όπου n το πλήθος κορυφών του γραφήματος. Ως βαθμική κεντρικότητα της κορυφής v ορίζεται το πηλίκο $\frac{\deg(v)}{n-1}$. Αν το δίκτυο είναι κατευθυνόμενο διακρίνουμε έσω-βαθμό (in-degree), δηλαδή πλήθος ακμών που κατευθύνονται προς την κορυφή, και έξω-βαθμό (out-degree), δηλαδή πλήθος ακμών που απομακρύνονται από την κορυφή. Στην περίπτωση αυτή ορίζεται αντίστοιχα εσω- ή έξω-βαθμική κεντρικότητα, αντικαθιστώντας το $\deg(v)$ με $\deg_{in}(v)$, ή $\deg_{out}(v)$, ανάλογα με το τι μας ενδιαφέρει. Ας δούμε την κεντρικότητα αυτή στα παρακάτω παραδείγματα:

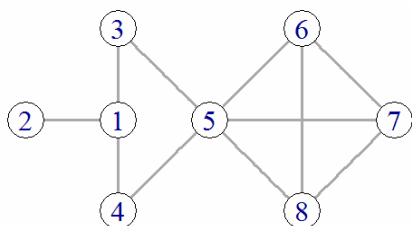
Παράδειγμα 1. Ας θεωρήσουμε το δίκτυο του Σχήματος 1. Το διάνυσμα βαθμικής κεντρικότητας των κορυφών του, είναι $d=(\frac{3}{7}, \frac{1}{7}, \frac{2}{7}, \frac{2}{7}, \frac{5}{7}, \frac{3}{7}, \frac{3}{7}, \frac{3}{7})'$ και υπολογίζεται από τον τύπο ή με χρήση κάποιου λογισμικού. Π.χ. με την εντολή `degree(g)` στο `igraph` του πακέτου `R`, έχουμε:

$$d=(0.429, 0.143, 0.286, 0.286, 0.714, 0.429, 0.429, 0.429)'$$

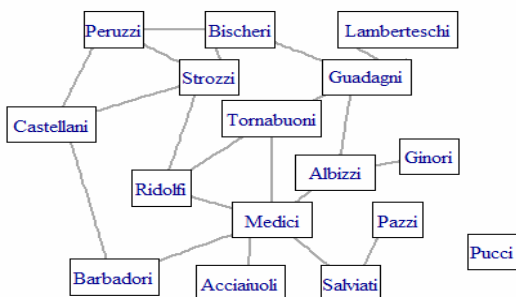
Στο γράφημα αυτό η κορυφή 5 έχει τη μεγαλύτερη βαθμική κεντρικότητα, ενώ ακολουθούν με την ίδια κεντρικότητα οι κορυφές 1, 6, 7 και 8. Όμως, η σημαντικότητα των κορυφών 6,7 και 8 φαίνεται και από το σχήμα ότι είναι μεγαλύτερη από αυτήν της κορυφής 1. Μία τυποποιημένη μορφή της βαθμικής κεντρικότητας προκύπτει με διαίρεση όλων των τιμών του d με το μέγιστο βαθμό. Στη μορφή αυτή η σημαντικότερη κορυφή έχει βαθμό 1. Στο παράδειγμά μας έχουμε:

$$d=(0.6, 0.2, 0.4, 0.4, 0.1, 0.6, 0.6, 0.6)'$$

Σχήμα 1: Ένα απλό δίκτυο.



Σχήμα 2: Το δίκτυο Των Μεδίκων



Παράδειγμα 2. Στα κοινωνικά δίκτυα μεγαλύτερος βαθμός μιας κορυφής (ατόμου) από άλλες με μικρότερο βαθμό, εκφράζει μεγαλύτερη επιρροή ή επίδραση ή σημαντικότητα αυτού του ατόμου στο συγκεκριμένο δίκτυο, από άλλα άτομα του δικτύου. Το Σχήμα 2 παριστάνει ένα γνωστό κοινωνικό δίκτυο με κόμβους κάποιες μεγάλες οικογένειες της Φλωρεντίας του 15^{ου} αιώνα και με συνδέσεις μεταξύ των οικογενειών που να παριστάνουν γάμους.

Στο δίκτυο αυτό φαίνεται ότι η οικογένεια των Medici κατέχει μια κεντρική θέση με γάμους με έξι άλλες οικογένειες, που μπορεί να εξηγήσει και την επίδραση αυτής της οικογένειας στα πράγματα της εποχής. Αμέσως μετά είναι οι οικογένειες Guadagni και Strozzi. Το διάνυσμα με τη βαθμική κεντρικότητα βρίσκεται ίσο με:

Med	Gua	Str	Alb	Bis	Cas	Per	Rid	Tor
0.400	0.267	0.267	0.200	0.200	0.200	0.200	0.200	0.200
Bar	Sal	Acc	Gin	Lam	Paz	Puc		
0.133	0.133	0.067	0.067	0.067	0.067	0.000		

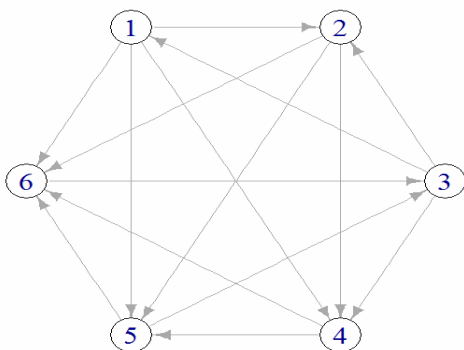
Παράδειγμα 3. Ο βαθμός κορυφής είναι το πρώτο μέτρο που λαμβάνει κανείς υπόψη του σε ένα round-robin τουρνουά, όπου κάθε παίκτης παίζει με όλους τους άλλους, για την ανάδειξη του νικητή. Ένα τέτοιο τουρνουά παριστάνεται από ένα πλήρες κατευθυνόμενο γράφημα (Σχήμα 3), όπου η κατεύθυνση της ακμής π.χ. από την κορυφή

3 στην 1, σημαίνει ότι ο παίκτης 3 νίκησε τον παίκτη 1. Στο γράφημα αυτό, το διάνυσμα βαθμικής κεντρικότητας (με τους έξω-βαθμούς) είναι:

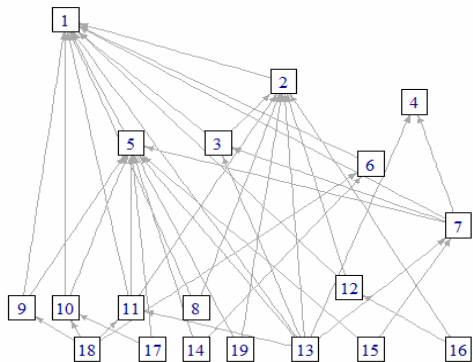
$$d=(0.8, 0.6, 0.6, 0.4, 0.4, 0.2)'$$

απ' όπου ο παίκτης 1 έχει τη μεγαλύτερη (έξω-) βαθμική κεντρικότητα, που στην περίπτωση αυτή σημαίνει ότι (ως προς αυτό το μέτρο) είναι ο νικητής.

Σχήμα 3: Ένα round robin τουρνουά



Σχήμα 4: Δίκτυο αναφορών



Παράδειγμα 4. Ο βαθμός κορυφών έχει κεντρική σημασία και σε μη κοινωνικά δίκτυα, όπως π.χ. τα δίκτυα αναφορών επιστημονικών εργασιών (citations), όπου οι κορυφές είναι εργασίες και οι συνδέσεις είναι οι αναφορές-παραπομπές που παραπέμπουν στη συγκεκριμένη εργασία (έσω βαθμός) ή αναφέρονται σ' αυτήν (έξω-βαθμός). Παράδειγμα τέτοιου δικτύου είναι στο Σχήμα 4, όπου ο έσω-βαθμός της εργασίας που αντιστοιχεί στην κορυφή 1 είναι 10, που σημαίνει ότι δέκα εργασίες αναφέρονται σ' αυτήν. Εδώ δηλαδή ενδιαφέρει περισσότερο η (εσω-) βαθμική κεντρικότητα, που βρίσκεται ίση με:

$$d=(0.579, 0.368, 0.105, 0.105, 0.421, 0.105, 0.105, 0.000, 0.053, 0.105, 0.105, 0.053, 0.000, 0.000, 0.000, 0.000, 0.000, 0.000, 0.000)'$$

που δίνει ως σημαντικότερη την εργασία 1, και ακολουθεί η εργασία 5.

1.2. Ιδιοκεντρικότητα (Eigenvector Centrality).

Φυσική γενίκευση της βαθμικής κεντρικότητας είναι, όπως περιγράφεται κατωτέρω, η ιδιοκεντρικότητα. Ας παρατηρήσουμε πρώτα το γράφημα του Σχήματος 1. Η κορυφή 1 έχει βαθμό 3, ενώ οι γειτονικές της κορυφές έχουν συνολικό άθροισμα βαθμών 5. Η κορυφή 6 έχει επίσης βαθμό 3, όμως οι γειτονικές της κορυφές έχουν συνολικό άθροισμα βαθμών 11. Αυτό μπορεί να ερμηνευτεί ότι η κορυφή 6 έχει μεγαλύτερη σημαντικότητα από την κορυφή 1. Όμοια, ας δούμε το γράφημα του σχήματος 3. Σ' αυτό όπως ήδη αναφέραμε ο παίκτης 1 έχει 4 νίκες, ενώ ο παίκτης 3 έχει τρεις νίκες. Αν όμως αθροίσουμε τις νίκες των παικτών που νίκησε ο παίκτης 1, θα βρούμε 8 «νίκες δεύτερης τάξης». Ενώ για τον παίκτη 3 βρίσκουμε 9 «νίκες δεύτερης τάξης». Δηλαδή, μπορούμε να ερμηνεύσουμε ότι ο παίκτης 3 έχει μεγαλύτερο δυναμικό από τον 1 και άρα αυτός πρέπει να είναι ο νικητής.

Στα κοινωνικά δίκτυα φιλίας, όπου αναζητούμε τα άτομα μεγάλης επιρροής, η παραπάνω παρατήρηση σημαίνει ότι πρέπει να λαμβάνουμε υπόψη μας όχι μόνο με πόσα άτομα του δικτύου συνδέονται, αλλά και το συνολικό πλήθος των ατόμων που συνδέονται με τα άτομα αυτά. Εμφανίζεται, λοιπόν η αναγκαιότητα για μια άλλη κεντρικότητα που να μπορεί να αποκαλύψει αυτή την έμμεση επιρροή.

Έστω A ο πίνακας αντιστοίχισης ενός γραφήματος. Με τον όρο αυτό εννοούμε τον $(0,1)$ -πίνακα με γραμμές και στήλες που παριστάνουν τις κορυφές και με στοιχεία $A_{ij}=1$ αν η κορυφή της i -γραμμής συνδέεται με την κορυφή της j -στήλης και 0 αλλιώς. Αν θέσουμε $x_0=(1, 1, 1, \dots, 1)'$ ένα διάνυσμα με όλα τα στοιχεία ίσα με 1, τότε το γινόμενο $x_1=Ax_0=d(v)$ δίνει του βαθμούς των κορυφών, από όπου βρίσκεται αμέσως η κορυφή (ή οι κορυφές) με μέγιστο βαθμό.

Αν θεωρήσουμε την αναδρομική σχέση $x_i=Ax_{i-1}$, $i=1, 2, \dots$, τότε, για κάθε i , θα βρούμε το άθροισμα βαθμών των κορυφών που τις προσεγγίζουμε με i βήματα. Π.χ. για $i=2$ βρίσκουμε το άθροισμα των βαθμών των άμεσα συνδεδεμένων κορυφών. Για $i=3$ βρίσκουμε το άθροισμα βαθμών των κορυφών που τις προσεγγίζουμε με δύο βήματα κ.ο.κ. Για $i=n$ έχουμε:

$$x_n=A^n x_0$$

Θέτοντας $x_0 = \sum_i c_i v_i$, όπου v_i τα ιδιοδιανύσματα του πίνακα αντιστοίχισης A , βρίσκουμε:

$$x_n = A^n \sum_i c_i v_i = \sum_i c_i A^n v_i = \sum_i c_i k_i^n v_i = k_1^n \sum_i c_i \left(\frac{k_i}{k_1}\right)^n v_i,$$

όπου k_i είναι οι ιδιοτιμές του A σε φθίνουσα σειρά, και επομένως $\frac{k_i}{k_1} < 1$. Για αρκούντως μεγάλο n μηδενίζονται όλοι οι όροι του τελευταίου αθροίσματος εκτός από αυτόν που περιέχει το v_1 δηλαδή το ιδιοδιάνυσμα που αντιστοιχεί στην μεγαλύτερη ιδιοτιμή λ_1 , το οποίο μάλιστα το κανονικοποιούμε. Το διάνυσμα v_1 δίνει επομένως τη ζητούμενη κεντρικότητα.

Η κεντρικότητα αυτή ονομάζεται ιδιοκεντρικότητα (eigenvalue centrality), λόγω του ορισμού της και τέθηκε αρχικά από τον Bonacich το 1987. Η ύπαρξη αυτής της κεντρικότητας, δηλαδή ενός ιδιοδιανύσματος, με όλα τα στοιχεία του θετικά, που αντιστοιχεί σε μια μοναδική μέγιστη ιδιοτιμή εξασφαλίζεται από την παρατήρηση ότι εφόσον ο A είναι συμμετρικός πίνακας και πρωταρχικός, δηλαδή υπάρχει k ώστε ο πίνακας A^k να έχει όλα τα στοιχεία του θετικά, ικανοποιεί τις προϋποθέσεις του θεωρήματος Peron-Frobenius. Ο πίνακας αντιστοίχισης είναι συμμετρικός στα απλά γραφήματα και αν δεν είναι ο ίδιος πρωταρχικός τότε θα είναι ο $A+I$, που έχει ακριβώς τα ίδια ιδιοδιανύσματα με τον A .

Θεωρητικά η ιδιοκεντρικότητα μπορεί να υπολογιστεί και για απλά και για κατευθυνόμενα γραφήματα. Εντούτοις, επειδή στα κατευθυνόμενα γραφήματα ο πίνακας αντιστοίχισης είναι μη-συμμετρικός, δεν ορίζονται μονοσήμαντα τα ιδιοδιανύσματα.

Αυτό έχει ως συνέπεια η ιδιοκεντρικότητα να μη λειτουργεί πολύ καλά στα κατευθυνόμενα γραφήματα.

Η βιβλιοθήκη `igraph` του πακέτου `R` έχει την εντολή `evcent` που υπολογίζει τις ιδιοτιμές και τα ιδιοδιανύσματα του πίνακα αντιστοίχισης και επιλέγοντας το πρώτο ιδιοδιάνυσμα που αντιστοιχεί στη μέγιστη ιδιοτιμή, προκύπτει η ζητούμενη ιδιοκεντρικότητα. Επίσης η βιβλιοθήκη `sna` έχει την εντολή `evcent` που υπολογίζει το ιδιοδιάνυσμα που αντιστοιχεί στη μέγιστη ιδιοτιμή, δηλαδή απευθείας την ιδιοκεντρικότητα.

Στα παραδείγματα που αναφέρθηκαν βρίσκουμε:

Για το παράδειγμα 1, η ιδιοκεντρικότητα δίνεται με το διάνυσμα:

$$ev = (0.147, 0.045, 0.214, 0.214, 0.544, 0.443, 0.443, 0.443)'$$

που αποκαλύπτει ότι η κορυφή 5 είναι η πιο σημαντική, ακολουθούν με ίδια σημαντικότητα οι κορυφές 6, 7, 8 και αρκετά πιο χαμηλά είναι η κορυφή 1.

Για το παράδειγμα 2. Εδώ οι ιδιοκεντρικοί βαθμοί είναι:

Med	Str	Rid	Tor	Gua	Bis	Per	Cas	Alb
0.430	0.356	0.342	0.326	0.289	0.283	0.276	0.259	0.244
Bar	Sal	Acc	Lam	Gin	Paz	Puc		
0.212	0.146	0.132	0.089	0.075	0.045	0.000		

από όπου προκύπτει ότι η οικογένεια των Medici που ήταν κεντρική στη βαθμική κεντρικότητα, εξακολουθεί να είναι κεντρική και στην ιδιοκεντρικότητα. Ενώ από τις δύο οικογένειες Guadagni και Strozzi που έχουν και οι δύο βαθμό 4 (δηλαδή ισοδυναμούν στη δεύτερη θέση ως προς τη βαθμική κεντρικότητα), ως προς την ιδιοκεντρικότητα κατέχουν η μεν Strozzi την 2^η θέση, η δε Guadagni την 5^η θέση. Και ότι σ' αυτήν την κεντρικότητα, οι δύο οικογένειες Ridolfi και Tornabuoni (με βαθμό κορυφής 3) ξεπερνούν την οικογένεια Guadagni.

Για το παράδειγμα 3. Επειδή το γράφημα είναι κατευθυνόμενο, η `evcent` της βιβλιοθήκης `igraph` δε δίνει σωστό αποτέλεσμα, διότι οι ιδιοτιμές δεν είναι πραγματικές. Όμως η `evcent` της `sna` δίνει ιδιοκεντρικότητα:

$$(0.556, 0.384, 0.540, 0.265, 0.350, 0.242)'$$

που είναι η σωστή. Αυτό ελέγχεται από τον υπολογισμό μιας μεγάλης σειράς διανυσμάτων από την αναδρομική σχέση:

$$x_i = Ax_{i-1}, i=1, 2, 3, \dots, \text{όπου } x_0 = (1, 1, 1, \dots, 1)'$$

τα οποία στη συνέχεια τα κανονικοποιούμε (δηλ. να έχουν μέτρο 1). Για i κοντά στο 30 το διάνυσμα σταθεροποιείται σε τρία δεκαδικά ψηφία με αυτό που δώσαμε πριν. Για την κανονικοποίηση μπορούμε να εργαστούμε είτε απ' ευθείας ή να χρησιμοποιήσουμε την εντολή `normalize` της βιβλιοθήκης `rpls` της `R`. Η ιδιοκεντρικότητα που βρέθηκε δείχνει, ότι καλύτερος παίκτης είναι ο παίκτης 1, παρόλο που αρχικά φαινόταν να είναι καλύτερος παίκτης, ο παίκτης 3 με δεύτερο τον παίκτη 1.

Για το παράδειγμα 4. Στο γράφημα αυτό, που είναι επίσης κατευθυνόμενο, μας ενδιαφέρει ο έσω-βαθμός και όχι ο έξω-βαθμός όπως στο παράδειγμα 3. Το διάνυσμα των έσω βαθμών

$$x_1=(11, 7, 2, 2, 8, 2, 2, 0, 1, 2, 2, 1, 0, 0, 0, 0, 0, 0, 0)'$$

αναδεικνύει ως κεντρική εργασία την 1^η και αμέσως σημαντικότερη την 5^η και μετά την 2^η. Για να βρούμε το πλήθος μονοπατιών μήκους δύο προς τις κορυφές αρκεί να πολλαπλασιάσουμε από αριστερά τον πίνακα αντιστοίχισης A με x_1 . Βρίσκουμε:

$$x_2=x_1'A=(19, 5, 2, 2, 7, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)'$$

που επιβεβαιώνει το προηγούμενο συμπέρασμα.

1.3. Κεντρικότητα Katz.

Είναι γενίκευση της ιδιοκεντρικότητας. Προκύπτει με τροποποίηση της αναδρομικής σχέσης που ορίζει την ιδιοκεντρικότητα με την εισαγωγή δύο θετικών σταθερών α και β , έτσι ώστε να δοθεί μία «ποσότητα κεντρικότητας» σε όλους τους κόμβους του γραφήματος. Η αναδρομική σχέση που ορίζει την Katz κεντρικότητα είναι η:

$$x_i = \alpha \sum_j A_{ij} x_j + \beta$$

, προτάθηκε το 1957 και ισοδυναμεί με τον τύπο:

$$x = \beta(I - \alpha A)^{-1} \mathbf{1}.$$

Ως β συνήθως παίρνουμε το 1. Για το α , παρατηρούμε ότι αν $\alpha < 1/\lambda_1$, τότε το διάνυσμα μηδενίζεται, οπότε είναι σκόπιμο να πάρουμε μία τιμή του α μικρότερη του $1/\lambda_1$ ώστε η διαδικασία να συγκλίνει. Η κεντρικότητα αυτή λύνει κάποια προβλήματα που εμφανίζονται στον υπολογισμό της ιδιοκεντρικότητας, ιδίως όταν υπάρχουν κορυφές με μηδενικό βαθμό κεντρικότητας.

Για το παράδειγμα 4 των αναφορών βρίσκουμε ότι $\lambda_1=5.49$, έτσι λαμβάνοντας π.χ. $\alpha=0.15$ βρίσκουμε Katz κεντρικότητα:

$$(0.485, 0.339, 0.210, 0.210, 0.369, 0.203, 0.203, 0.156, 0.180, 0.203, \\ 0.203, 0.180, 0.156, 0.156, 0.156, 0.156, 0.156, 0.156, 0.156)'$$

που οδηγεί στο ίδιο συμπέρασμα με τις άλλες κεντρικότητες: ότι δηλαδή η εργασία 1 έχει το μεγαλύτερο βαθμό, με δεύτερη την εργασία 5. Άλλες τιμές του α δίνουν διαφορετικά διανύσματα, όμως το συμπέρασμα παραμένει ίδιο.

1.4 Βαθμική κεντρικότητα Page (PageRank centrality).

Είναι βελτίωση της Katz κεντρικότητας για δίκτυα ιστοσελίδων και δόθηκε αυτό το όνομα από τον Larry Page, έναν από τους ιδρυτές της Google, που με αυτήν την κεντρικότητα αξιολόγησε τις συνδέσεις των σελίδων της. Η κεντρικότητα Katz έχει το εξής μειονέκτημα: αν ένας κόμβος έχει μεγάλη κεντρικότητα, τότε και όσοι συνδέονται με αυτήν έχουν επίσης μεγάλη κεντρικότητα, χωρίς να διακρίνει ποιοτικά τις συνδέσεις. Ένας τρόπος στάθμισης της κεντρικότητας είναι το να διαιρέσουμε με τον έξω-βαθμό, θεωρώντας σημαντικότερες τις σελίδες που παραπέμπουν σε μικρότερο πλήθος άλλων σελίδων, ενώ δέχονται πολλές παραπομπές οι ίδιες. Για παράδειγμα κάποια σελίδα με λίγες παραπομπές από άλλες σελίδες, στην οποία όμως παραπέμπει π.χ. η Google, θα έχει μεγαλύτερη κεντρικότητα Page από άλλη σελίδα στην οποία παραπέμπουν πολλές «μικρότερες» σελίδες. Η σχέση που ορίζει τη βαθμική κεντρι-

κότητα Page είναι η:

$$x_i = \alpha \sum_j A_{ij} \frac{x_j}{k_j^{out}} + \beta.$$

Στη σχέση αυτή έχουμε πρόβλημα όταν υπάρχουν κορυφές με έξω-βαθμό 0. Επειδή όμως αυτές οι κορυφές δεν προσφέρουν τίποτε στην κεντρικότητα, μπορούμε να αντικαταστήσουμε το βαθμό αυτών των κορυφών με 1. Με μορφή πινάκων η κεντρικότητα Page, υπολογίζεται από τον τύπο $x = \alpha AD^{-1}x + \beta \mathbf{X}$, όπου D διαγώνιος πίνακας με στοιχεία $D_{ii} = \max\{1, k_i^{out}\}$ και η λύση (για $\beta=1$) γράφεται:

$$x = D(D - \alpha A)^{-1} \mathbf{X}.$$

Όπως και στην κεντρικότητα Katz, πρέπει και εδώ να θέσουμε μια τιμή στην ελεύθερη παράμετρο α . Μια καλή επιλογή είναι να πάρουμε α μικρότερο του αντιστρόφου της μεγαλύτερης ιδιοτιμής του πίνακα AD^{-1} . Η Google παίρνει αυθαίρετα την τιμή $\alpha=0.85$.

Για το παράδειγμα 4 των αναφορών βρίσκουμε βαθμική κεντρικότητα Page:

(0.261, 0.105, 0.038, 0.038, 0.114, 0.043, 0.041, 0.026, 0.032, 0.043,
0.035, 0.038, 0.026, 0.026, 0.026, 0.026, 0.026, 0.026, 0.026)

που δίνει το ίδιο συμπέρασμα όπως και πριν.

Τα τέσσερα μέτρα κεντρικότητας συνδέονται με το παρακάτω σχήμα:

	με σταθερό όρο	Χωρίς σταθερό όρο
διαίρεση με έξω-βαθμό	$x = D(D - \alpha A)^{-1} \mathbf{X}$ βαθμική κεντρικότητα Page	$x = \alpha D^{-1} \mathbf{X}$ βαθμική κεντρικότητα
χωρίς διαίρεση	$x = (I - \alpha A)^{-1} \mathbf{X}$ Katz κεντρικότητα	$x = \lambda_1^{-1} A \mathbf{X}$ ιδιοκεντρικότητα

1.5 Κεντρικότητα εγγύτητας (closeness centrality).

Αν για την κορυφή i η τιμή d_{ij} , $j=1, 2, \dots, d_i$ παριστάνει την απόστασή της από τις άλλες κορυφές, τότε η ποσότητα $l_i = \frac{1}{n-1} \sum_{j \neq i} d_{ij}$ είναι μικρή για τις κεντρικές κορυφές, μεγαλύτερη για τις απόμακρες, και δίνει ένα μέτρο για το πόσο μακριά είναι η κορυφή i . Για την περίπτωση που δεν υπάρχει μονοπάτι μεταξύ δύο κορυφών θέτουμε ως απόσταση το πλήθος όλων των κορυφών (αντί του απείρου). Η ποσότητα $C_i = 1/l_i$, δίνει ένα μέτρο για το πόσο κοντά είναι η κορυφή i και ονομάζεται κεντρικότητα εγγύτητας. Η κεντρικότητα αυτή χρησιμοποιείται πολύ σε κοινωνικά δίκτυα, όμως, δεν μπορεί να κάνει καλό διαχωρισμό και ταξινόμηση των κορυφών, ιδιαίτερα σε μεγάλα δίκτυα, επειδή το εύρος των τιμών είναι πολύ μικρό. Άλλο μειονέκτημα είναι ότι μεταβάλλεται εύκολα με την προσθήκη ή διαγραφή κορυφών.

Έχουν προταθεί και άλλοι ορισμοί, όπως το να χρησιμοποιηθεί ο αρμονικός μέσος των αποστάσεων, ο οποίος δεν έχει πρόβλημα με κορυφές που δεν συνδέονται, δηλαδή με άπειρη απόσταση, αφού το αντίστροφό της είναι 0 και δεν μεταβάλλει την κεντρικότητα.

1.6 Ενδιάμεση κεντρικότητα (betweenness centrality).

Η έννοια της ενδιάμεσης κεντρικότητας έχει να κάνει με το κατά πόσον μια κορυφή βρίσκεται σε μονοπάτια μεταξύ άλλων κορυφών. Προτάθηκε από τον Freeman το 1977, αν και όπως ο ίδιος γράφει είχε ήδη προταθεί από τον Anthonisse σε μία αδημοσίευτη εργασία του.

Αν $n_{st}^i = 1$ όταν η κορυφή i βρίσκεται στη γεωδαισιακή που συνδέει τις κορυφές s και t ενώ είναι 0 σε άλλη περίπτωση, τότε η ενδιάμεση κεντρικότητα ορίζεται ως

$$x = (x_i), \text{ όπου } x_i = \sum_{s,t} n_{st}^i.$$

Ένας καλύτερος ορισμός είναι να θεωρήσουμε ως n_{st}^i το πλήθος των γεωδαισιακών από το s προς το t που περνούν από το i , και, ως g_{st} το πλήθος των γεωδαισιακών από το s προς το t , οπότε η ενδιάμεση κεντρικότητα της κορυφής i γράφεται:

$$x_i = \sum_{s,t} \frac{n_{st}^i}{g_{st}}, \text{ όπου δεχόμαστε ότι } \frac{n_{st}^i}{g_{st}} = 0, \text{ όταν και οι δύο όροι του είναι 0.}$$

2. ΣΥΓΚΡΙΣΗ ΤΩΝ ΜΕΤΡΩΝ ΚΕΝΤΡΙΚΟΤΗΤΑΣ

Είδαμε ότι τα διάφορα μέτρα κεντρικότητας δίνουν διαφορετική πληροφορία για τη σημαντικότητα των κορυφών. Για να συγκρίνουμε τα διάφορα μέτρα πήραμε το πρώτο παράδειγμα με τις 8 κορυφές και υπολογίσαμε την κεντρικότητα για τα 6 μέτρα που περιγράψαμε, χρησιμοποιώντας τις τυποποιημένες μορφές ώστε να είναι συγκρίσιμα. Έτσι προέκυψε ο παρακάτω πίνακας, όπου φαίνεται ότι όλα τα μέτρα στο παράδειγμα αυτό δίνουν σημαντικότερη την κορυφή 5, ενώ οι άλλες κορυφές έχουν διαφορετική διάταξη ως προς τη σημαντικότητα στα διάφορα μέτρα, όπως φαίνεται και στο Σχήμα 5.

Κεντρικότητα	1	2	3	4	5	6	7	8
degree	0.6	0.2	0.4	0.4	1.0	0.6	0.6	0.6
eigenvector	0.269	0.083	0.393	0.393	1.000	0.814	0.814	0.814
Katz	0.471	0.265	0.515	0.515	1.000	0.794	0.794	0.794
PageRank	0.747	0.301	0.471	0.471	1.000	0.600	0.600	0.600
closeness	0.714	0.500	0.833	0.833	1.000	0.714	0.714	0.714
betweenness	0.52	0.00	0.32	0.32	1.00	0.00	0.00	0.00

Στη συνέχεια υπολογίσαμε τη συσχέτιση των 6 αυτών διανυσμάτων και πήραμε τον παρακάτω πίνακα

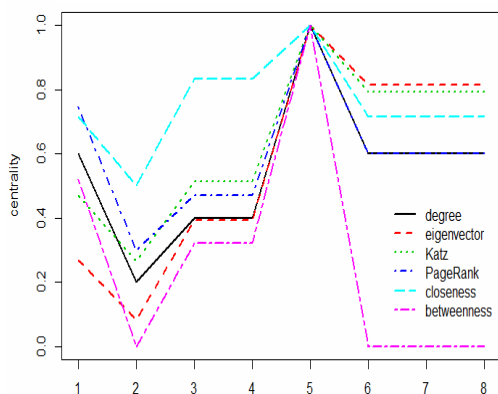
	degree	eigenvector	Katz	PageRank	closeness	betweenness
degree	1.000	0.820	0.885	0.972	0.718	0.653
eigenvector		1.000	0.992	0.666	0.560	0.189
Katz			1.000	0.754	0.634	0.308
PageRank				1.000	0.708	0.762
closeness					1.000	0.777

betweenness						1.000

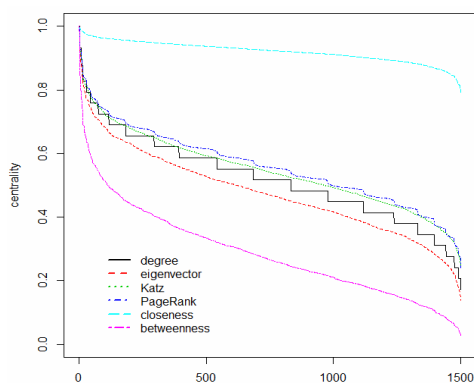
όπου φαίνεται ότι τα τέσσερα πρώτα μέτρα έχουν κάπως μεγαλύτερη σχέση μεταξύ τους και διαφέρουν από τα δύο τελευταία τα οποία διαφέρουν και μεταξύ τους. Ανάλογα συμπεράσματα βγαίνουν και από τα άλλα παραδείγματα.

Τη σύγκριση αυτή την εφαρμόσαμε και στα τυχαία δίκτυα, στα οποία διατάξαμε τα διανύσματα κεντρικότητας σε φθίνουσα σειρά (Σχήμα 6). Διαπιστώθηκε ότι τα

Σχήμα 5: Οι κεντρικότητες για το παράδειγμα 1

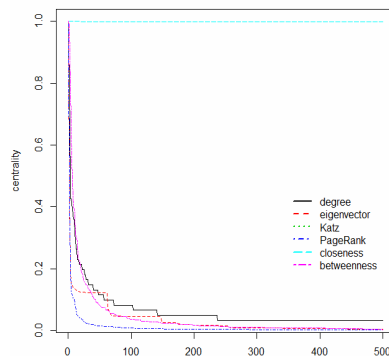


Σχήμα 6: Οι κεντρικότητες σε τυχαία δίκτυα



τέσσερα μέτρα που σχετίζονται με το βαθμό κορυφής έχουν μια παρόμοια κατανομή. Η ενδιαμέση κεντρικότητα διαφέρει κάπως, αλλά είναι ανάλογη με τις υπόλοιπες. Η κεντρικότητα εγγύτητας είναι πολύ διαφορετική και δεν ευνοεί τη διάκριση (όλες οι τιμές περίπου ίδιες), όπως φαίνεται στο Σχήμα 6. Η εύρεση της συσχέτισης μεταξύ των μέτρων έδωσε μεγάλες γραμμικές συσχετίσεις με καλύτερες της βαθμικής με την Page και της ιδιοκεντρικότητας με την Katz.

Σχήμα 7: Οι κεντρικότητες σε αυτο-όμοια δίκτυα



Ανάλογη σύγκριση των μέτρων αυτών έγινε και στα αυτο-όμοια (ανεξάρτητα κλίμακος) δίκτυα (Σχήμα 7) και διαπιστώθηκε ότι η κατανομή για τα 5 από τα 6 μέτρα, εκτός της κεντρικότητας εγγύτητας, είναι παρόμοια, αλλά πολύ διαφορετική, με αυτήν των τυχαίων δικτύων. Εδώ βρέθηκαν μικρότερες συσχετίσεις, με καλύτερη τη σχέση βαθμικής με την Katz.

Αυτό το γεγονός είχε παρατηρηθεί για την κατανομή των βαθμών, εδώ όμως διαπιστώνεται ότι ισχύει και για τα άλλα μέτρα κεντρικότητας.

ABSTRACT

For studying and understanding of the structure of the graphs we need different metrics and distances. Many of these were originally defined in social networks (social networks), and therein have their names, but today they have widespread use in other networks, such as biological networks, computer networks, random networks, etc. In this work a didactic approach to centrality measures (their definition, importance) is attempted, and hints are given for the calculation and the limitations of their applications. A comparison of these measures is attempted in simple and in random networks.

ΑΝΑΦΟΡΕΣ

- Anthonisse, J. M. (1971). The rush in a directed graph, *Technical Report* BN 9/71, Stichting Mathematisch Centrum, Amsterdam.
- Cawkell, A. E. (1971). Science Citation Index: Effectiveness in locating articles in the anaesthetics field: 'Perturbation of ion transport'. *British J. of Anaesthesia*, **43**: 814
- Coleman, J. S., Katz, E., and Menzel, H. (1957). The diffusion of an innovation among physicians, *Sociometry* **20**, 253-270
- Freeman, L. C. (1977). A set of measures of centrality based upon betweenness, *Sociometry* **40**, 35-41.
- Newman Mark (2010). Networks: An Introduction
- Phillip Bonacich (1987) Power and Centrality: A Family of Measures, *American Journal Of Sociology* **92**(5): 1170-1182

ΑΙΤΙΟΤΗΤΑ ΚΑΤΑ GRANGER ΣΕ ΑΠΑΡΙΘΜΗΤΕΣ ΧΡΟΝΟΣΕΙΡΕΣ ΚΑΙ ΕΦΑΡΜΟΓΗ ΣΕ ΙΣΤΟΡΙΚΑ ΔΕΔΟΜΕΝΑ ΠΡΟΣΤΑΣΙΑΣ ΤΗΣ ΒΙΟΠΟΙΚΙΛΟΤΗΤΑΣ

Κ.Γ. Παπασπυρόπουλος^{1,3}, Δ. Κουγιουμτζής²

¹Εργαστήριο Δασικής Οικονομικής, Σχολή Δασολογίας και Φυσικού Περιβάλλοντος,
Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης

²Πολυτεχνική Σχολή, Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης,

³Διεύθυνση Έρευνας και Τεκμηρίωσης, ΣΤ΄ Κυνηγετική Ομοσπονδία Μακεδονίας
Θράκης, Θεσσαλονίκη

kodafype@for.auth.gr, dkugiu@gen.auth.gr

ΠΕΡΙΛΗΨΗ

Η παρούσα έρευνα εξετάζει την εκτίμηση της αιτιότητας κατά Granger σε απαριθμητές (count) χρονοσειρές, συγκρίνοντας δύο προσεγγίσεις. Η πρώτη είναι η κλασική προσέγγιση, που ορίζει τη Granger αιτιότητα με χρήση διανυσματικών αυτοπαλινδρομούμενων μοντέλων (VAR) και εφαρμόζεται σε συνεχή δεδομένα. Εδώ εφαρμόζουμε την ίδια προσέγγιση σε απαριθμητά δεδομένα και την ονομάζουμε DVAR. Η δεύτερη προσέγγιση είναι από το χώρο της βιοστατιστικής, βασίζεται σε πολυμεταβλητά μοντέλα επιτήρησης και εφαρμόζεται σε απαριθμητά δεδομένα (τη συμβολίζουμε MSURV). Ορίζονται οι δείκτες Granger αιτιότητας από τις δύο προσεγγίσεις και εξετάζεται η δυνατότητά τους να εντοπίζουν σωστά την Granger αιτιότητα σε απαριθμητές χρονοσειρές που παράγονται μέσω προσομοίωσης από ένα VAR(1) σύστημα συνεχών δεδομένων. Οι δύο προσεγγίσεις συγκρίνονται τόσο ως προς το βαθμό εύρεσης της αιτιότητας κατά Granger και της υπό συνθήκη αιτιότητας κατά Granger, όσο και ως προς το βαθμό εύρεσης της έντασης των δύο αυτών δεικτών αιτιότητας. Στο τέλος, με βάση αυτά τα αποτελέσματα, γίνεται εφαρμογή σε πραγματικές χρονοσειρές που τηρούνται με τη μορφή απαριθμητών δεδομένων και σχετίζονται με την προστασία της βιοποικιλότητας.

Λέξεις Κλειδιά: απαριθμητές χρονοσειρές, Granger-αιτιότητα, πολυμεταβλητή ανάλυση χρονοσειρών, δασοπονία

1. ΕΙΣΑΓΩΓΗ

Η ανάλυση μονομεταβλητών χρονοσειρών είναι πολύ διαδεδομένη σε ένα ευρύ φάσμα της επιστημονικής βιβλιογραφίας, όμως η ανάλυση πολυμεταβλητών χρονοσειρών είναι αυτή που περιγράφει καλύτερα κάποια φαινόμενα όταν παρατηρούμε ταυτόχρονα πολλές μεταβλητές (Box et al. 2008, Chatfield 2004,

Lütkepohl 2005, Brandt & Williams 2007). Υπάρχουν διάφοροι τρόποι μελέτης των χρονοσειρών στην πολυμεταβλητή ανάλυση με πιο διαδεδομένο αυτόν των Διανυσματικών Αυτοπαλινδρομούμενων Μοντέλων (Vector Autoregressive Models, VAR models) (Sims 1980). Μέσω των VAR μοντέλων ο ερευνητής, παράλληλα, μπορεί να εξετάσει εάν υπάρχει σχέση αιτιότητας μεταξύ των μεταβλητών-χρονοσειρών. Όταν υπάρχει μια χρονοσειρά της οποίας οι παρελθούσες τιμές βελτιώνουν την πρόβλεψη των τιμών μιας άλλης χρονοσειράς, τότε η πρώτη προκαλεί κατά Granger τη δεύτερη χρονοσειρά (Granger 1969). Τα VAR μοντέλα εφαρμόζονται σε συνεχή δεδομένα χρονοσειρών. Τα τελευταία χρόνια έχουν διερευνηθεί από διάφορους ερευνητές ειδικά μοντέλα χρονοσειρών που εφαρμόζονται αποκλειστικά σε απαριθμητά δεδομένα (Kedem & Fokianos 2002, Fokianos 2012, Jung et al. 2006). Το επιστημονικό αυτό πεδίο εξελίσσεται συνεχώς, αλλά οι περισσότερες καινοτομίες του, μέχρι στιγμής, αφορούν τη μελέτη μονομεταβλητών χρονοσειρών απαριθμητών δεδομένων. Παρόλα αυτά, υπάρχουν στην πρόσφατη βιβλιογραφία μοντέλα πολυμεταβλητής ανάλυσης χρονοσειρών απαριθμητών δεδομένων (Pedeli & Karlis 2012, Jung et al. 2011, Held et al. 2005, Heinen & Rengifo 2007, Paul et al. 2008, Paul & Held 2011), ενώ ελάχιστες είναι προς το παρόν οι προσπάθειες για τη διερεύνηση της αιτιότητας κατά Granger σε χρονοσειρές απαριθμητών δεδομένων (Boudreault & Charpentier, 2011).

Λαμβάνοντας τα παραπάνω υπόψη, οι σκοποί της παρούσας έρευνας είναι: α) να συγκρίνει τα Διανυσματικά Αυτοπαλινδρομούμενα Μοντέλα με τα μοντέλα πολυμεταβλητών χρονοσειρών απαριθμητών δεδομένων ως προς την αιτιότητα κατά Granger, β) να μελετήσει πώς συμπεριφέρονται και λειτουργούν τα VAR μοντέλα σε συνάρτηση με το εύρος τιμών των απαριθμητών δεδομένων σχετικά με την εκτίμηση της αιτιότητας κατά Granger, αν δηλαδή καταφέρνουν να εκτιμήσουν την αιτιότητα κατά Granger στον ίδιο βαθμό με το αντίστοιχο συνεχές VAR μοντέλο, και γ) να εφαρμόσει την κατάλληλη μέθοδο σε απαριθμητές χρονοσειρές που σχετίζονται με την προστασία της βιοποικιλότητας.

2. GRANGER ΑΙΤΙΟΤΗΤΑ ΚΑΙ ΑΠΑΡΙΘΜΗΤΕΣ ΧΡΟΝΟΣΕΙΡΕΣ

2.1 Πολυμεταβλητή ανάλυση χρονοσειρών συνεχών δεδομένων

Ένα VAR(p) μοντέλο που προσαρμόζεται σε χρονοσειρές K μεταβλητών μήκους T , $\{y_{k,t}\}$, $k=1,\dots,K$ και $t=1,\dots,T$, περιγράφεται από την ακόλουθη σχέση (Lütkepohl, 2005)

$$y_t = \nu + A_1 y_{t-1} + \dots + A_p y_{t-p} + e_t, \quad t = p+1, \dots, T \quad (1)$$

όπου y_t , e_t και ν είναι διανύσματα διάστασης $K \times 1$ και A_i , $i=1,\dots,p$, είναι πίνακες διάστασης $K \times K$. Το y_t είναι το διάνυσμα των K μεταβλητών τη χρονική στιγμή t και το e_t είναι το διάνυσμα μεταβλητών λευκού θορύβου και ισχύει $E(e_t) = \mathbf{0}$ ($\mathbf{0}$ είναι το μηδενικό $K \times 1$ διάνυσμα), $E(e_t e_s') = \mathbf{0}$ για $s \neq t$ ($\mathbf{0}$ είναι ο μηδενικός $K \times K$ πίνακας) και $E(e_t e_t') = \Sigma_e$, όπου Σ_e είναι ο πίνακας διασπορών συνδιασπορών του e_t .

Η εκτίμηση των παραμέτρων ενός τέτοιου μοντέλου μπορεί να γίνει με τη μέθοδο των ελαχίστων τετραγώνων, καθώς και με τη μέθοδο της μέγιστης πιθανοφάνειας και τον εκτιμητή Yule-Walker (Lütkepohl 2005). Η τάξη p του VAR μοντέλου εκτιμάται με κάποιο κριτήριο τάξης, όπως το κριτήριο πληροφορίας Akaike (AIC) και το Μπεϋζιανό κριτήριο πληροφορίας (BIC) (Lütkepohl 2005, Brandt & Williams 2007). Η επάρκεια του μοντέλου εξετάζεται με ελέγχους αντίστοιχους με αυτούς για τη μονομεταβλητή περίπτωση, όπως τον έλεγχο Portmantaeu των αυτοσυσχετίσεων των υπολοίπων από την προσαρμογή του μοντέλου. Η εκτίμηση της τιμής της μεταβλητής k σε χρόνο t , $y_{k,t}$, δίνεται από τη δεσμευμένη μέση τιμή $\mu_{k,t}$ ως

$$\mu_{k,t} = v_k + \sum_{j=1}^K A_{1,k,j} y_{j,t-1} + \dots + \sum_{j=1}^K A_{p,k,j} y_{j,t-p} + e_{k,t}, \quad t = p+1, \dots, T,$$

όπου v_k είναι η τιμή στη θέση k του διανύσματος $\mathbf{v}$ και $A_{i,k,j}$ είναι η τιμή στη θέση (k,j) του πίνακα A_i .

2.2 Πολυμεταβλητή ανάλυση χρονοσειρών απαριθμητών δεδομένων

Στην επιστήμη της βιοστατιστικής προτάθηκαν από τους Held et al. (2005) πολυμεταβλητά μοντέλα ανάλυσης χρονοσειρών απαριθμητών δεδομένων. Οι Jung et al. (2011) περιγράφουν το μοντέλο των Held et al. (2005) και Paul et al. (2008) ως ένα μοντέλο που ομοιάζει με τα VAR μοντέλα για τις δεσμευμένες μέσες τιμές. Αυτός είναι και ένας από τους σημαντικούς λόγους που οδήγησαν την παρούσα έρευνα στη χρήση του παρακάτω μοντέλου για σύγκριση με τα VAR μοντέλα. Σύμφωνα με τους Paul et al. (2008), η απλή μορφή του προτεινόμενου μοντέλου είναι η εξής

$$\mu_{k,t} = \lambda y_{k,t-l} + \exp(\eta_{k,t}) \quad (2)$$

Το παρατηρούμενο σύστημα μπορεί να χωριστεί σε δύο μέρη. Το πρώτο μέρος είναι το επιδημικό μέρος, όπου λ είναι η άγνωστη αυτοπαλινδρομούμενη παράμετρος και l η τιμή της υστέρησης, και το ενδημικό μέρος $v_{k,t} = \exp(\eta_{k,t})$, όπου $\eta_{k,t}$ λευκός θόρυβος. Το επιδημικό μέρος είναι συνήθως ικανό να εξηγήσει τις μεταβολές και τα ξεσπάσματα του συστήματος, ενώ το ενδημικό μέρος αποτελεί τη «σταθερά», δηλαδή ένα σταθερό χρονικό αριθμό εμφανίσεων του γεγονότος. Το μοντέλο αυτό μπορεί να εφαρμοστεί τόσο σε χρονοσειρές που ακολουθούν αρνητική διωνυμική κατανομή, όσο και σε χρονοσειρές απαριθμητών δεδομένων Poisson κατανομής. Το επιδημικό μέρος μπορεί να επεκταθεί και στην ύπαρξη μιας ξεχωριστής παραμέτρου λ_k για κάθε χρονοσειρά, καθώς και στην συμπερίληψη των παρελθουσών τιμών των άλλων χρονοσειρών του συστήματος, είτε με μια ξεχωριστή παράμετρο ϕ_k , ή με μια κοινή παράμετρο ϕ , οπότε η (2) γίνεται (Paul et al., 2008)

$$\mu_{k,t} = \lambda y_{k,t-l} + \phi_k \sum_{j=1, j \neq k}^K w_{jk} y_{j,t-l} + v_{k,t} \quad (3)$$

Τα βάρη w_{jk} υπονοούν σχέση μεταξύ των χρονοσειρών και στην πιο απλή περίπτωση παίρνουν την τιμή 1 για όλα τα $j \neq k$, ενώ μπορούν να πάρουν και άλλες τιμές (Held & Paul, 2012). Η παράμετρος ϕ_k ποσοτικοποιεί την επίδραση των παρατηρήσεων των

άλλων χρονοσειρών στη δεσμευμένη τιμή της μιας χρονοσειράς και είναι αυτή η παράμετρος που μπορεί να αποκαλύψει σχέσεις αιτιότητας μεταξύ των χρονοσειρών. Το ενδημικό μέρος μπορεί να εκφραστεί με τον ακόλουθο τρόπο για να συμπεριλάβει τάσεις και εποχικότητα του φαινομένου σε κάθε χρονοσειρά (Paul et al., 2008)

$$\log(v_{k,t}) = \eta_{k,t} = \alpha_k + \beta_k t + \sum_{s=1}^{S_k} (\gamma_{k,t} \sin(\omega_s t) + \delta_{k,t} \cos(\omega_s t))$$

όπου S_k είναι ο αριθμός των αρμονικών για τη k χρονοσειρά (μεταβλητή), ω_s οι Fourier συχνότητες, α_k ο σταθερός όρος και β_k ο συντελεστής τάσης για την k χρονοσειρά. Οι παράμετροι του παραπάνω μοντέλου μπορούν να εκτιμηθούν με μεθόδους μέγιστης πιθανοφάνειας όταν $\lambda = 0$ και υπάρχει ίδια μεγάλη διακύμανση σε όλες τις χρονοσειρές, ενώ όταν αυτό δεν ισχύει, γίνεται αριθμητική βελτιστοποίηση της λογαριθμο-πιθανοφάνειας (log-likelihood) (Paul et al., 2008).

2.3 Granger αιτιότητα

Η αιτιότητα κατά Granger έχει γνωρίσει μεγάλο ενδιαφέρον τα τελευταία έτη και ενώ αρχικά αναπτύχθηκε στο πλαίσιο της πρόβλεψης με διανυσματικά γραμμικά αυτοπαλινδρομούμενα μοντέλα VAR(p) στη συνέχεια έγιναν επεκτάσεις σε μη-γραμμικά μοντέλα και προτάθηκαν επίσης και άλλα μέτρα αιτιότητας κατά Granger, π.χ. βασισμένα σε μέτρα πληροφορίας (Parana et al, 2013). Εδώ επικεντρωνόμαστε στην κλασική προσέγγιση για τη διμεταβλητή αιτιότητα και την υπό συνθήκη αιτιότητα κατά Granger. Η εκτίμηση της γραμμικής διμεταβλητής αιτιότητας, όπου έστω οι δύο μεταβλητές X και Y και η χρονοσειρά $\{x_t, y_t\}_{t=1}^T$, επιτυγχάνεται μέσω της ανάλυσης δύο αυτοπαλινδρομούμενων μοντέλων. Για να διαπιστωθεί εάν η X οδηγεί τη Y εξετάζεται πρώτα το αυτοπαλινδρομούμενο μοντέλο

$$y_t = \sum_{i=1}^p a_i y_{t-i} + e_{1,t} \quad (4)$$

όπου p ο μέγιστος αριθμός υστερήσεων που συμπεριλαμβάνονται στο αυτοπαλινδρομούμενο (AR) μοντέλο, a_i οι συντελεστές της εξίσωσης και $e_{1,t}$ τα σφάλματα της πρόβλεψης. Το μοντέλο αυτό ονομάζεται περιορισμένο (restricted) γιατί συμπεριλαμβάνει μόνο τη μια από τις δύο χρονοσειρές του συστήματος. Στη συνέχεια εξετάζεται το λεγόμενο μη-περιορισμένο (unrestricted) μοντέλο στο οποίο συμπεριλαμβάνονται και οι παρελθούσες τιμές της δεύτερης χρονοσειράς του συστήματος, δηλαδή

$$y_t = \sum_{i=1}^p a_i y_{t-i} + \sum_{i=1}^p b_i x_{t-i} + e_{2,t} \quad (5)$$

όπου $e_{2,t}$ τα σφάλματα πρόβλεψης του δεύτερου μοντέλου. Αυτό είναι το VAR(p) μοντέλο στην (1) για δύο μεταβλητές. Εάν η διακύμανση του $e_{1,t}$ μειώνεται με την ενσωμάτωση της χρονοσειράς της X στο μοντέλο, τότε η X Granger-οδηγεί (G-causes) τη Y . Μπορεί μάλιστα να εκτιμηθεί και ο δείκτης της αιτιότητας κατά Granger (Granger Causality Index, GCI) (Geweke 1982) που ορίζεται ως

$$GCI_{X \rightarrow Y} = \ln \frac{\text{Var}(\hat{e}_{1,t})}{\text{Var}(\hat{e}_{2,t})} \quad (6)$$

όπου $\hat{e}_{1,t}$ και $\hat{e}_{2,t}$ τα εκτιμώμενα σφάλματα από την προσαρμογή του περιορισμένου μοντέλου στην (4) και του μη-περιορισμένου μοντέλου στην (5) αντίστοιχα. Ο δείκτης αυτός μηδενίζεται όταν το σφάλμα πρόβλεψης είναι το ίδιο και στα δύο μοντέλα, ενώ όσο μεγαλύτερος του μηδενός είναι, τόσο εντονότερη είναι η σχέση αιτιότητας από τη X στη Y . Αντίστοιχα με τη διμεταβλητή περίπτωση, εκτιμάται και η υπό συνθήκη αιτιότητα κατά Granger (Conditional Granger Causality Index, CGCI) (Geweke et al 1982) από τη X στη Y δεδομένου των άλλων παρατηρούμενων μεταβλητών που δηλώνονται συλλογικά με τη διανυσματική μεταβλητή Z . Ο δείκτης CGCI εκτιμάται όπως και ο GCI ως

$$CGCI_{X \rightarrow Y} = \ln \frac{\text{Var}(\hat{e}'_{1,t})}{\text{Var}(\hat{e}'_{2,t})} \quad (7)$$

όπου $\hat{e}'_{1,t}$ και $\hat{e}'_{2,t}$ τα εκτιμώμενα σφάλματα από την προσαρμογή των μοντέλων στην (4) και (5), αντίστοιχα, προσαυξημένα όμως με τους όρους υστερήσεων των υπολοίπων μεταβλητών στη Z . Συνολικά σε ένα σύστημα K χρονοσειρών μπορούν να εκτιμηθούν $K(K-1)$ δείκτες CGCI. Σε κάθε σύστημα, φυσικά, πρέπει πρώτα να έχει εκτιμηθεί η τάξη p του VAR μοντέλου, π.χ. με το κριτήριο AIC ή BIC. Γενικά οι σχέσεις αιτιότητας μπορούν να διερευνηθούν από τους συντελεστές του VAR μοντέλου. Έτσι για τη στατιστική σημαντικότητα του δείκτη $GCI_{X \rightarrow Y}$ και $CGCI_{X \rightarrow Y|Z}$, δηλαδή αν η X προκαλεί ή δεν προκαλεί κατά Granger την Y διερευνούμε αν υστερήσεις της X δεν εμφανίζονται στην εξίσωση για το μη-περιορισμένο μοντέλο της Y , δηλαδή ελέγχουμε τη μηδενική υπόθεση $H_0: b_i=0$ για όλες τις υστερήσεις $i=1, \dots, p$. Ο έλεγχος γίνεται με το λεγόμενο F -στατιστικό και αποτελεί τον παραμετρικό έλεγχο σημαντικότητας των δεικτών GCI και CGCI (Brandt & Williams 2007). Για K μεταβλητές έχουμε $K(K-1)$ πολλαπλούς ελέγχους και για αυτό χρησιμοποιείται επίσης και η διόρθωση των ψευδούς ποσοστού ανακαλύψεων (False Discovery Rate, FDR) (Benjamini & Hochberg, 1995). Η ίδια διαδικασία υπολογισμού και ελέγχου για τα GCI και CGCI εφαρμόστηκε για το περιορισμένο και μη-περιορισμένο μοντέλο επιτήρησης της (3).

3. ΣΧΕΔΙΑΣΜΟΣ ΠΡΟΣΟΜΟΙΩΤΙΚΗΣ ΜΕΛΕΤΗΣ & ΕΦΑΡΜΟΓΗΣ

Αρχικά έγινε προσομοιωτική προσέγγιση στο ζητούμενο ερώτημα κάνοντας χρήση κατά κύριο λόγο του υπολογιστικού περιβάλλοντος Matlab. Δημιουργήθηκαν χρονοσειρές συνεχών δεδομένων από ένα διανυσματικό αυτοπαλινδρομούμενο σύστημα τάξης ένα, VAR(1). Επειδή οι σχέσεις που προέρχονται από τα VAR(1) συστήματα είναι γνωστές, θεωρήθηκε ως βάση σύγκρισης η αιτιότητα κατά Granger που αποκαλυπτόταν από αυτά τα συστήματα. Στις πραγματοποιήσεις του VAR(1) εξετάστηκε η αιτιότητα κατά Granger από τους δείκτες GCI και CGCI στην (6) και (7), η σημαντικότητα της οποίας ελέγχθηκε μέσω του F -ελέγχου και της διόρθωσης

FDR για επίπεδο σημαντικότητας $\alpha = 0,05$. Η προσέγγιση αυτή συμβολίζεται ως VAR. Στη συνέχεια οι ίδιες πραγματοποιήσεις διακριτοποιούνταν, με ελάχιστη τιμή το 0, και εξεταζόταν στις νέες διακριτές χρονοσειρές η αιτιότητα κατά Granger. Η προσέγγιση αυτή συμβολίζεται ως DVAR. Τέλος στις ίδιες διακριτοποιημένες πραγματοποιήσεις γινόταν προσαρμογή του περιορισμένου και μη-περιορισμένου πολυμεταβλητού μοντέλου επιτήρησης και υπολογίζονταν οι δείκτες GCI και CGCI καθώς και στατιστική τους σημαντικότητα. Αυτή η προσέγγιση συμβολίζεται ως MSURV. Σημειώνεται πως τα μοντέλα επιτήρησης που χρησιμοποιήθηκαν στην εργασία είναι από το πακέτο Surveillance (Höhle 2007, Höhle et al. 2012) στη γλώσσα R (R Core Team 2012). Η επιλογή των μοντέλων για VAR, DVAR και MSURV γινόταν με το AIC. Συνολικά δημιουργούνταν κάθε φορά 100 πραγματοποιήσεις χρονοσειρών από σύστημα VAR(1) με πίνακα A

$$A = \begin{bmatrix} 0,7 & 0,3 & 0,2 \\ 0 & 0,2 & 0,5 \\ 0 & 0,4 & 0,6 \end{bmatrix}$$

Η διαδικασία της προσομοίωσης πραγματοποιήθηκε συνολικά δέκα φορές για τους συνδυασμούς δύο μηκών χρονοσειρών και 5 τιμών τυπικής απόκλισης (SD) του λευκού θορύβου εισόδου (δες Πίνακα 1). Η μεταβολή του SD έχει ως αποτέλεσμα τη μεταβολή του εύρους τιμών της διακριτής χρονοσειράς μετά από στρογγυλοποίηση των συνεχών τιμών της χρονοσειράς σε ακέραιο μέρος, όπως δίνεται στον Πίνακα 1. Στόχος αυτών των αλλαγών ήταν να διαπιστωθεί η απόδοση των προσεγγίσεων DVAR και MSURV στα διάφορα εύρη τιμών των απαριθμητών δεδομένων, καθώς και αν αυτή επηρεάζεται από το μήκος της χρονοσειράς.

Πίνακας 1. Εύρη διακριτών τιμών ανάλογα με το θόρυβο και το μήκος χρονοσειράς

Θόρυβος (SD)	SD = 0,2	SD = 0,5	SD = 0,8	SD = 1,2	SD = 2,5
Μήκος $T = 100$	[0 4]	[0 7]	[0 11]	[0 16]	[0 35]
Μήκος $T = 1000$	[0 5]	[0 8]	[0 13]	[0 19]	[0 40]

Με βάση τα ευρήματα της προσομοίωσης, έγινε εφαρμογή της πιο κατάλληλης μεθοδολογίας για να εφαρμοστεί σε πραγματικά δεδομένα. Αντιστοιχίστηκε δηλαδή το εύρος τιμών των πραγματικών δεδομένων με μια από τις 5 επαναλήψεις της προσομοίωσης και εφαρμόστηκε η κατάλληλη μέθοδος επεξεργασίας χρονοσειρών απαριθμητών δεδομένων για την εκτίμηση της αιτιότητας κατά Granger. Επιλέχθηκαν χρονοσειρές από την επιστήμη της δασολογίας και του φυσικού περιβάλλοντος, οι οποίες σχετίζονται με την προστασία της βιοποικιλότητας και έχουν απαριθμητές παρατηρήσεις. Τα δεδομένα προέρχονται από το αρχείο της ΣΤ΄ Κυνηγετικής Ομοσπονδίας Μακεδονίας Θράκης και σχετίζονται με το έργο της Ομοσπονδιακής Θηροφυλακής στην πρόληψη και καταστολή της λαθροθηρίας και την εν γένει προστασία του φυσικού περιβάλλοντος (PapaspYROPOULOS et al., 2012). Στην παρούσα έρευνα ελήφθησαν τα δεδομένα που αφορούν τις μηνύσεις που υποβάλλουν συνολικά οι Ομοσπονδιακοί Θηροφύλακες στις τρεις περιφέρειες της περιοχής ευθύνης: Δυτική Μακεδονία, Κεντρική Μακεδονία και Ανατολική

Μακεδονία-Θράκη. Συγκεκριμένα καταγράφηκε ο συνολικός αριθμός μηνύσεων ανά μήνα και περιφέρεια από τον Μάρτιο του 2000 έως τον Οκτώβριο του 2012. Πρόκειται, δηλαδή, για ένα σύστημα 3 χρονοσειρών (μία ανά Περιφέρεια) με 152 παρατηρήσεις η κάθε μια (152 μήνες). Οι μηνύσεις που καταθέτουν οι Ομοσπονδιακοί Θηροφύλακες μπορεί να περιλαμβάνουν ένα ή περισσότερα παραπτώματα, έναν ή περισσότερους μηνυόμενους, ή να είναι κατά αγνώστων. Τα παραπτώματα μπορεί να είναι προληπτικά (για παράδειγμα μεταφορά όπλου εκτός θήκης εντός καταφυγίου άγριας ζωής), ή κατασταλτικά (θήρα ζαρκαδιού).

4. ΑΠΟΤΕΛΕΣΜΑΤΑ - ΣΥΖΗΤΗΣΗ

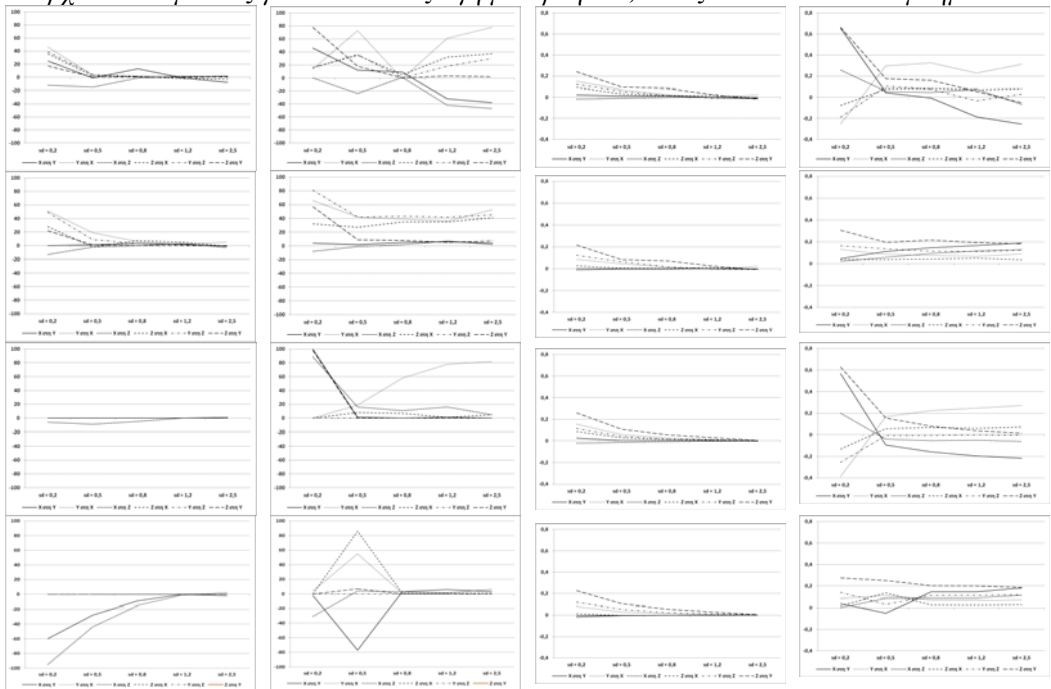
4.1 Αποτελέσματα προσομοιώσεων

Λόγω των τιμών $A_{21}=0$ και $A_{31}=0$ του πίνακα A αναμένεται εκ των προτέρων η X να μην προκαλεί κατά Granger την Y και τη Z αντίστοιχα. Οι υπόλοιπες περιπτώσεις αναμένεται να εμφανίζουν κάποια αιτιότητα. Παρακάτω παρουσιάζονται τα αποτελέσματα ως συγκρίσεις ανά δύο των μεθόδων (VAR με DVAR και VAR με MSURV) τόσο για μήκος 100 όσο και 1000 για όλα τα εύρη τιμών του Πίνακα 1. Η σύγκριση αφορά αρχικά το αν εντοπίζουν σωστά τη σχέση αιτιότητας στα ζεύγη μεταβλητών. Εξετάζεται το ποσοστό απορρίψεων για τον έλεγχο σημαντικότητας του GCI και CGCI για κάθε μια από τις προσεγγίσεις VAR, DVAR και MSURV. Η Εικόνα 1 παρουσιάζει τις διαφορές των απορρίψεων ως συνάρτηση του εύρους τιμών για κάθε ένα από τα 6 ζεύγη μεταβλητών, και για κάθε δείκτη και μήκος χρονοσειρών. Αρνητικές τιμές σημαίνει ότι η DVAR ή η MSURV απορρίπτουν περισσότερες φορές την H_0 , βρίσκουν δηλαδή σε μεγαλύτερο ποσοστό αιτιότητα, ενώ θετικές τιμές το αντίστροφο. Όσον αφορά την DVAR μέθοδο για $T=100$, τόσο το ποσοστό απορρίψεων μέσω του δείκτη GCI, όσο και του CGCI φαίνεται να συγκλίνει σε εύρος τιμών $[0\ 13]$. Αυτή η σύγκλιση εμφανίζεται και στα ζεύγη μεταβλητών που είναι γνωστό εξ αρχής ότι δεν υπάρχει αιτιότητα. Οι δύο μέθοδοι διαπράττουν σφάλμα τύπου I όσον αφορά το δείκτη GCI. Στο δείκτη CGCI η DVAR μέθοδος βρίσκει τη μη αιτιότητα μόνο σε εύρος τιμών $[0\ 19]$. Όσον αφορά την MSURV μέθοδο σύγκλιση υπάρχει μόνο ως εύρος τιμών $[0,13]$ σε όλες τις περιπτώσεις. Εξάιρεση αποτελεί ο δείκτης CGCI για μήκος χρονοσειρών που δίνει σύγκλιση και σε μεγαλύτερα εύρη τιμών. Αξιοσημείωτο είναι ότι η MSURV μέθοδος βρίσκει καλύτερα τις σχέσεις μη αιτιότητας σε μικρό θόρυβο σε σχέση με τη VAR μέθοδο.

Στην Εικόνα 1, επίσης, παρουσιάζεται η σύγκριση όσον αφορά την ένταση της σχέσης αιτιότητας. Εξετάζεται δηλαδή πέρα από την επιτυχή αναγνώριση της σχέσης αιτιότητας, κατά πόσο εκτιμάται το ίδιο και η ένταση αυτής. Διαπιστώνεται ότι η DVAR δίνει καλύτερη εκτίμηση από την MSURV. Για το μέγιστο εύρος τιμών η προσέγγιση DVAR καταφέρνει να εκτιμήσει την ίδια ένταση με το συνεχές αντίστοιχο σε όλες τις περιπτώσεις εκτίμησης. Για μικρό εύρος τιμών παρουσιάζουν αποκλίσεις από $-0,02$ έως $0,26$ μονάδες περίπου, ενώ για μεγάλο εύρος τιμών με 1000 παρατηρήσεις οι τιμές συγκλίνουν απολύτως, ενώ στις 100 παρατηρήσεις υπάρχει μια μικρή απόκλιση μόνο για ένα ζεύγος μεταβλητών. Αντίθετα η MSURV δεν συγκλίνει

σχεδόν ποτέ. Για την περίπτωση των 100 παρατηρήσεων και για σχετικά μικρό εύρος τιμών υπάρχει για κάποιες σχέσεις σφάλμα κοντά στο μηδέν, αλλά στη συνέχεια αυτό μεγαλώνει όσο μεγαλώνει και το εύρος τιμών. Πάντως, παρόλο που τα μοντέλα αυτά εκτιμούν καλύτερα την ύπαρξη αιτιότητας σε πολύ μικρό εύρος τιμών και για 100 παρατηρήσεις, στις περιπτώσεις αυτές η ένταση της αιτιότητας αποκλίνει αρκετά σε σύγκριση με το συνεχές VAR σύστημα. Με εξαίρεση ενός ζεύγους μεταβλητών, όπου αυτή η απόκλιση είναι περίπου 0,05 μονάδες, στις υπόλοιπες, η απόκλιση είναι μεγαλύτερη των 0,2 μονάδων. Τέλος, σύγκλιση για ζεύγη μεταβλητών παρατηρείται για το δείκτη GCI, αλλά για μικρός εύρος τιμών και 1000 παρατηρήσεις.

Εικόνα 1. Σύγκριση μεθόδων μέσω των διαφορών των απορρίψεων της H_0 (πρώτη-δεύτερη στήλη) και μέσω των διαφορών των εκτιμήσεων των δεικτών αιτιότητας (τρίτη-τέταρτη στήλη). Πρώτη-τρίτη στήλη: Σύγκριση VAR - DVAR. Δεύτερη-τέταρτη στήλη: Σύγκριση VAR - MSURV. Πρώτη γραμμή: GCI, $T=100$. Δεύτερη γραμμή: CGCI, $T=100$. Τρίτη γραμμή: GCI, $T=1000$. Τέταρτη γραμμή: CGCI, $T=1000$. Σε κάθε σχήμα υπάρχουν 6 καμπύλες για τα όλα τα ζεύγη μεταβλητών, όπως δίνονται στο υπόμνημα.



4.2 Αποτελέσματα εφαρμογής

Με βάση τα αποτελέσματα της προσομοίωσης, διαφάνηκε ότι στις περισσότερες περιπτώσεις, εκτός μερικών εξαιρέσεων, τα DVAR μοντέλα λειτουργούν καλά και μάλιστα καλύτερα από τα πολυμεταβλητά μοντέλα επιτήρησης, τα οποία έχουν κατασκευαστεί για απεριθμητά δεδομένα. Οι τιμές των παρατηρήσεων στην εφαρμογή κυμαίνονται μεταξύ 0 και 60 με τις περισσότερες παρατηρήσεις να βρίσκονται στο διάστημα 0 έως 20. Για αυτό το εύρος τιμών η προσέγγιση DVAR

βρέθηκε στην προσομοιωτική μελέτη να είναι καλύτερη για την εκτίμηση της αιτιότητας κατά Granger και για αυτό προτιμήθηκε. Επίσης στην προσέγγιση DVAR μπορούμε να επιλέξουμε $p > 1$, ενώ η υλοποίηση της MSURV στη γλώσσα R που επιλέξαμε δεν το επέτρεπε.

Με X συμβολίζεται η χρονοσειρά της Δ. Μακεδονίας, με Y της Κεντρικής και με Z της Α. Μακεδονίας-Θράκης. Στον Πίνακα 2 παρουσιάζονται τα αποτελέσματα για το δείκτη της διμεταβλητής αιτιότητας κατά Granger (GCI). Βρέθηκαν τρεις στατιστικά σημαντικές σχέσεις διμεταβλητής αιτιότητας, από τις οποίες η μία είναι οριακή και δεν είναι στατιστικά σημαντική μετά τη διόρθωση. Οι σχέσεις που αναδεικνύονται με το δείκτη GCI είναι ότι η παραβατικότητα στην Κ. Μακεδονία προκαλεί κατά Granger την παραβατικότητα στη Δ. Μακεδονία, η παραβατικότητα στην Α. Μακεδονία και Θράκη προκαλεί κατά Granger την παραβατικότητα στην Κ. Μακεδονία, ενώ το αντίστροφο συμβαίνει οριακά. Στον Πίνακα 2 παρουσιάζονται τα αποτελέσματα της υπό συνθήκη αιτιότητας κατά Granger (CGCI). Ο δείκτης αυτός επιβεβαιώνει δύο σχέσεις σύμφωνα με τον έλεγχο σημαντικότητας και μία μετά τη διόρθωση FDR. Οι σχέσεις είναι αυτές που βρέθηκαν και στη διμεταβλητή αιτιότητα.

Πίνακας 2. Δείκτες αιτιότητας για το VAR(12) διακριτό σύστημα

	Σχέση	Τιμή GCI	Τιμή p ελέγχου	Τιμή ελέγχου FDR	Μέση τιμή CGCI	Τιμή p ελέγχου	Τιμή ελέγχου FDR
1	$X \rightarrow Y$	0,1576	0,0873	0	0,1613	0,1298	0
2	$Y \rightarrow X$	0,3041	0,0002*	1	0,2549	0,0060*	1
3	$X \rightarrow Z$	0,0893	0,5465	0	0,0907	0,6266	0
4	$Z \rightarrow X$	0,1717	0,0540	0	0,1226	0,3443	0
5	$Y \rightarrow Z$	0,1754	0,0475*	0	0,1768	0,0826	0
6	$Z \rightarrow Y$	0,2223	0,0081*	1	0,2260	0,0169*	0

* στατιστικά σημαντική σχέση αιτιότητας στο επίπεδο σημαντικότητας $\alpha = 0,05$

Το αποτέλεσμα από την εφαρμογή των μοντέλων VAR για την εκτίμηση της αιτιότητας κατά Granger στα δεδομένα της Ομοσπονδιακής Θηροφυλακής της ΣΤ' ΚΟΜΑΘ είναι πολύ σημαντικά από διαχειριστική άποψη για την προστασία της βιοποικιλότητας. Η γνώση της κατεύθυνσης της παραβατικότητας μέσα στο έτος μπορεί να δώσει τη δυνατότητα για καλύτερη οργάνωση των περιπολιών και καλύτερη οργάνωση του προσωπικού στις υπόψη περιοχές για την πρόληψη των παράνομων πράξεων, χωρίς να χρειαστεί η καταστολή τους. Η πρόληψη στην περιβαλλοντική διαχείριση είναι μια πολύ σημαντική αρχή, διότι η τέλεση της παράνομης περιβαλλοντικής πράξης συχνά δεν είναι αντιστρέψιμη (για παράδειγμα παράνομη υλοτομία, παράνομη εκχέρσωση, παράνομη θήρα). Η γνώση του ότι η πρόβλεψη των περιστατικών παράβασης του νόμου σε μια περιοχή μπορεί να βελτιωθεί από ιστορικά δεδομένα άλλης περιοχής κάνει επιβεβλημένη την τήρηση των ιστορικών στοιχείων, ενώ δίνει τη δυνατότητα να επιτευχθούν καλύτερες προβλέψεις, ώστε να οργανωθεί πιο σωστά το σύστημα της Ομοσπονδιακής Θηροφυλακής. Εργαλεία σαν την αιτιότητα κατά Granger αποδεικνύονται πολύτιμα

στο σύγχρονο διαχειριστή του περιβάλλοντος, ο οποίος έχει σκοπό τη διαχείριση και προστασία της βιοποικιλότητας με αειφορικό τρόπο.

5. ΣΥΜΠΕΡΑΣΜΑΤΑ

Η ανάλυση χρονοσειρών απαριθμητών δεδομένων είναι ένας κλάδος με ιδιαίτερη άνθιση τα τελευταία χρόνια. Υπάρχουν πολλά μοντέλα μονομεταβλητής ανάλυσης χρονοσειρών απαριθμητών δεδομένων, αλλά πολύ λίγα πολυμεταβλητής ανάλυσης, που αναπτύχθηκαν όλα πρόσφατα. Στη βιβλιογραφία δε φαίνεται να έχει μελετηθεί το ζήτημα της αιτιότητας κατά Granger σε χρονοσειρές απαριθμητών δεδομένων.

Η προσομοιωτική μελέτη σε ένα σύστημα VAR(1) έδειξε πως τα VAR μοντέλα σε απαριθμητές χρονοσειρές (προσέγγιση DVAR) φαίνεται να αποκαλύπτουν κατά τον ίδιο τρόπο με τα VAR μοντέλα συνεχών δεδομένων (προσέγγιση VAR) τις σχέσεις αιτιότητας όταν το εύρος των τιμών είναι πάνω από το επίπεδο 10 τιμών. Όταν το μήκος των χρονοσειρών είναι μεγάλο, οι δύο προσεγγίσεις VAR και DVAR συγκλίνουν σε όλα τα εύρη τιμών τιμών για την περίπτωση της ύπαρξης αιτιότητας, ενώ στη διμεταβλητή αιτιότητα (δείκτης GCI) και οι δύο διαπράττουν σφάλμα τύπου I. Το σφάλμα αυτό διορθώνεται στην υπό συνθήκη αιτιότητα (δείκτης CGCI) για τη VAR μέθοδο, αλλά η DVAR μέθοδος το διορθώνει από εύρος τιμών [0 19] και μετά.

Τα πολυμεταβλητά μοντέλα επιτήρησης (MSURV μέθοδος), τα οποία είναι μοντέλα που έχουν αναπτυχθεί στον κλάδο της βιοστατιστικής ειδικά για την επεξεργασία πολυμεταβλητών χρονοσειρών απαριθμητών δεδομένων, συγκλίνουν ικανοποιητικά με τη VAR μέθοδο μόνο όταν το μήκος των χρονοσειρών είναι μεγάλο και εκτιμάται η διμεταβλητή αιτιότητα κατά Granger σε εύρος τιμών το πολύ [0 9] και η υπό συνθήκη αιτιότητα κατά Granger σε εύρος τιμών το πολύ [0 16]. Η MSURV μέθοδος βρίσκει καλύτερα τις σχέσεις μη αιτιότητας σε τιμές [0 4] σε σχέση με τη DVAR μέθοδο. Όσον αφορά την ένταση των σχέσεων αιτιότητας, οι VAR και DVAR παρουσιάζουν αποκλίσεις για μικρά εύρη τιμών και συγκλίνουν με την αύξηση του εύρους. Η MSURV όμως δεν καταφέρνει να διορθώσει αυτό το πρόβλημα. Μόνο στην περίπτωση του δείκτη CGCI, για μήκος χρονοσειράς 1000 και για πολύ μικρά εύρη τιμών οι αποκλίσεις βρέθηκαν να είναι μικρές, όχι όμως μηδενικές σε σχέση με τις τιμές του συνεχούς VAR συστήματος.

Τα αποτελέσματα αυτά υποδηλώνουν πως χρειάζεται περαιτέρω διερεύνηση σε μεθόδους πολυμεταβλητής ανάλυσης χρονοσειρών απαριθμητών δεδομένων για να μπορεί να προβλεφθεί σωστά τόσο η ύπαρξη όσο και ο βαθμός της έντασης της αιτιότητας κατά Granger.

Η χρήση της αιτιότητας κατά Granger σε ιστορικά δεδομένα προστασίας της βιοποικιλότητας μπορεί να αποδειχθεί πολύ χρήσιμο εργαλείο στους διαχειριστές του περιβάλλοντος για την πρόληψη παράνομων περιστατικών περιβαλλοντικής υποβάθμισης και διαχείρισης και προστασίας της άγριας ζωής.

ABSTRACT

The work is about Granger causality for time series of counts, a topic that has been little studied so far. Two methodologies are examined: the classic one of the Vector

Autoregressive Models (VAR), designed for continuous-valued time series but applied here also to count time series, denoted DVAR, and the modern one, which originates from biostatistics and specifically the multivariate surveillance methodology, designed for count data. The two methodologies are examined as to their ability to estimate Granger causality in multivariate count time series, which are derived by discretization of time series generated by a VAR(1) system. They are compared first, on their ability to detect the true causality relationships and, second, to estimate the strength of these relationships. It is shown that VAR models for counts (DVAR method) detect the Granger causality in the same way as the VAR models for continuous data, when the time series of counts is long and the range of counts is relatively large. When the range of counts is small, say $[0, 2]$, DVAR models do not perform well either in identifying the causal relationship or in estimating its strength. In such a small range of counts, the multivariate surveillance method (MSURV) identifies well the non-causal relationships, but cannot estimate adequately the strength of all the causal relationships. It turns out that there is need for improving the methods of multivariate time series of counts for the estimation of Granger causality, especially for small values of counts. Based on these results, the studied methodology is applied to real time series of counts from biodiversity conservation. It is proved that Granger causality can be a useful tool to natural environment managers for the prevention of degradation and the management and conservation of wildlife.

Ευχαριστίες: Ο πρώτος συγγραφέας ευχαριστεί τη ΣΤ΄ Κυνηγετική Ομοσπονδία Μακεδονίας Θράκης για την παραχώρηση των χρονοσειρών

ΑΝΑΦΟΡΕΣ

- 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 (Methodology)*, 57: 289–300.
- Boudreault, M. and Charpentier, A. (2011). *Multivariate integer-valued autoregressive models applied to earthquake counts*. Quantact Research Group. (available at <http://arxiv.org/pdf/1112.0929v1.pdf>, accessed on Oct 6th, 2012)
- Box, G.E.P., Jenkins, G.M. and Reinsel, G.C. (2008). *Time Series Analysis: Forecasting and Control*. 4th Edition. Wiley.
- Brandt P.T. and Williams J.T. (2007). *Multiple Time Series Models*, Sage Publications.
- Chatfield, C. (2004). *The Analysis of Time Series: An Introduction*. 6th Edition. Chapman & Hall/CRC Texts in Statistical Science.
- Fokianos, K. (2012). *Count Time Series Models*, in: Tata Subba Rao, Suhasini Subba Rao and C.R. Rao, Editors, *Handbook of Statistics 30*: 315-347.
- Geweke, J. (1982) Measurement of linear dependence and feedback between multiple time series. *Journal of the American Statistical Association*, 77: 304-313.
- Granger, C.W.J. (1969). Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *Econometrica* 37(3): 424-438.

- Heinen, A. and Rengifo, E. (2007). Multivariate autoregressive modelling of time series count data using copulas. *Journal of Empirical Finance*, 14:564-583.
- Held, L. and Paul, M. (2012). Modeling seasonality in space-time infectious disease surveillance data. *Biometrical Journal* 54(6): 824-843.
- Held, L., Höhle, M. and Hofmann, M. (2005). A statistical framework for the analysis of multivariate infectious disease surveillance counts. *Statistical Modelling*, 5: 187-199.
- Höhle, M. (2007). Surveillance: An R package for the monitoring of infectious diseases. *Computational Statistics*, 22(4): 571-582.
- Höhle, M., Meyer, S. and Paul, M. with contributions from T. Correa, M. Hofmann, C. Lang, A. Riebler, D. Sabanés Bové, S. Steiner, M. Virtanen and V. Wimmer (2012). *Package 'surveillance': Temporal and Spatio-Temporal Modeling and Monitoring of Epidemic Phenomena. Version 1.4-2*. Date 2012-08-17.
- Jung, R.C., Kukuk, M., Liesenfeld, R. (2006). Time series of count data: modeling, estimation and diagnostics. *Computational Statistics & Data Analysis*, 51: 2350-2364.
- Jung, R.C., Liesenfeld, R., Richard, J.F. (2011). Dynamic Factor Models for Multivariate Count Data: An Application to Stock-Market Trading Activity. *Journal of Business & Economic Statistics*, 29(1): 73-85.
- Kedem, B., Fokianos, K. (2002). *Regression Models for Time Series Analysis*. Wiley, Hoboken.
- Lütkepohl, H. (2005). *New Introduction to Multiple Time Series Analysis*. Springer.
- Papana, A. Kyrtsov, C., Kugiumtzis, D. and Diks, C. (2013). Simulation Study of Direct Causality Measures in Multivariate Time Series, *Entropy*, 15(7): 2635-2661
- Papaspyropoulos, K.G., Blioumis, V., Christodoulou, A.S., Birtsas, P.K. and Skordas, K.E. (2012). Challenges in implementing environmental management accounting tools: the case of a non-profit forestry organization. *Journal of Cleaner Production*, 29-30: 132-143.
- Paul, M. and Held, L. (2011). Predictive assessment of a non-linear random effects model for multivariate time series of infectious disease counts. *Statistics in Medicine*, 30: 1118-1136
- Paul, M., Held, L. and Toschke, A. M. (2008). Multivariate modelling of infectious disease surveillance data. *Statistics in Medicine*, 27: 6250-6267.
- Pedeli, X. and Karlis, D. (2012). On composite likelihood estimation of a multivariate INAR(1) model. *Journal of Time Series Analysis*. Early View, DOI: 10.1111/jtsa.12003.
- R Core Team (2012). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. (<http://www.R-project.org/>)
- Sims, C.A. (1980). Macroeconomics and reality. *Econometrica* 48: 1-48.



ΒΕΛΤΙΣΤΟΙ ΠΕΙΡΑΜΑΤΙΚΟΙ ΣΧΕΔΙΑΣΜΟΙ ΜΕ ΕΞΑΡΤΗΜΕΝΕΣ ΠΑΡΑΤΗΡΗΣΕΙΣ

Κατερίνα Περικλέους¹, Χρήστος Κουκουβίνος¹, Στρατής Κουνιάς²

¹Εθνικό Μετσόβιο Πολυτεχνείο

katerina.pericleous@gmail.com, ckoukouv@math.ntua.gr

² Πανεπιστήμιο Αθηνών, skounias@math.uoa.gr

ΠΕΡΙΛΗΨΗ

Οι πειραματικές μονάδες (μονάδες) είναι σε διάταξη γραμμής και σε κάθε μονάδα εφαρμόζεται μια από τις n αγωγές. Το πρόβλημα είναι να βρεθεί ο σχεδιασμός που δίνει βέλτιστες εκτιμήσεις των επιδράσεων των αγωγών. Το μοντέλο που χρησιμοποιείται είναι κύριων επιδράσεων, ομοιογενών μονάδων, όταν οι παρατηρήσεις έχουν εξάρτηση αυτοπαλινδρομούμενου μοντέλου πρώτης τάξης (AR(1)). Δίνονται καθολικά βέλτιστοι και φ-βέλτιστοι σχεδιασμοί και χρησιμοποιούνται κανόνες (φίλτρα) ούτως ώστε να μειωθεί το πλήθος των προς εξέταση σχεδιασμών για φ-βελτιστοποίηση. Στην εργασία αυτή εξετάζεται λεπτομερώς ο D-βέλτιστος σχεδιασμός για τέσσερις αγωγές, ενώ γίνεται εξαντλητική μελέτη για την περίπτωση των τριών αγωγών, όπου η παράμετρος α του AR(1) είναι αρνητική. Εάν το πλήθος των μονάδων είναι πολλαπλάσιο του 3 δίνεται ο φ-βέλτιστος σχεδιασμός, διαφορετικά παρουσιάζονται οι A-, D-, E-βέλτιστοι σχεδιασμοί.

Λέξεις κλειδιά: Καθολικά βέλτιστος, φ-βέλτιστος, A-, D-, E-βέλτιστοι σχεδιασμοί

1. ΕΙΣΑΓΩΓΗ

Υπάρχουν αρκετά παραδείγματα μονοδιάστατων εξαρτημένων σχεδιασμών όπως αναφέρονται από το Martin (1998). Σε αυτή την εργασία εξετάζουμε σχεδιασμούς γραμμής με ομοιογενή πληθυσμό, όπου οι πειραματικές μονάδες είναι τοποθετημένες σε σειρά και οι παρατηρήσεις ακολουθούν πρώτης τάξης αυτοπαλινδρομούμενο μοντέλο (AR(1)) με παράμετρο α . Εξετάζουμε το μοντέλο κύριων επιδράσεων και σε κάθε μονάδα εφαρμόζεται μια από τις n αγωγές. Μας ενδιαφέρει να βρούμε τους βέλτιστους σχεδιασμούς εκτίμησης της επίδρασης κάθε αγωγού. Η εργασία αυτή είναι συνέχεια της εργασίας Περικλέους και Κουνιάς (2010) και Περικλέους και Κουνιάς (2012), στις οποίες αναφερόμαστε ως (Π.Κ.2010) και (Π.Κ.2012) αντίστοιχα.

Ο Kiefer (1975) εισήγαγε τους καθολικά βέλτιστους σχεδιασμούς, μια ανασκόπηση των αποτελεσμάτων σχετικά με τους σχεδιασμούς γραμμής-στήλης δόθηκε από τους Hedayat, Jacroux και Majumdar (1988). Βέλτιστοι σχεδιασμοί με

εξάρτηση μελετήθηκαν από πολλούς ερευνητές, όπως Kiefer και Wynn (1981,1984), Martin (1986, 1998), Martin και Eccleston (1991,1993), Morgan και Uddin (1991, 2003), Uddin (1997), Pericleous και Kounias (2012). Α-βέλτιστοι σχεδιασμοί γραμμής-στήλης με τρεις αγωγές εξετάστηκαν από τον Morgan και Parvu (2007). Για την έννοια της κυριαρχίας και της εφαρμογής της αναφέρονται τα βιβλία των Marshall και Olkin (1979) και Pukelsheim (1993).

Ο ορισμός της συνάρτησης πληροφορίας $\varphi(\mathbf{C}) \rightarrow \Re \quad \forall \mathbf{C} \in nnd(k)$ δόθηκε στην (Π.Κ.2010), όπου, $nnd(k)$ είναι μη αρνητικά ορισμένος και $pd(k)$ είναι 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 $\varphi(\mathbf{C}_{d^*})$ είναι καθολικά βέλτιστος, στην κλάση των σχεδιασμών F , αν μεγιστοποιεί την $\varphi(\mathbf{C}_d)$, $\forall d \in F$ για κάθε συνάρτηση πληροφορίας $\varphi(\mathbf{C}_d)$.

Ορισμός 1.2 Ο σχεδιασμός d^* με πίνακα πληροφορίας $\mathbf{C}_{d^*}$ είναι φ-βέλτιστος, στην κλάση των σχεδιασμών F , αν ελαχιστοποιείτο άθροισμα $\varphi(\lambda_1) + \dots + \varphi(\lambda_k)$, όπου $\lambda = (\lambda_1, \dots, \lambda_k)$ οι ιδιοτιμές του πίνακα πληροφορίας $\mathbf{C}_d$, $d \in F$ για κάθε συνεχή φθίνουσα και κυρτή συνάρτηση $\varphi(\lambda)$.

Ο τελευταίος ορισμός ισχύει αν και μόνον αν το διάνυσμα $\lambda = (\lambda_1, \dots, \lambda_k)$ των ιδιοτιμών του $\mathbf{C}_d^*$ κυριαρχείται από το διάνυσμα $\mu = (\mu_1, \dots, \mu_k)$ των ιδιοτιμών του πίνακα πληροφορίας $\mathbf{C}_d$, $\forall d \in F$ (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.

Σε αυτήν την εργασία εξετάζουμε την περίπτωση όπου τα σφάλματα των παρατηρήσεων ακολουθούν ένα αυτοπαλινδρομούμενο μοντέλο πρώτης τάξης με παράμετρο α . Εξετάζουμε την περίπτωση που η παράμετρος α παίρνει τιμές στο διάστημα $(-1,0)$. Όταν η παράμετρος α βρίσκεται στο διάστημα $(0,1)$ είναι πιο απαιτητικό πρόβλημα και χρειάζεται μια τελείως διαφορετική προσέγγιση.

Στην παράγραφο 2 δίνουμε το μοντέλο, στην υποπαράγραφο 2.1 δίνουμε τον D-βέλτιστο σχεδιασμό για τέσσερις αγωγές και για όλες τις τιμές του n . Η προσέγγιση αυτή είναι αρκετή για να βρεθεί ο D-βέλτιστος σχεδιασμός για πέντε ή περισσότερες αγωγές. Στην υποπαράγραφο 2.2 παρουσιάζονται τεχνικές φιλτραρίσματος ώστε να μειωθούν οι προς εξέταση σχεδιασμοί.

Στην παράγραφο 3 δίνονται οι κλάσεις των σχεδιασμών που συγκρίνονται για τρεις αγωγές και στην παράγραφο 4 δίνονται οι βέλτιστοι σχεδιασμοί για όλες τις τιμές του n . Όταν το πλήθος n των μονάδων είναι $n \equiv 0 \pmod 3$ οι σχεδιασμοί είναι φ-βέλτιστοι, διαφορετικά οι σχεδιασμοί είναι A-D-E βέλτιστοι. Μερικές αποδείξεις Λημμάτων απαιτούν πολλές πράξεις και παραλείπονται.

2. ΟΜΟΙΟΓΕΝΕΙΣ ΜΟΝΑΔΕΣ, ΕΞΑΡΤΗΜΕΝΕΣ ΠΑΡΑΤΗΡΗΣΕΙΣ

Όταν έχουμε v αγωγές, το μοντέλο σε διανυσματική μορφή είναι,

$$\mathbf{y} = \mu_1 \mathbf{x}_1 + \dots + \mu_v \mathbf{x}_v + \mathbf{e}, \quad E(\mathbf{e}\mathbf{e}') = \sigma^2 \mathbf{V} \quad (2.1)$$

Το $n \times 1$ διάνυσμα $\mathbf{x}_j = (x_{1j}, \dots, x_{nj})'$ έχει $x_{ij} = 1$ αν στην μονάδα i εφαρμόζεται η αγωγή T_j , $j = 1, \dots, v$ και 0 αλλού, οπότε $\mathbf{x}_s' \mathbf{x}_t = \begin{cases} n_s & \text{αν } s = t \\ 0 & \text{αν } s \neq t \end{cases}$, όπου n_s το πλήθος των εμφανίσεων της αγωγής T_s στις v παρατηρήσεις.

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 $\text{var}(e) = \sigma^2 \mathbf{V}$. Ο πίνακας διασποράς $\mathbf{V}$ και ο αντίστροφός του δίνονται στην εργασία (Π.Κ 2010).

Ο πίνακας πληροφορίας $\mathbf{Q} = \mathbf{X}' \mathbf{V}^{-1} \mathbf{X}$, $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_v)$, όπου $\mathbf{Q} = \mathbf{A} + a^2 \mathbf{B} - a \mathbf{C}$,

$$\mathbf{A} = \begin{bmatrix} n_1 & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & n_v \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} \tilde{n}_1 & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \tilde{n}_v \end{bmatrix}, \quad \mathbf{C} = \begin{bmatrix} 2n_{11} & n_{12} + n_{21} & \dots & n_{1v} + n_{v1} \\ n_{12} + n_{21} & 2n_{22} & \ddots & \vdots \\ \vdots & \ddots & \ddots & n_{v-1,v} + n_{v,v-1} \\ n_{1v} + n_{v1} & \dots & n_{v-1,v} + n_{v,v-1} & 2n_{vv} \end{bmatrix}.$$

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), ..., (n-1,n), n_{ij} $i \neq j$ είναι το πλήθος των εμφανίσεων του ζεύγους των αγωγών $T_j T_i$ $i \neq j$ στις θέσεις (1,2), (2,3), ..., (n-1,n).

2.1 D-ΒΕΛΤΙΣΤΟΙ ΣΧΕΔΙΑΣΜΟΙ ΓΙΑ v ΑΓΩΓΕΣ

Όπως έχει αποδειχθεί και στην (ΠΚ2010) για ανεξάρτητες παρατηρήσεις και v αγωγές ο καθολικά βέλτιστος σχεδιασμός έχει $|n_i - n_j| \leq 1$, για $i \neq j$, $i, j = 1, \dots, v$. Ο πίνακας πληροφορίας $\mathbf{Q}$ δίνεται από τη σχέση (2.3), και έτσι $\text{trace}(\mathbf{Q}) = n + a^2(n-2) - 2a(n_{11} + \dots + n_{vv})$. Για την περίπτωση μας θέλουμε να βρούμε το σχεδιασμό με μέγιστο ίχνος του πίνακα πληροφορίας. Αυτή είναι και η αναγκαία συνθήκη για να είναι ένας σχεδιασμός φ-βέλτιστος. Εδώ δίνουμε τη λύση όταν $-1 < a < 0$ και $v = 4$, με όμοιο τρόπο λύνουμε το πρόβλημα για $v > 4$.

Θεώρημα: 2.1 Για $v=4$ αγωγές, όταν $-1 < a < 0$ και $\text{trace}(\mathbf{Q})$ επιτυγχάνει τη μέγιστη τιμή, ο D-βέλτιστος σχεδιασμός $d^* : \underbrace{T_1 T_1 \dots T_1}_{n_1} \underbrace{T_2 T_2 \dots T_2}_{n_2} \underbrace{T_3 T_3 \dots T_3}_{n_3} \underbrace{T_4 T_4 \dots T_4}_{n_4}$

έχει:

(i) $n_1 \geq n_4 \geq n_3 \geq n_2$, (ii) Εάν $n \equiv 0 \pmod{4}$, τότε $n_1 = n_4 = n_3 = n_2 = n/4$, (iii) Εάν $n \equiv 1 \pmod{4}$, τότε $n_1 = (n+3)/4$, $n_2 = n_3 = n_4 = (n-1)/4$, (iv) Εάν $n \equiv 2 \pmod{4}$, τότε $n_1 = n_4 = (n+2)/4$, $n_2 = n_3 = (n-2)/4$, (v) Εάν $n \equiv 3 \pmod{4}$, τότε $n_1 = n_3 = n_4 = (n+1)/4$, $n_2 = (n-3)/4$.

Απόδειξη. Από την (2.3) ο πίνακας πληροφορίας $\mathbf{Q}$ έχει μέγιστο ίχνος όταν τα $n_{11}, \dots, n_{vv}$ επιτυγχάνουν τη μέγιστη τιμή $n_{ii} = n_i - 1, i = 1, \dots, v$ και

$$d : \underbrace{T_1 T_1 \dots T_1}_{n_1} \underbrace{T_2 T_2 \dots T_2}_{n_2} \dots \underbrace{T_v T_v \dots T_v}_{n_v}, \text{ τότε,}$$

$$\mathbf{Q} = \begin{bmatrix} \tilde{x}_1 & -a & 0 & \dots & 0 \\ -a & x_2 & -a & \ddots & \vdots \\ & -a & \ddots & \ddots & 0 \\ & & \ddots & x_{v-1} & -a \\ sym & & & -a & \tilde{x}_v \end{bmatrix}, x_i = n_i(1-a)^2 + 2a, \tilde{x}_i = x_i - a^2. \quad (2.3)$$

Εάν $\det(\mathbf{Q}) = |\mathbf{Q}|$, $v = 4$, τότε

$$|\mathbf{Q}| = \tilde{x}_1 x_2 x_3 \tilde{x}_4 - a^2(x_2 \tilde{x}_1 + \tilde{x}_1 \tilde{x}_4 + \tilde{x}_4 x_3) + a^4. \quad (2.4)$$

Η αντίστροφη ακολουθία έχει την ίδια ορίζουσα. Έτσι για τη συνδιασπορά παίρνουμε $\tilde{x}_1 \geq \tilde{x}_4$, αυτό σημαίνει ότι, $n_1 \geq n_4$. Η μέγιστη τιμή $|\mathbf{Q}^*|$ της $|\mathbf{Q}|$ λαμβάνεται

όταν, i) $n_1 \geq n_4 \geq n_3 \geq n_2$. Αυτό διότι εναλλάσσονται τα x_2, x_3 εξασφαλίζοντας

$$|\mathbf{Q}_1| = \tilde{x}_1 x_2 x_3 \tilde{x}_4 - a^2(x_3 \tilde{x}_1 + \tilde{x}_1 \tilde{x}_4 + \tilde{x}_4 x_2) + a^4 \quad \text{και}$$

$$|\mathbf{Q}^*| - |\mathbf{Q}_1| = -a^2(\tilde{x}_1 - \tilde{x}_4)(x_2 - x_3) \geq 0 \text{ τότε } \tilde{x}_1 \geq \tilde{x}_4, \text{ και έτσι } x_2 \leq x_3 \Leftrightarrow n_2 \leq n_3.$$

ii) Εάν $n_1 \geq n_3$ εναλλάσσοντας τα x_1, x_3 παίρνουμε,

$$|\mathbf{Q}_2| = \tilde{x}_3 x_2 x_1 \tilde{x}_4 - a^2(x_2 \tilde{x}_3 + \tilde{x}_3 \tilde{x}_4 + \tilde{x}_4 x_1) + a^4 \text{ και}$$

$$|\mathbf{Q}^*| - |\mathbf{Q}_2| = a^2 x_2 (x_1 - x_3)(\tilde{x}_4 - 1) \geq 0. \text{ Έτσι } x_1 - x_3 \geq 0 \Leftrightarrow n_1 \geq n_3 \text{ διότι}$$

$$\tilde{x}_4 - 1 = (n_4(1-a)^2 - a^2 + 2a) - 1 \geq ((1-a)^2 - a^2 + 2a) - 1 = 0.$$

iii) Εάν $n_4 \geq n_3$ εναλλάσσοντας τα x_3, x_4 παίρνουμε,

$$|\mathbf{Q}_4| = \tilde{x}_1 x_2 x_4 \tilde{x}_3 - a^2(x_2 \tilde{x}_1 + \tilde{x}_1 \tilde{x}_3 + x_4 \tilde{x}_3) + a^4 \text{ και}$$

$$|\mathbf{Q}^*| - |\mathbf{Q}_4| = a^2(x_4 - x_3)(x_2 \tilde{x}_1 - (\tilde{x}_1 + a^2)) \geq 0. \text{ Όμως } \tilde{x}_1 \geq (1-a)^2 - a^2 + 1 = 1,$$

$x_2 = n_2(1-a)^2 + 2a \geq (1-a)^2 + 2a = 1+a^2$, επομένως $x_4 \geq x_3$. Άρα $n_2 \leq n_3 \leq n_4 \leq n_1$.

Εάν $n_1 \geq n_3 + 2$ τότε $|\mathbf{Q}_d|$ είναι βελτιωμένος, αφού θέτοντας $n_1 - 1$ και $n_3 + 1$ στη θέση των n_1, n_3 γράφουμε $|\mathbf{Q}(n_1 - 1, n_3 + 1)|$ και παίρνουμε τη (2.5).

$$|\mathbf{Q}(n_1 - 1, n_3 + 1)| = (\tilde{x}_1 - (1-a)^2)x_2(x_3 + (1-a)^2)\tilde{x}_4 + a^4 - a^2((\tilde{x}_1 - (1-a)^2)x_2 + (\tilde{x}_1 - (1-a)^2)\tilde{x}_4 + (x_3 + (1-a)^2)\tilde{x}_4) \Rightarrow$$

$$|\mathbf{Q}(n_1 - 1, n_3 + 1)| - |\mathbf{Q}| = x_2\tilde{x}_4(1-a)^2((1-a)^2(n_1 - n_3 - 1) - a^2) + a^2(1-a)^2x_2 > 0.$$

Επίσης όταν $n_4 \geq n_2 + 2$, τότε $|\mathbf{Q}(n_4 - 1, n_2 + 1)| - |\mathbf{Q}| > 0$.

Εάν $n_1 \geq n_2 + 2$, τότε $(n_1 - n_2 - 1)(1-a)^2 \geq 1 - a^2$,

$x_3 \geq (1-a)^2 + 2a = 1 + a^2 > 0$, $\tilde{x}_4 \geq 1$ και

$$|\mathbf{Q}(n_1 - 1, n_2 + 1)| - |\mathbf{Q}| = (1-a)^2((n_1 - n_2 - 1)(1-a)^2 - a^2)(x_3\tilde{x}_4 - a^2) + a^2\tilde{x}_4 > 0.$$

Άρα ο D-βέλτιστος σχεδιασμός έχει $|n_i - n_j| \leq 1$, $i \neq j, i = 1, 2, 3, 4$. •

3. ΤΡΕΙΣ ΑΓΩΓΕΣ S, T, R

Με τρεις αγωγές είναι μεγάλο το πλήθος των προς σύγκριση σχεδιασμών και αυξάνεται όσο αυξάνει το πλήθος η των μονάδων.

Μια διέξοδος είναι να χρησιμοποιήσουμε τα φίλτρα 1, 2, 3, 4 και 5, που δόθηκαν στο ΠΚ2012, όπου μπορεί να αναζητηθούν από τους ενδιαφερόμενους.

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.

Σημείωση. Εάν d_1, d_2 είναι δύο σχεδιασμοί με πίνακες πληροφορίας $\mathbf{Q}_1, \mathbf{Q}_2$ και: (i) $g(\mathbf{Q}_2) \geq g(\mathbf{Q}_1)$, για όλες τις συναρτήσεις πληροφορίας g, τότε γράφουμε $d_1 \rightarrow d_2$ ή ότι d_2 υπερκυριαρχεί τον d_1 .

(ii) Εάν $g(\mathbf{Q}_2) = g(\mathbf{Q}_1)$ γράφουμε $d_1 \Leftrightarrow d_2$.

Λήμμα 3.1 Εάν $-1 < a < 0$, ο σχεδιασμός $d_1 : \underbrace{SS\dots S}_{n_S+1} \underbrace{TT\dots T}_{n_T} \underbrace{RR\dots R}_{n_R+1}$ υπερκυριαρχεί του σχεδιασμού $d_2 : \underbrace{SS\dots S}_{n_S} \underbrace{TT\dots T}_{n_T} \underbrace{RR\dots R}_{n_R} SR$, δηλαδή $g(\mathbf{Q}_1) \geq g(\mathbf{Q}_2)$, όπου $g(\cdot)$

είναι συνάρτηση πληροφορίας.

Απόδειξη.

$$\mathbf{Q}_1 = \begin{bmatrix} n_S + 1 & 0 \\ & n_T \\ 0 & n_R + 1 \end{bmatrix} + a^2 \begin{bmatrix} n_S & 0 \\ & n_T \\ 0 & n_R \end{bmatrix} - a \begin{bmatrix} 2n_S & 1 & 0 \\ 1 & 2(n_T - 1) & 1 \\ 0 & 1 & 2n_R \end{bmatrix} \quad \text{και}$$

$$\mathbf{Q}_2 = \begin{bmatrix} n_S + 1 & 0 \\ & n_T \\ 0 & n_R + 1 \end{bmatrix} + a^2 \begin{bmatrix} n_S & 0 \\ & n_T \\ 0 & n_R \end{bmatrix} - a \begin{bmatrix} 2(n_S - 1) & 1 & 2 \\ 1 & 2(n_T - 1) & 1 \\ 2 & 1 & 2(n_R - 1) \end{bmatrix},$$

$$\text{δηλαδή, } \mathbf{Q}_1 - \mathbf{Q}_2 = -a \begin{bmatrix} 2 & 0 & -2 \\ 0 & 0 & 0 \\ -2 & 0 & 2 \end{bmatrix} \geq \mathbf{0}, \text{ άρα } g(\mathbf{Q}_1) \geq g(\mathbf{Q}_2) \quad \bullet$$

Θεώρημα 3.1 Εάν $-1 < a < 0$ και το συνολικό πλήθος των μονάδων είναι n , μόνο οι πιο κάτω 5 κλάσεις σχεδιασμών προς σύγκριση λαμβάνονται υπόψη:

$$F_0 : \underbrace{SS \dots}_{n_S} \underbrace{STT \dots}_{n_T} \underbrace{TRR \dots}_{n_R} R, \quad F_1 : \underbrace{SS \dots}_{n_S} \underbrace{TT \dots}_{n_T} \underbrace{RR \dots}_{n_R} RS,$$

$$F_2 : \underbrace{SS \dots}_{n_S} \underbrace{STT \dots}_{n_T} \underbrace{RR \dots}_{n_R} RT, \quad F_3 : \underbrace{SS \dots}_{n_S} \underbrace{STT \dots}_{n_T} \underbrace{RR \dots}_{n_R} RTS,$$

$$F_4 : \underbrace{TSS \dots}_{n_S} \underbrace{TTT \dots}_{n_T} \underbrace{RR \dots}_{n_R} RT.$$

Απόδειξη. Σύμφωνα με τους κανόνες φιλτραρίσματος, οι πιθανές ακολουθίες (κλάσεις σχεδιασμών), με n μονάδες, είναι:

$$(1) \underbrace{SS \dots}_{n_S} \underbrace{STT \dots}_{n_T} \underbrace{TRR \dots}_{n_R} R, \quad (2) \underbrace{SS \dots}_{n_S} \underbrace{STT \dots}_{n_T} \underbrace{TRR \dots}_{n_R} RS, \quad (3) \underbrace{SS \dots}_{n_S} \underbrace{STT \dots}_{n_T} \underbrace{TRR \dots}_{n_R} RT,$$

$$(4) \underbrace{SS \dots}_{n_S} \underbrace{STT \dots}_{n_T} \underbrace{TRR \dots}_{n_R} RTS, \quad (5) \underbrace{SS \dots}_{n_S} \underbrace{STT \dots}_{n_T} \underbrace{TRR \dots}_{n_R} RTSR,$$

$$(6) \underbrace{SS \dots}_{n_S} \underbrace{STT \dots}_{n_T} \underbrace{TSRR \dots}_{n_R} R, \quad (7) \underbrace{SS \dots}_{n_S} \underbrace{STT \dots}_{n_T} \underbrace{TSRR \dots}_{n_R} RS, \quad (8) \underbrace{SS \dots}_{n_S} \underbrace{STT \dots}_{n_T} \underbrace{TSRR \dots}_{n_R} RT,$$

$$(9) \underbrace{TSS \dots}_{n_S} \underbrace{TTT \dots}_{n_T} \underbrace{RR \dots}_{n_R} RT.$$

Οι κλάσεις (5), (6), (7), (8) είναι εκτός σύγκρισης (Λήμμα 3.1) και

$$(5) \underbrace{SS \dots}_{n_S} \underbrace{STT \dots}_{n_T} \underbrace{TRR \dots}_{n_R} RTSR \Leftrightarrow \underbrace{SS \dots}_{n_S} \underbrace{STT \dots}_{n_T} \underbrace{TRR \dots}_{n_R} STR \rightarrow \underbrace{SS \dots}_{n_S+1} \underbrace{STT \dots}_{n_T+1} \underbrace{TRR \dots}_{n_R+1} R,$$

$$(6) \underbrace{SS \dots}_{n_S} \underbrace{STT \dots}_{n_T} \underbrace{TSRR \dots}_{n_R} R \Leftrightarrow \underbrace{STT \dots}_{n_T} \underbrace{TSS \dots}_{n_S} \underbrace{RR \dots}_{n_R} R \Leftrightarrow \underbrace{TSS \dots}_{n_S} \underbrace{STT \dots}_{n_T} \underbrace{TRR \dots}_{n_S},$$

όμως,

$$\begin{aligned}
 (7) & \underbrace{T SS \dots S}_{n_S} \underbrace{TT \dots T}_{n_T} \underbrace{RR \dots R}_{n_R} \Leftrightarrow \underbrace{RR \dots R}_{n_R} \underbrace{TT \dots T}_{n_T} \underbrace{SS \dots S}_{n_S} \Leftrightarrow \underbrace{SS \dots S}_{n_S} \underbrace{TT \dots T}_{n_T} \underbrace{RR \dots R}_{n_R}, \\
 (8) & \underbrace{SS \dots S}_{n_S} \underbrace{TT \dots T}_{n_T} \underbrace{TS RR \dots RS}_{n_R} \Leftrightarrow \underbrace{STT \dots T}_{n_T} \underbrace{SS \dots S}_{n_S} \underbrace{RR \dots RS}_{n_R} \Leftrightarrow \underbrace{T SS \dots S}_{n_S} \underbrace{TT \dots T}_{n_T} \underbrace{RR \dots RT}_{n_R}, \\
 (8) & \underbrace{SS \dots S}_{n_S} \underbrace{TT \dots T}_{n_T} \underbrace{TS RR \dots RT}_{n_R} \rightarrow \underbrace{SS \dots S}_{n_S} \underbrace{S RR \dots RTT \dots TT}_{n_R} \Leftrightarrow \underbrace{SS \dots S}_{n_S+1} \underbrace{RR \dots RTT \dots T}_{n_R} \underbrace{TT \dots T}_{n_T+1}.
 \end{aligned}$$

Στην (8) εφαρμόζεται το Φίλτρο 3. ●

4. ΣΥΓΚΡΙΣΕΙΣ ΜΕΣΑ ΚΑΙ ΜΕΤΑΞΥ ΤΩΝ ΚΛΑΣΕΩΝ

Σε κάθε κλάση και για κάθε τιμή του $n=3m, 3m+1, 3m+2$ υπάρχει ένας σχεδιασμός που είναι φ-βέλτιστος είτε υπάρχουν περισσότεροι από ένας σχεδιασμοί οι οποίοι συγκρίνονται για A-, D-, E-βελτιστοποίηση.

4.1 ΒΕΛΤΙΣΤΟΙ ΣΧΕΔΙΑΣΜΟΙ ΜΕΣΑ ΣΤΙΣ ΚΛΑΣΕΙΣ

Θεώρημα 4.1 Εάν $-1 < a < 0$, για όλους τους σχεδιασμούς της κλάσης F_0 τα πιο κάτω είναι προς εξέταση: (i) Όταν $n=3m$, $d_0^* : \underbrace{SS \dots S}_{m} \underbrace{TT \dots T}_{m} \underbrace{RR \dots R}_{m}$, (ii) όταν

$$\begin{aligned}
 n=3m+1, & d_{11}^* : \underbrace{SS \dots S}_{m} \underbrace{TT \dots T}_{m+1} \underbrace{RR \dots R}_{m}, d_{12}^* : \underbrace{SS \dots S}_{m+1} \underbrace{TT \dots T}_{m-1} \underbrace{RR \dots R}_{m+1}, \\
 d_{13}^* : & \underbrace{SS \dots S}_{m+1} \underbrace{TT \dots T}_{m} \underbrace{RR \dots R}_{m} \text{ και (iii) όταν } n=3m+2, d_{21}^* : \underbrace{SS \dots S}_{m+1} \underbrace{TT \dots T}_{m} \underbrace{RR \dots R}_{m+1}, \\
 d_{22}^* : & \underbrace{SS \dots S}_{m+1} \underbrace{TT \dots T}_{m+1} \underbrace{RR \dots R}_{m}.
 \end{aligned}$$

Η απόδειξη παραλείπεται.

Θεώρημα 4.2 Εάν $-1 < a < 0$, από όλους τους σχεδιασμούς της κλάσης F_1 , μόνο ο σχεδιασμός $d_{14}^* : \underbrace{SS \dots S}_{m} \underbrace{TT \dots T}_{m} \underbrace{RR \dots RS}_{m}$, με $n=3m+1$, λαμβάνεται υπόψη.

Η απόδειξη παραλείπεται.

Θεώρημα 4.3 Εάν $-1 < a < 0$, από όλους τους σχεδιασμούς της κλάσης F_2 , υπάρχουν σχεδιασμοί της κλάσης F_0 που είναι A-D-E καλύτεροι.

Η απόδειξη παραλείπεται.

Θεώρημα 4.4 Εάν $-1 < a < 0$, από όλους τους σχεδιασμούς της κλάσης F_3 υπάρχουν σχεδιασμοί των κλάσεων F_0 και F_1 που είναι A-D-E καλύτεροι.

Η απόδειξη παραλείπεται.

Θεώρημα 4.5 Εάν $-1 < a < 0$, από όλους τους σχεδιασμούς της κλάσης F_4 , υπάρχουν σχεδιασμοί της κλάσης F_0 που είναι A-D-E καλύτεροι.

Η απόδειξη παραλείπεται.

Συμπερασματικά από τα Θεωρήματα 4.1 - 4.4 απομένουν για σύγκριση οι πιο κάτω σχεδιασμοί.

$$(i) n = 3m : d_0^* : \underbrace{SS \dots}_{m} \underbrace{STT \dots}_{m} \underbrace{T RR \dots}_{m} R.$$

$$(ii) n = 3m + 1 : d_{11}^* : \underbrace{SS \dots}_{m} \underbrace{STT \dots}_{m+1} \underbrace{T RR \dots}_{m} R, d_{12}^* : \underbrace{SS \dots}_{m+1} \underbrace{STT \dots}_{m-1} \underbrace{T RR \dots}_{m+1} R$$

$$d_{13}^* : \underbrace{SS \dots}_{m+1} \underbrace{STT \dots}_{m} \underbrace{T RR \dots}_{m} R, d_{14}^* : \underbrace{SS \dots}_{m} \underbrace{STT \dots}_{m} \underbrace{T RR \dots}_{m} R S.$$

$$(iii) n = 3m + 2 : d_{21}^* : \underbrace{SS \dots}_{m+1} \underbrace{STT \dots}_{m} \underbrace{T RR \dots}_{m+1} R, d_{22}^* : \underbrace{SS \dots}_{m+1} \underbrace{STT \dots}_{m+1} \underbrace{T RR \dots}_{m} R.$$

Στο επόμενο υποκεφάλαιο συγκρίνουμε τους σχεδιασμούς μεταξύ των κλάσεων για να βρούμε βέλτιστους σχεδιασμούς για κάθε τιμή του n.

4.2 ΒΕΛΤΙΣΤΟΙ ΣΧΕΔΙΑΣΜΟΙ ΜΕΤΑΞΥ ΤΩΝ ΚΛΑΣΕΩΝ

Θεώρημα 4.6 Εάν $-1 < a < 0$ και $n=3m$, ανάμεσα σε όλους τους σχεδιασμούς, ο σχεδιασμός $d_0^* : \underbrace{SS \dots}_{m} \underbrace{STT \dots}_{m} \underbrace{T RR \dots}_{m} R$ είναι φ-βέλτιστος.

Απόδειξη. Είναι ο μόνος Φ-βέλτιστος σχεδιασμός για $n=3m$. ●

Θεώρημα 4.7 Εάν $-1 < a < 0$ και $n=3m+1$ υπάρχουν τέσσερις σχεδιασμοί για σύγκριση:

(i) Ο A-βέλτιστος είναι ο σχεδιασμός $d_{13}^* : \underbrace{SS \dots}_{m+1} \underbrace{STT \dots}_{m} \underbrace{T RR \dots}_{m} R$ για $m \geq 2$ και για

$m=1$ και $-1 < a \leq -0.618034$. Για $m=1$ και $-0.618034 \leq a < 0$ ο A-βέλτιστος σχεδιασμός είναι ο $d_{14}^* : STRS$.

(ii) Ο D- βέλτιστος είναι ο σχεδιασμός $d_{13}^* : \underbrace{SS \dots}_{m+1} \underbrace{STT \dots}_{m} \underbrace{T RR \dots}_{m} R$ για $m \geq 2$. Για

$m=1$ οι σχεδιασμοί $d_{13}^* : SS TR$ και $d_{11}^* : ST TR$ είναι D-βέλτιστοι.

(iii) Ο E- βέλτιστος είναι ο σχεδιασμός d_{11}^* για $-1 < a < -0.618$ και ο σχεδιασμός d_{14}^* είναι E-βέλτιστος για $-0.618 < a < 0$, $m \geq 1$.

Η απόδειξη παραλείπεται.

Θεώρημα 4.8 Εάν $-1 < a < 0$ και $n=3m+2$, ο σχεδιασμός $d_{21}^* : \underbrace{SS \dots S}_{m+1} \underbrace{TT \dots T}_m \underbrace{RR \dots R}_{m+1}$ είναι E- και D-βέλτιστος, η A-βελτιστοποίηση των d_{21}^* ή d_{22}^* εξαρτάται από την τιμή του a όπως δίνεται στον παρακάτω πίνακα 1. Η απόδειξη παραλείπεται.

ΠΙΝΑΚΑΣ 1
A-βελτιστοποίηση των d_{21}^* , d_{22}^* , $n=3m+2$

m	1	2	5	10	20	50
d_{21}	$-1 < a < -0.402$	$-1 < a < -0.535$	$-1 < a < -0.642$	$-1 < a < -0.683$	$-1 < a < -0.705$	$-1 < a < -0.718$
d_{22}	$-0.402 < a < 0$	$-0.535 < a < 0$	$-0.642 < a < 0$	$-0.683 < a < 0$	$-0.705 < a < 0$	$-0.718 < a < 0$

ABSTRACT

The experimental units (units) are placed in a row and to every unit one out of v treatments is allocated. The problem is to find the design which gives optimal estimates of the treatments effects. We use the main effects model with homogeneous population, when the observations follow a first order autoregressive (AR(1)) scheme. Universal optimality and Φ -optimality are defined and a filtering procedure is applied to reduce the number of competing Φ -optimal designs. In this paper the D-optimal design for four treatments is examined in detail and exhaustive search is developed for the case of three treatments, where the parameter α of AR(1) is negative. If the number n of units is a multiple of 3 the Φ -optimal design is given, A-, D-, E-optimal designs are presented for all other values of n .

ΑΝΑΦΟΡΕΣ

- Hedayat, A.S., Jacroux. M., Majumbar, D. (1988). Optimal designs for comparing test treatments with controls. *Statistical Science Vol. 3, No 4, 462-491*.
- Kiefer, J. (1975) *Construction and optimality of generalized Youden designs*. In a Survey of Statistical Design and Linear Models, Edited by J. N. Srivastava, 333-353. Amsterdam, North Holland.
- Kiefer, J. and Wynn, H.P. (1981). Optimum balanced block and Latin square designs for correlated observations. *Ann. Statist.*, 9, 737-757.
- Kiefer, J. and Wynn, H. P. (1984). Optimum and minimax exact treatment designs for one-dimensional autoregressive error processes. *Ann. Statist.*, 12, 431-450.

- Marshall, A.W. and Olkin, I. (1979). *Inequalities: Theory of majorization and its applications*. Academic Press, New York.
- Martin, R.J. (1986). On the design of experiments under spatial correlation. *Biometrika* 73, 247-277.
- Martin, R.J. (1998). Optimal designs for small-sized blocks under dependence. *Journal of Combinatorics, Information and System Science* Vol. 23, 95-104.
- Martin, R. J. and Eccleston, J.A. (1991). Optimal incomplete block designs for general dependent structures. *J. Statist. Plann. Inf.* 28, 67-81.
- Martin, R. J. and Eccleston, J.A. (1993). Incomplete block designs with special layouts, when observations are dependent. *J. Statist. Plann. Inf.* 35, 77-92.
- Morgan, J.P. and Uddin, N. (1991). Two-dimensional; designs for correlated errors. *The Annals of Statistics*, vol. 19,4, 2160-2182.
- Morgan, J.P. and Uddin, N. (2003). Optimal row-column designs for two treatments. *J. Statist. Plann. Inf.* 115, 603-622.
- Morgan, J.P. and Parvu V. (2007). Optimal row-column design for three treatments. *J. Statist. Plann. Inf.* 137, 1474-1487.
- Pericleous, K. and Kounias, S. (2012). The concept of majorization in experimental design, *Communications in Statistics-Theory and Methods*, 41:2836-2850.
- Pukelsheim, F. (1993). *Optimal designs of experiments*. John Wiley and Sons, Inc.
- Uddin, N. (1997). Row-column designs for comparing two treatments in the presence of correlated errors. *J. Statist. Res.* 31, 75-81.
- Περικλέους, Κ., Κουνιάς, Σ. (2010). *Η Έννοια της Κυριαρχίας στους Πειραματικούς Σχεδιασμούς*. Πρακτικά 22ου Πανελληνίου Συνεδρίου Στατιστικής, Χανιά, Ελλάς, σελ.167-174.
- Περικλέους, Κ., Κουνιάς, Σ. (2012). *Σχεδιασμοί Γραμμής-Στήλης με τρεις αγωγές και εξαρτημένες παρατηρήσεις*, Πρακτικά 24ου Πανελληνίου Συνεδρίου Στατιστικής, Πάτρα, Ελλάς, σελ.280-287.



3^k ΒΕΛΤΙΣΤΟΙ ΣΧΕΔΙΑΣΜΟΙ ΓΙΑ ΤΗΝ ΕΚΤΙΜΗΣΗ ΓΡΑΜΜΙΚΩΝ ΚΑΙ ΤΕΤΡΑΓΩΝΙΚΩΝ ΑΝΤΙΘΕΣΕΩΝ, ΟΤΑΝ $N \equiv 0 \pmod{3}$

**Κατερίνα Περικλέους¹, Σταύρος Χατζόπουλος², Φωτεινή Κολυβά-Μαχαίρα³,
Στρατής Κουνιάς⁴**

¹Σχολή Εφαρμοσμένων Μαθηματικών και Φυσικών Επιστημών, Εθνικό Μετσόβιο
Πολυτεχνείο, katerina.pericleous@gmail.com,

²Τμήμα Μαθηματικών, Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης,
cstavros@math.auth.gr,

³Τμήμα Μαθηματικών, Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης
fkolyva@math.auth.gr,

⁴Τμήμα Μαθηματικών, Εθνικό Καποδιστριακό Πανεπιστήμιο Αθηνών,
skounias@math.uoa.gr

ΠΕΡΙΛΗΨΗ

Βρίσκονται βέλτιστοι σχεδιασμοί, χρησιμοποιώντας τη μέθοδο κυριαρχίας διανυσμάτων, για την εκτίμηση των τυποποιημένων γραμμικών και τετραγωνικών αντιθέσεων, ενός γραμμικού μοντέλου. Δείχνεται ότι οι Φ βέλτιστοι σχεδιασμοί για $N \equiv 0 \pmod{9}$ είναι οι ορθογώνιοι σχηματισμοί (Orthogonal Arrays), για $N \equiv 3 \pmod{9}$ και $N \equiv 6 \pmod{9}$ δίνονται οι E-A-D βέλτιστοι σχεδιασμοί. Η περίπτωση $N=12$ χρησιμοποιείται ως υπόδειγμα και μέχρι 5 παράγοντες δίνονται οι βέλτιστοι σχεδιασμοί και οι πίνακες πληροφορίας. Για την κατασκευή των βέλτιστων σχεδιασμών με το μέγιστο αριθμό παραγόντων χρησιμοποιούνται αλγόριθμοι, που εξαντλητική έρευνα όλων των περιπτώσεων.

Λέξεις και φράσεις κλειδιά: Φ -A-E-D βέλτιστοι σχεδιασμοί, ισορροπημένοι και μερικώς ισορροπημένοι σχεδιασμοί, κυριαρχία διανυσμάτων.

1.ΕΙΣΑΓΩΓΗ.

Η εργασία αυτή είναι συνέχεια της εργασίας που παρουσιάσαμε στο 25ο συνέδριο του ΕΣΙ στο Βόλο με τον ίδιο περίπου τίτλο (Πρακτικά ΕΣΙ 25^{ου} συνεδρίου, σελ. 179-187), θα την συμβολίζουμε ΕΣΙ-25°. Στην παρούσα εργασία μελετάμε τους βέλτιστους πειραματικούς σχεδιασμούς στην περίπτωση που υπάρχουν k παράγοντες και κάθε παράγοντας μετέχει σε κάθε παρατήρηση σε μια από τρεις στάθμες, που τις συμβολίζουμε 0,1,2, το πλήθος των παρατηρήσεων είναι $N \equiv 0 \pmod{3}$.

Ορισμένες αποδείξεις και ορισμοί που περιέχονται στο ESI-25^ο παραλείπονται.

Σε προηγούμενες εργασίες οι βέλτιστοι σχεδιασμοί, βρίσκονταν με την επαύξηση των γραμμών ενός ορθογώνιου σχηματισμού OA(N, k, 3, 2), όπως D-optimal (Kolyva-Machaira, 1989a), Goptimal (Kolyva-Machera, 1989b), και βέλτιστοι κάτω από τα κριτήρια τύπου 1 του Cheng (Mukerjee, 1999).

Τα OA(N,k,3,2) είναι $N \times k$ σχηματισμοί που τα στοιχεία τους είναι 0,1,2, το πλήθος των 0,1,2 σε κάθε στήλη είναι N/3 και το πλήθος των 00,01,02,10,11, 12, 20,21,22 σε κάθε δύο στήλες είναι ίσο με N/9. Τα OA(N,k,3,2) έχουν κατασκευαστεί για πολλές τιμές του $N \equiv 0 \pmod{9}$

Στην εργασία αυτή χρησιμοποιούμε την έννοια της κυριαρχίας (majorization), όπως εφαρμόζεται στους καθολικά βέλτιστους σχεδιασμούς (Pukelsheim, 1993 p.139-142, 352-358). Μας ενδιαφέρει η εκτίμηση των τυποποιημένων γραμμικών και τετραγωνικών αντιθέσεων των παραγόντων

Εισάγονται οι ορισμοί των ισορροπημένων σχηματισμών (Balanced Arrays, BA(N,k)) και των μερικώς ισορροπημένων σχηματισμών (Partially Balanced Arrays, PBA(N,k-r,r)).

Το μοντέλο για την εκτίμηση των τυποποιημένων γραμμικών και τετραγωνικών αντιθέσεων, σε διανυσματική μορφή, ESI-25ο, είναι:

$$\mathbf{y} = \mu \mathbf{I}_N + \sum_{i=1}^k \left[\frac{q_i - l_i}{\sqrt{2}} \mathbf{g}_i + \frac{q_i - 2u_i + l_i}{\sqrt{6}} \mathbf{h}_i \right] \quad (1.1)$$

Όπου μ είναι ο γενικός μέσος, $l_i, u_i, q_i, i = 1, \dots, k$ είναι η επίδραση του παράγοντα F_i όταν μετέχει σε μια παρατήρηση σε στάθμη 0,1,2 αντίστοιχα, $d(I,s)$ είναι η στάθμη με την οποία μετέχει ο παράγοντας F_i στην s παρατήρηση, και $\mathbf{g}_i = (g_{is}), \mathbf{h}_i = (h_{is})$ είναι $N \times 1$ διανύσματα με

$$g_{ij} = \begin{cases} -1 & \text{αν } d(i,s) = 0 \\ 0 & \text{αν } d(i,s) = 1 \\ 1 & \text{αν } d(i,s) = 2 \end{cases}, \quad h_{ij} = \begin{cases} 1 & \text{αν } d(i,s) = 0 \\ 0 & \text{αν } d(i,s) = 1 \\ 1 & \text{αν } d(i,s) = 2 \end{cases}$$

$\frac{q_i - l_i}{\sqrt{2}}$ είναι η τυποποιημένη γραμμική αντίθεση (contrast) και $\frac{q_i - 2u_i + l_i}{\sqrt{6}}$ η τυποποιημένη τετραγωνική αντίθεση.

Αν $\sigma^2 \mathbf{V}$ είναι ο $(2k) \times (2k)$ πίνακας διασποράς των εκτιμητών των $2k$ παραμέτρων, ο πίνακας $\mathbf{Q} = \mathbf{V}^{-1}$ λέγεται πίνακας πληροφορίας των υπό εκτίμηση παραμέτρων. Για την εύρεση του πίνακα πληροφορίας είναι αρκετό να γνωρίζουμε τον πίνακα σχεδιασμού, δηλαδή τα $2k$ διανύσματα $\mathbf{g}_1, \mathbf{g}_2, \dots, \mathbf{g}_k, \mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_k$.

Στην παράγραφο 2 της παρούσας εργασίας δίνονται θεωρήματα και λήμματα για την εύρεση των βέλτιστων σχεδιασμών, όταν $N=3,6 \bmod 9$, στην παράγραφο 3 δίνονται θεωρήματα και λήμματα για την εύρεση βέλτιστων σχεδιασμών όταν $N=0 \bmod 3$. Στην παράγραφο 4 παρουσιάζεται υπόδειγμα και δίνονται οι βέλτιστοι σχεδιασμοί για $N=12$ και πλήθος παραγόντων $k=2,3,4,5$.

2. ΒΕΛΤΙΣΤΟΙ ΣΧΕΔΙΑΣΜΟΙ ΓΙΑ $N=0 \bmod 3$.

Ο πίνακας πληροφορίας των υπό εκτίμηση παραμέτρων είναι,

$$\mathbf{Q} = \begin{bmatrix} \mathbf{Q}_1 & \mathbf{Q}_{12} \\ \mathbf{Q}_{12}' & \mathbf{Q}_2 \end{bmatrix}, \mathbf{Q}_1 = \mathbf{A} - \frac{\mathbf{a}\mathbf{a}'}{N}, \mathbf{Q}_2 = 9(\mathbf{C} - \frac{\mathbf{b}\mathbf{b}'}{9}), \mathbf{Q}_{12} = 3(\mathbf{B} - \frac{\mathbf{a}\mathbf{b}'}{N}) \quad (2.1)$$

Οι πίνακες αυτοί είναι συναρτήσεις των $\mathbf{g}_i, \mathbf{h}_i, i = 1, \dots, k$, δες ΕΣΙ-25°.

Με $\lambda(\mathbf{Q})$ συμβολίζεται το διάνυσμα των ιδιοτιμών του πίνακα $\mathbf{Q}$.

Οι πίνακες $\mathbf{Q}_1, \mathbf{Q}_{12}, \mathbf{Q}_2$ είναι $k \times k$ και $\mathbf{Q}_0$ είναι ο πίνακας που προκύπτει αν $\mathbf{Q}_{12} = \mathbf{0}$, τότε το διάνυσμα των ιδιοτιμών του πίνακα $\mathbf{Q}_0$ κυριαρχείται από το διάνυσμα ιδιοτιμών του αρχικού πίνακα (Marshall and Olkin (1979), p.225). Επομένως στους σχεδιασμούς μας θα φροντίζουμε να είναι $\mathbf{Q}_{12} = \mathbf{0}$ και αυτό συμβαίνει μόνο αν ο πίνακας σχεδιασμού έχει $n_i^0 = n_i^2, n_{ij}^{01} = n_{ij}^{21}, n_{ij}^{10} = n_{ij}^{12}, i \neq j, i, j = 1, \dots, k$. δες ΕΣΙ-25°.

Λήμμα 2.1 Αν $k \geq 2$ και $n_i^0 = n_i^2, n_{ij}^{01} = n_{ij}^{21}, n_{ij}^{10} = n_{ij}^{12} \quad 1 \leq i \neq j \leq k$, τότε ο πίνακας

πληροφορίας (2.1) γίνεται $\mathbf{Q}_0 = \begin{bmatrix} \mathbf{Q}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{Q}_2 \end{bmatrix}, \mathbf{Q}_1 = \mathbf{A} = (a_{ij}), \mathbf{Q}_2 = (q_{ij})$ με

$$a_{ii} = (N - n_i^1) / 2, \quad a_{ij} = (n_{ij}^{00} + n_{ij}^{22} - n_{ij}^{20} - n_{ij}^{02}) / 2, \quad q_{ii} = 3(N - n_i^1)n_i^1 / (2N), \\ q_{ij} = 3(Nn_{ij}^{11} - n_i^1n_j^1) / (2N).$$

Η απόδειξη παραλείπεται. •

Θεώρημα 2.1. Αν σε ένα μη αρνητικά ορισμένο συμμετρικό $k \times k$ πίνακα $\mathbf{A} = (a_{ij})$ πάρουμε το μέσο όρο a των διαγωνίων στοιχείων και το μέσο όρο b των μη διαγωνίων στοιχείων, τότε ο πίνακας $\mathbf{A}^* = (a - b)\mathbf{I}_k + b\mathbf{J}_k$ έχει $\lambda(\mathbf{A}^*) \prec \lambda(\mathbf{A})$.

Απόδειξη. $\mathbf{A}^* = \frac{1}{k!} \sum_{i=1}^{k!} \mathbf{P}_i \mathbf{A} \mathbf{P}_i'$, όπου $\mathbf{P}_i$ είναι μεταθετικοί πίνακες και

$$\sum_{i=1}^j \lambda_i(\mathbf{A}^*) = \max_{\mathbf{U}\mathbf{U}'=\mathbf{I}_j} \text{tr}(\mathbf{U}\mathbf{A}^*\mathbf{U}') \leq \left(\frac{1}{k!} \sum_{i=1}^{k!} \max_{\mathbf{U}\mathbf{U}'=\mathbf{I}_j} \text{tr}(\mathbf{U}\mathbf{P}_i\mathbf{A}\mathbf{P}_i'\mathbf{U}') \right) = \sum_{i=1}^j \lambda_i(\mathbf{A})$$

Διότι $\lambda(\mathbf{P}_i\mathbf{A}\mathbf{P}_i') = \lambda(\mathbf{A})$ και $\lambda_1(\mathbf{A}) \geq \lambda_2(\mathbf{A}) \geq \dots \geq \lambda_j(\mathbf{A})$, $j \leq k$, τότε

$\lambda(\mathbf{A}^*) \prec \lambda(\mathbf{A})$ (Marshall and Olkin p.511).•

Ορισμός 2.1 Ο $N \times k$ πίνακας με στοιχεία 0,1,2, που έχει (i) σε κάθε στήλη τα ίδια n^0, n^1, n^2 (ii) σε κάθε δύο στήλες έχει τα ίδια $n^{00}, n^{01}, n^{02}, n^{10}, n^{11}, n^{12}, n^{20}, n^{21}, n^{22}$ (iii) $n^0 = n^2, n^{01} = n^{21} = n^{12} = n^{10}, n^{00} = n^{22}, n^{02} = n^{20}, n^1 - n^{11} = 0 \pmod{2}$, $0 < n^{11} < n^1$, λέγεται ισορροπημένος σχηματισμός και συμβολίζεται $\text{BA}(N, k)$ •

Επομένως οι σχεδιασμοί που είναι $\text{BA}(N, k)$ δίνουν πίνακες πληροφορίας, σύμφωνα με το Λήμμα 2.1, που είναι

$$\mathbf{Q}_0 = \begin{bmatrix} \mathbf{Q}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{Q}_2 \end{bmatrix}, \quad \mathbf{Q}_1 = \mathbf{A} = (a_{ij}), \quad \mathbf{Q}_2 = (q_{ij}) \quad (2.2)$$

Με $a_{ii} = (N - n^1) / 2$, $a_{ij} = (n^{00} - n^{02})$ $i \neq j$, $q_{ii} = 3(N - n^1)n^1 / (2N)$, $q_{ij} = 3(Nn^{11} - n^1n^1) / (2N)$, $i \neq j = 1, 2, \dots, k$, $n^{01} = n^{21} = (n^1 - n^{11}) / 2$.

Καταλήγουμε επομένως στις σχέσεις,

$$\mathbf{Q}_1 = ((N - n^1) / 2 - (n^{00} - n^{02}))\mathbf{I}_k + (n^{00} - n^{02})\mathbf{J}_k \quad (2.3)$$

$$\mathbf{Q}_2 = \frac{3}{2}(n^1 - n^{11})\mathbf{I}_k + \frac{3}{2N}(Nn^{11} - n^1n^1)\mathbf{J}_k \quad (2.4)$$

Οι ιδιοτιμές του πίνακα $a_i\mathbf{I}_k + b_i\mathbf{J}_k$ είναι $\lambda_1 = \dots = \lambda_{k-1} = a_i, \lambda_k = a_i + kb_i$ και ο Φ-βέλτιστος έχει το μέγιστο a_i και το μέγιστο $a_i + b_i$.

Αν $N \equiv 0 \pmod{3}$, το μέγιστο ίχνος του $\mathbf{Q}_0$ πετυχαίνεται όταν $n^1 = N / 3$, με μέγιστο ίχνος $2kN / 3$ και παραμέτρους,

$$n^{11}, 0 < n^{11} < n^1 = N / 3, n^1 - n^{11} = 0 \pmod{2} \quad (2.5)$$

$$n^{00} + n^{02} = n^0 - n^{01} = (N + 3n^{11}) / 6 \quad (2.6)$$

$$\lambda_1(\mathbf{Q}_1) = \dots = \lambda_{k-1}(\mathbf{Q}_1) = \frac{N}{3} - (n^{00} - n^{02}), \quad \lambda_k(\mathbf{Q}_1) = \frac{N}{3} + (k-1)(n^{00} - n^{02}) \quad (2.7)$$

$$\lambda_1(\mathbf{Q}_2) = \dots = \lambda_{k-1}(\mathbf{Q}_2) = \frac{N - 3n^{11}}{2}, \lambda_k(\mathbf{Q}_2) = \frac{N - 3n^{11}}{2} + \frac{k(9n^{11} - N)}{6} \quad (2.8)$$

3. Βέλτιστοι ισορροπημένοι σχεδιασμοί BA(N,k), όταν $N \equiv 0 \pmod{3}$

Οι σχεδιασμοί BA(N,k), όταν $N \equiv 0 \pmod{3}$, με μέγιστο ίχνος έχουν $n^1 = N/3$ και το μέγιστο ίχνος είναι το ίδιο για όλες τις τιμές του n^{11} (σχέση (2.5)), επομένως πρέπει να βρούμε για ποια τιμή του n^{11} ο σχεδιασμός BA(N,k) είναι βέλτιστος.

Για την εύρεση του βέλτιστου BA(N,k) χρειαζόμαστε τα επόμενα λήμματα.

Λήμμα 3.1. (i) Αν $s_1 : y_1 = y_2 = \dots = y_k = y$, $s_2 : x_1, x_2, \dots, x_k$, $x_1 + \dots + x_k = ky$ και τα $x_1, \dots, x_k$ δεν είναι όλα ίσα, τότε $s_1 \prec s_2$.

(ii) Αν $s_3 : x_1 = \dots = x_{k-1} = x$, $x_k = x + kb$, $b > 0$, $s_4 : y_1 = \dots = y_{k-1} = y$, $y_k = y + kc$, $c > 0$ και $x + b = y + c$, $b < c$, τότε $s_3 \prec s_4$.

(iii) Αν $s_5 : x_1 = \dots = x_{k-1} = x$, $x_k = x - kb$, $b > 0$, $s_6 : y_1 = \dots = y_{k-1} = y$, $y_k = y - kc$, $c > 0$ και $x - b = y - c$, $c > b$, τότε $s_5 \prec s_6$.

Η απόδειξη παραλείπεται.

Θεώρημα 3.1 Αν $N \equiv 0 \pmod{9}$ και το πλήθος των παραγόντων είναι $k \geq 2$, ο Φ βέλτιστος σχεδιασμός είναι OA(N,k,3,2), που σημαίνει $n^{pq} = N/9$, $p, q = 0, 1, 2$.

Και ο πίνακας πληροφορίας είναι $\mathbf{Q}_0 = (N/3)\mathbf{I}_{2k}$.

Απόδειξη. Αν $\delta(\mathbf{Q}_0)$ είναι το διάνυσμα των διαγωνίων στοιχείων του $2k \times 2k$ πίνακα $\mathbf{Q}_0$ και $\lambda(\mathbf{Q}_0)$ το διάνυσμα των ιδιοτιμών του $\mathbf{Q}_0$, τότε $\delta(\mathbf{Q}_0) \prec \lambda(\mathbf{Q}_0)$ (Pukelsheim p. 146), και $\delta(\mathbf{Q}_0) = \lambda(\mathbf{Q}_0)$ μόνο αν ο πίνακας $\mathbf{Q}_0$ είναι διαγώνιος.

Από τους διαγώνιους πίνακες με το ίδιο ίχνος, Φ βέλτιστος είναι αυτός που έχει όλα τα διαγώνια στοιχεία ίσα (Λήμμα 3.1 (i)). Επομένως αν ο πίνακας πληροφορίας είναι διαγώνιος, με ίσα διαγώνια στοιχεία και μέγιστο ίχνος $2kN/3$, τότε κάθε διαγώνιο στοιχείο είναι ίσο με $(2kN/3)/(2k) = N/3$.

Αυτό συμβαίνει μόνο αν $a_{ij} = n^{00} - n^{02} = 0$, $q_{ij} = 0 \Leftrightarrow n^{11} = N/9$ (Λήμμα 2.1) και τότε $n^{01} = n^{21} = n^{12} = n^{10} = ((N/3) - (N/9))/2 = N/9$ (σχέση 2.6), τότε $n^{00} = n^{22} = n^{02} = n^{20} = N/9$, αυτό όμως είναι OA(N,k,3,2).•

3.1 Μερικώς ισορροπημένοι σχεδιασμοί, PBA(N,k-r,r)

Οι βέλτιστοι ισορροπημένοι σχεδιασμοί BA(N,k) δεν είναι και βέλτιστοι σχεδιασμοί, αυτό διαπιστώνεται αν ο πίνακας πληροφορίας (2.2) του BA(N,k) πάρει τη μορφή,

$$\mathbf{Q}_0 = \begin{bmatrix} \mathbf{Q}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{Q}_2 \end{bmatrix}, \mathbf{Q}_1 = \begin{bmatrix} \mathbf{Q}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{Q}_{12} \end{bmatrix}, \mathbf{Q}_2 = \begin{bmatrix} \mathbf{Q}_{21} & \mathbf{0} \\ \mathbf{0} & \mathbf{Q}_{22} \end{bmatrix} \quad (3.1)$$

Όπου $\mathbf{Q}_{11}, \mathbf{Q}_{21}$ είναι $(k-r) \times (k-r)$ και $\mathbf{Q}_{12}, \mathbf{Q}_{22}$ είναι $r \times r$, $0 \leq r < k$. Αυτό συμβαίνει διότι $\lambda \left(\begin{bmatrix} \mathbf{H}_{11} & \mathbf{0} \\ \mathbf{0} & \mathbf{H}_{22} \end{bmatrix} \right) \prec \lambda \left(\begin{bmatrix} \mathbf{H}_{11} & \mathbf{H}_{12} \\ \mathbf{H}_{12}' & \mathbf{H}_{22} \end{bmatrix} \right)$, Marshall and Olkin p. 225.

Ορισμός 3.1. Ένας σχεδιασμός $N \times k$ με στοιχεία 0,1,2, που αποτελείται (i) από ένα $d_1 : BA(N, k-r)$, (ii) Από ένα $d_2 : BA(N, r)$, (iii) Οι παράμετροι του d_1 είναι $n_1^0 = n_1^2, n_1^{01} = n_1^{21}, n_1^{10} = n_1^{12}$, του d_2 είναι $n_2^0 = n_2^2, n_2^{01} = n_2^{21}, n_2^{10} = n_2^{12}$ (iv) Κάθε στήλη του d_1 με κάθε στήλη του d_2 έχουν $n_{12}^{11} = n_2^1 / 3, n_2^1 = 0 \bmod 3, N - n_2^1 = 0 \bmod 2$, λέγεται μερικώς ισιρροπημένος σχεδιασμός PBA(N,k-r,r)•

Στον ορισμό αυτό το (iii) είναι απαραίτητο για να είναι ο αρχικός πίνακας $\mathbf{B}=\mathbf{0}$, το (iv) χρειάζεται για να έχει τη μορφή (3.1) ο πίνακας $\mathbf{Q}_0$.

Λήμμα 3.2 Αν ο σχεδιασμός $d_1 : BA(N, k-r)$ έχει πίνακα πληροφορίας με μέγιστο ίχνος, τότε ο σχεδιασμός $d_2 : BA(N, r)$ έχει μέγιστο ίχνος:

- (i) Αν $N \equiv 3 \bmod 9 = 9m+3$, τότε $n_2^1 = 3m+3, n_2^0 = n_2^2 = 3m, n_{12}^{11} = m+1$
- (ii) Αν $N \equiv 6 \bmod 9 = 9m+6$, τότε $n_2^1 = 3m, n_2^0 = n_2^2 = 3m+3, n_{12}^{11} = m$

Η απόδειξη από έλλειψη χώρου παραλείπεται.

Ο Πίνακας πληροφορίας του $d_1 : BA(n, k-r)$ έχει $n_1^1 = N/3$ και βρίσκεται από τις (2.3), (2.4) θέτοντας $n_1^1 = N/3$

4. Βέλτιστοι σχεδιασμοί για N=12

Στην εύρεση βέλτιστων σχεδιασμών θα υπολογίσουμε τους πίνακες $\mathbf{Q}_1, \mathbf{Q}_2$ από τις σχέσεις (2.3), (2.4). Επίσης οι βέλτιστοι BA(N,k) έχουν $n^{00} - n^{02}$ ίσο με 0,1 ή -1, όπως προκύπτει από τις ιδιοτιμές στη (2.7-2.8) και το Λήμμα 3.1.

Από τον ορισμό 3.1 και το λήμμα 3.2, ο σχεδιασμός d_1 έχει $n_1^1 = n_1^0 = n_1^2 = (N/3)$,

Βρίσκουμε τα n^{11} που δίνουν τους βέλτιστους σχεδιασμούς για τις διάφορες τιμές του k. Γίνεται υποδειγματική εφαρμογή για N=12.

N=12, k=2,3,4,5

$$n^1 = n^0 = n^2 = 4, n^{11} = 2, n^{01} = n^{21} = n^{10} = n^{12} = (4 - 2) / 2 = 1, \\ n^{00} + n^{02} = 4 - 1 = 3, \quad \text{τότε} \quad n^{00} = n^{22} = 2, n^{02} = n^{20} = 1 \\ \text{ή } n^{00} = n^{22} = 1, n^{02} = n^{20} = 2$$

Οπότε έχουμε δύο BA(12,k), ένα όταν $n^{00} - n^{02} = 1$ και δεύτερο με $n^{00} - n^{02} = -1$, από (2.3), (2.4) παίρνουμε,

$$(i) \mathbf{Q}_1 = 3\mathbf{I}_k + \mathbf{J}_k, \quad \mathbf{Q}_2 = 3\mathbf{I}_k + \mathbf{J}_k, \quad (ii) \mathbf{Q}_1 = 5\mathbf{I}_k - \mathbf{J}_k, \quad \mathbf{Q}_2 = 3\mathbf{I}_k + \mathbf{J}_k$$

Θεώρημα 4.1 Αν $s_1 : x_1 = \dots = x_{k-1} = x, x_k = x + bk, x > b > 0, k \geq 3$,

$s_2 : y_1 = \dots = y_{k-1} = x + 2b, y_k = x + 2b - bk, y_k > 0, k \geq 3$, τότε

(i) Αν $k=2$, οι $s_1 = s_2$ είναι ισοδύναμες (έχουν τις ίδιες ιδιοτιμές), (ii) Αν $3 \leq k < 2 + (x / b)$, τότε η s_1 είναι E,A,D βέλτιστη, συγκρινόμενη με την s_2 .

Η απόδειξη παραλείπεται από έλλειψη χώρου.

Επομένως από τα (i) και (ii) το BA(12,k) στο (i) είναι E,A,D βέλτιστο για $k=2,3,4$.

Για $k=5$ ο σχεδιασμός (ii) έχει ιδιοτιμές ίσες με 0 και διαγράφεται.

Βρίσκουμε το βέλτιστο BA(12,r) με $n_2^1 = 6, n_2^0 = n_2^2 = 3$, το $n_2^{11} = 2, n_2^{11} = 4$.

$$n_2^{11} = 2, n_2^{01} = n_2^{10} = n_2^{21} = n_2^{12} = (6 - 2) / 2 = 2, n_2^{00} + n_2^{02} = n_2^0 - n_2^{01} = 3 - 2 = 1$$

$$\mathbf{Q}_{12} = 2\mathbf{I}_k + \mathbf{J}_k \text{ όταν } n_2^{00} = 1, n_2^{02} = 0, \mathbf{Q}_{12} = 4\mathbf{I}_k - \mathbf{J}_k \text{ όταν } n_2^{00} = 0, n_2^{02} = 1$$

$$\mathbf{Q}_{22} = 6\mathbf{I}_k - \frac{3}{2}\mathbf{J}_k. \text{ Όμως ο } \mathbf{Q}_{12} = 2\mathbf{I}_k + \mathbf{J}_k \text{ είναι E,A,D. Βέλτιστος συγκρινόμενος}$$

με τον $\mathbf{Q}_{12} = 4\mathbf{I}_k - \mathbf{J}_k$ (Θεώρημα 4.1).

$$n_2^{11} = 4, n_2^{01} = n_2^{10} = n_2^{21} = n_2^{12} = (6 - 4) / 2 = 1, n_2^{00} + n_2^{02} = n_2^0 - n_2^{01} = 3 - 1 = 2$$

$$\mathbf{Q}_{12} = 3\mathbf{I}_k \text{ όταν } n_2^{00} = 1, n_2^{02} = 1 \text{ και } \mathbf{Q}_{22} = 3\mathbf{I}_k + \frac{3}{2}\mathbf{J}_k$$

Έχουμε 2 σχεδιασμούς προς σύγκριση,

$$(i) \mathbf{Q}_{12} = 2\mathbf{I}_k + \mathbf{J}_k, \mathbf{Q}_{22} = 6\mathbf{I}_k - \frac{3}{2}\mathbf{J}_k, (ii) \mathbf{Q}_{12} = 3\mathbf{I}_k, \mathbf{Q}_{22} = 3\mathbf{I}_k + \frac{3}{2}\mathbf{J}_k$$

Στον επόμενο πίνακα συμβολίζουμε με $\lambda_{\min}$ τη μικρότερη ιδιοτιμή,

$$\text{με } A = \sum_{i=1}^{2k} (1 / \lambda_i) \text{ και } D = \prod_{i=1}^{2k} \lambda_i.$$

Πίνακας 1: Βέλτιστο BA(12,r)

BA	r	Ιδιοτιμές	$\lambda_{\min}$	A	D
(i)	1	3,9/2	3	0.555	13.5
(ii)	1	3, 9/2	3	0.555	13.5
(i)	2	2,4,6,3	2	1.25	144
(ii)	2	3,3,3,6	3	1.167	162
(i)	3	2,2,5,6,6,3/2	3/2	2.2	1080
(ii)	3	3,3,3,3,3,15/2	3	1.8	1822.5

Επομένως ο βέλτιστος σχεδιασμός για BA(12,r) είναι: Για r=1 οι δύο είναι ισοδύναμοι, για r=2,3 ο βέλτιστος σχεδιασμός είναι το (ii) που έχει $n_2^{11} = 4$.

Τέλος βρίσκουμε το βέλτιστο σχεδιασμό PBA(12,k-r,r)

Στον πίνακα 2 και 2 το 5_3 είναι συντομογραφία του 5,5,5.

Πίνακας 2: Βέλτιστο BA(12,r)

PBA(12, k-r,r)	k	r	Ιδιοτιμές	$\lambda_{\min}$	A	D
BA(12,2)	2	0	3,5,3,5	3	1.067	225
PBA(12,1,1)	2	1	4,4,3,9/2	3	1.055	216
BA(12,3)	3	0	$3_4, 6_2$	2	1.667	2916
PBA(12,2,1)	3	1	$3_3, 5_2, 9 / 2$	3	1.622	3037.5
PBA(12,1,2)	3	2	$3_3, 4_2, 6_1$	3	1.667	2592
BA(12,4)	4	0	$3_6, 7_2$	3	2.285	35721
PBA(12,3,1)	4	1	$3_5, 6_2, 9 / 2$	3	2.222	39366
PBA(12,2,2)	4	2	$3_5, 5_2, 6_1$	3	2.233	36450
PBA(12,1,3)	4	3	$3_5, 5_2, 15 / 2$	3	2.400	29160
BA(12,5)	5	0	$3_8, 8_2$	3	2.917	4199040

PBA(12,4,1)	5	1	$3_7, 7_2, 9/2$	3	2.841	482233.5
PBA(12,3,2)	5	2	$3_7, 6_3$	3	2.833	472392.0
PBA(12,2,3)	5	3	$3_7, 5_2, 15/2$	3	2.867	410062.5
PBA(12,1,4)	5	4	$4_2, 3_7, 9$	3	2.944	314928.0

k=2: E,D βέλτιστος σχεδιασμός είναι BA(12,2) με $\mathbf{Q}_1 = \mathbf{Q}_2 = 3\mathbf{I}_2 + \mathbf{J}_2$ και

d: 0 0 0 0 1 1 1 1 2 2 2 2, 0 0 1 2 0 2 1 1 1 0 2 2

E,A Βέλτιστος σχεδιασμός είναι PBA(12,1,1) με $\mathbf{Q}_{11} = \mathbf{Q}_{21} = 4$, $\mathbf{Q}_{12} = 3$, $\mathbf{Q}_{22} = 9/2$ και d: 0 0 0 0 1 1 1 1 2 2 2 2, 1 1 0 2 1 1 0 2 1 1 0 2,

k=3: E,A,D βέλτιστος σχεδιασμός είναι PBA(12,2,1) με $\mathbf{Q}_{11} = \mathbf{Q}_{21} = 3\mathbf{I}_2 + \mathbf{J}_2$, $\mathbf{Q}_{12} = 3$, $\mathbf{Q}_{22} = 9/2$ και είναι OA(9,3,3,2) σύν τρεις γραμμές 001,111,221,

d: 0 0 0 0 1 1 1 1 2 2 2 2, 0 0 1 2 1 1 0 2 0 1 2 1, 1 1 0 2 1 1 2 0 0 2 1 1

k=4: E,A,D βέλτιστος σχεδιασμός είναι PBA(12,3,1) με $\mathbf{Q}_{11} = \mathbf{Q}_{21} = 3\mathbf{I}_3 + \mathbf{J}_3$, $\mathbf{Q}_{12} = 3$, $\mathbf{Q}_{22} = 9/2$. Αυτό είναι ένα OA(9,4,3,2) επαυξημένο κατά τρεις γραμμές: 0001,1111,2221. Ο σχεδιασμός είναι d: 0 1 2 0 0 1 1 2 2 0 1 2,

0 1 2 0 2 0 1 1 2 1 2 0, 0 1 2 1 2 0 2 1 0 0 1 2, 0 0 0 1 1 1 1 1 1 2 2 2.

k=5: E,A,D βέλτιστος σχεδιασμός είναι PBA(12,4,1) με $\mathbf{Q}_{11} = \mathbf{Q}_{21} = 3\mathbf{I}_4 + \mathbf{J}_4$, $\mathbf{Q}_{12} = 3$, $\mathbf{Q}_{22} = 9/2$ κατασκευάστηκε με αλγόριθμο διότι δεν υπάρχει OA(9,5,3,2). Ο άλλος βέλτιστος σχεδιασμός που είναι PBA(12,3,2) δεν υπάρχει, όπως μας έδειξε η εξέταση όλων των περιπτώσεων. Ο σχεδιασμός είναι

d: 0 1 2 0 0 1 1 2 2 0 1 2, 0 1 2 0 2 0 1 1 2 1 2 0, 0 1 2 1 2 0 2 1 0 0 1 2,

1 2 0 0 2 0 1 1 2 0 1 2, 0 0 0 1 1 1 1 1 1 2 2 2.

ABSTRACT

Optimal designs are studied for the estimation of standardized linear and quadratic contrasts, when there are k factors, each at three levels in a linear model. The concept of majorization is applied. If $N \equiv 0 \pmod{9}$ the OA(N,k,3,2) is Φ optimal. If $N \equiv 3 \pmod{9}$ or $N \equiv 6 \pmod{9}$, balanced arrays and partial balanced arrays are introduced and E,A,D optimal designs are constructed. The case $N=12$ is studied in detail giving the optimal designs when the number of factors is 2,3,4,5. Algorithms are employed for the construction of optimal designs.

ΑΝΑΦΟΡΕΣ

- Kiefer, J. (1958). On the nonrandomized optimality and randomized nonoptimality of symmetrical designs. *Ann. Math. Statist.*, **29**, 675-600.
- Kolyva-Machera, F. (1989a). D-Optimality in 3^k Designs for $N \equiv 1 \pmod{9}$ observations. *J. Statist. Plann. Inference*, **22**, 95-103.
- Kolyva-Machera, F. (1989b) Fractional factorial designs and G-optimality, In: *Proceedings, forth Prague symposium on asymptotic statistics*. Charles University Press, Prague, 349-358.
- Marshall, A., and Olkin, I. (1979). *Inequalities: Theory of majorization and its applications*, New York. Academic press.
- Mukerjee, R.(1999). On the optimality of orthogonal arrays plus one run plans. *Ann. Statist.* **27**, 1256-1271.
- Pukelsheim, F. (1993). *Optimal Design of Experiments*, John Wiley and Sons, Inc.
- Περικλέους Κ., Χατζόπουλος Σ., Κολυβά-Μαχαίρα Φ., Κουνιάς Σ. (2012). Πρακτικά 25^{ου} Πανελληνίου Συνεδρίου Στατιστικής, σελ. 179-187.

ΠΑΡΑΓΟΝΤΕΣ ΠΟΥ ΕΠΗΡΕΑΖΟΥΝ ΤΗΝ ΛΙΑΝΙΚΗ ΤΙΜΗ ΑΜΟΛΥΒΔΗΣ ΒΕΝΖΙΝΗΣ 95 ΟΚΤΑΝΙΩΝ: ΑΠΡ. 2011-ΔΕΚ. 2012

Σωτήριος Πέτρος¹, Πρόδρομος Προδρομίδης^{2,3}

¹ Υπουργείο Ανάπτυξης και Ανταγωνιστικότητας

spetros@gge.gr

² Κέντρο Προγραμματισμού και Οικονομικών Ερευνών

pjprodr@kepe.gr

ΠΕΡΙΛΗΨΗ

Περιγράφονται οι τιμές αμόλυβδης βενζίνης που καταχωρηθήκαν μεταξύ Απριλίου 2011-Δεκεμβρίου 2012 στο Παρατηρητήριο Υγρών Καυσίμων (ΠΥΚ) του Υπουργείου Ανάπτυξης από πρατήρια ανά την επικράτεια. Εκτιμώνται οικονομετρικά η τάση, η εποχικότητα, οι επιδράσεις της ημέρας, της τιμής διυλιστηρίου, της περιοχής της χώρας, καθώς και του αριθμού πρατηρίων των δραστηριοποιούμενων εταιρειών εμπορίας και των απεργιών στα μέσα μεταφοράς. Εν συνεχεία, με ελέγχους αιτιότητας (συστηματικής προήγησης) τιμών διερευνάται μια χρήσιμη διάσταση του ανταγωνισμού μεταξύ των εταιρειών εμπορίας σε μεγάλους δήμους της χώρας.

Λέξεις Κλειδιά: τιμή βενζίνης, τοπικές ιδιαιτερότητες, ανταγωνισμός, απεργίες.

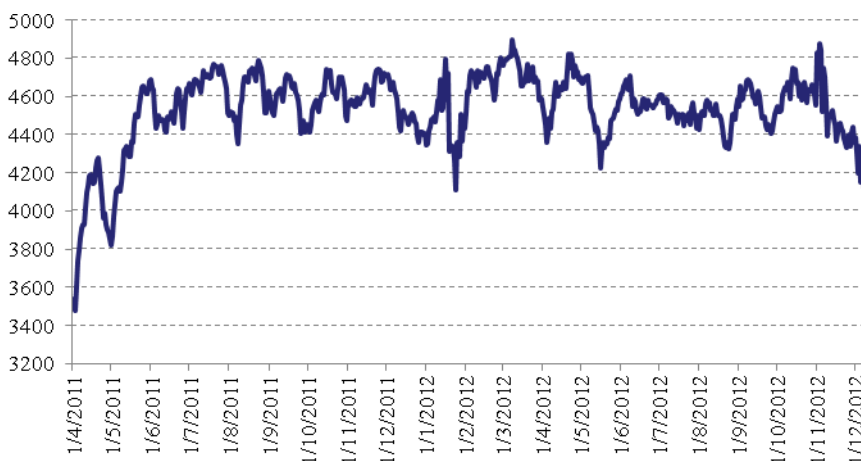
1. ΕΙΣΑΓΩΓΗ

Στην Ελλάδα η ζήτηση βενζίνης εξυπηρετείται από 18 εταιρείες εμπορίας (αλυσίδες πρατηρίων), καθώς και ανεξάρτητα πρατήρια που την προμηθεύονται από δύο ομίλους διυλίσεως. Ο ευρέως χρησιμοποιούμενος τύπος βενζίνης είναι των 95 οκτανίων. Στις επόμενες σελίδες διερευνούμε τους παράγοντες που επηρεάζουν την μέση λιανική τιμή ανά δήμο, καθώς και την λειτουργία του ανταγωνισμού σε μεγάλες τοπικές αγορές της χώρας. Προς τούτο αξιοποιούμε τα στοιχεία τιμών που εκαταχώρησαν πρατήρια ανά την επικράτεια στο ΠΥΚ (www.fuelprices.gr) μεταξύ 1^{ης} Απριλίου 2011 (οπότε η συμμετοχή των πρατηρίων στο ΠΥΚ ξεπέρασε το 50% των λειτουργούντων πρατηρίων) και 31^{ης} Δεκεμβρίου 2012. Όπως φαίνεται από το Διάγραμμα 1, κατά το εν λόγω χρονικό διάστημα ο αριθμός των συμμετεχόντων πρατηρίων ανήλθε από 3.536 πρατήρια σε 4.189.

2. ΠΕΡΙΓΡΑΦΗ ΤΩΝ ΣΤΟΙΧΕΙΩΝ

Η βάση δεδομένων περιέχει 1.250.122 ημερήσιες παρατηρήσεις υπό την μορφή μοναδικών τιμοληψιών ανά 24ώρο ή 193.949 ημερήσιες εκτιμήσεις μέσω τιμών σε επίπεδο δήμων. Εξυπακούεται ότι οι παρατηρήσεις θα ήσαν πολλαπλάσιες αν υπολογίζοντο σε ωριαία βάση. Εν τούτοις, η προκαταρκτική διερεύνηση των στοιχείων έδειξε ότι δεν υπάρχουν πολλαπλές αλλαγές τιμών εντός της ημέρας εκ μέρους των συμμετεχόντων πρατηρίων.

Διάγραμμα 1: Εξέλιξη του συνολικού αριθμού πρατηρίων βάσει των ενημερώσεων που προωθούν στο ΠΥΚ



Σημ.: Η μικρότερη συχνότητα καταγράφεται την 3^η Απριλίου 2011 (3.481 πρατήρια).

Από τα καταχωρηθέντα πρατήρια τα μισά (το 50% των πρατηρίων κατά το 2011, το 49% κατά το 2012) εμπορεύονται υπό τα εμπορικά σήματα των ΕΚΟ και ΒΡ του ομίλου διυλίσεως των ΕΛΠΕ ή τα εμπορικά σήματα της Shell, Avin, Cyclon του ομίλου διυλίσεως της ΜΟΗ. Τα υπόλοιπα πρατήρια προμηθεύονται καύσιμα ανάλογα με την περίπτωση από τους προαναφερομένους ομίλους. Εκ των εταιρειών εμπορίας:

- Η Shell στην αρχή (τέλος) της περιόδου διατηρούσε σημαντική παρουσία, άνω του 10% των πρατηρίων, σε ένδεκα (εννέα) από τις 13 διοικητικές περιφέρειες της χώρας: Συγκεκριμένα σε όλες πλην του Β. Αιγαίου και της Στ. Ελλάδος-Ευβοίας (Β. Αιγαίου, Κ. Μακεδονίας, Θεσσαλίας, νοτίου και ανατολικής Πελοποννήσου).
- Η ΕΚΟ διατηρούσε αντίστοιχα σημαντική παρουσία σε εννέα περιφέρειες (και αναλογικά μειωμένη παρουσία κατά μήκος της Δ. Ελλάδος, Ηπείρου, Δ. και Κ. Μακεδονίας).
- Η ΒΡ διατηρούσε ανάλογη παρουσία σε εννέα περιφέρειες (και συγκριτικά μειωμένη παρουσία σε Κ. και Δ. Μακεδονία, Ν. Αιγαίο, Ιονίους νήσους).
- Η Elinoil διατηρούσε σημαντική παρουσία αρχικώς σε τέσσερις και τελικώς σε επτά περιφέρειες: την Δ. Μακεδονία, την Α. Μακεδονία-Θράκη, το Β. και Ν. Αι-

γαίο στην πρώτη περίπτωση· τις προηγούμενες, καθώς και την Θεσσαλία, την Στ. Ελλάδα-Εύβοια και τις Ιονίους νήσους στην δεύτερη περίπτωση.

- Η Jetoil διατηρούσε σημαντική παρουσία σε Ιονίους νήσους, Β. και Ν. Αιγαίο.
- Η Aegean διατηρούσε σημαντική παρουσία σε Κ. Μακεδονία και Α. Μακεδονία και Θράκη, τόσο στην αρχή όσο και στο τέλος της υπό εξέταση περιόδου.
- Η Avin διατηρούσε σημαντική παρουσία αρχικώς στην νότιο, κεντρική και ανατολική Πελοπόννησο και προς το τέλος της περιόδου τόσο στην εκεί όσο και την Ήπειρο.

Τέλος, οι ανεξάρτητοι πρατηριούχοι (ΑΠ), οι οποίοι αρχικώς διατηρούσαν σημαντική παρουσία στην Θεσσαλία, προς το τέλος της περιόδου εμφανίζουν ανάλογα ποσοστά και στις όμορες περιφέρειες της Δ. Ελλάδος και Δ. Μακεδονίας.

3. ΕΜΠΕΙΡΙΚΗ ΑΝΑΛΥΣΗ ΤΗΣ ΜΕΣΗΣ ΤΙΜΗΣ ΑΝΑ ΔΗΜΟ

Στην διεθνή βιβλιογραφία το ύψος της λιανικής τιμής της βενζίνης, καθώς και οι όποιες διαφορές στην τιμή (α) αποδίδονται σε ιδιαιτερότητες των εταιρειών εμπορίας, την τάση και την εποχικότητα, την περιοχή (ή το μέσο εισόδημα στην περιοχή), την αναλογία ανεξαρτήτων προς συνασπισμένων πρατηρίων (αλυσίδων πρατηρίων), την τιμή του διυλιστηρίου και την μεταβολή στον φόρο, τις παρελθούσες λιανικές τιμές, τον αριθμό και τις τιμές των ανταγωνιστών στην περιοχή ή/και (β) ερμηνεύονται από την πυκνότητα του πληθυσμού (εναλλακτικά, αν η περιοχή είναι αστική ή αγροτική), τον αριθμό των πρατηρίων ανά κάτοικο, τα χαρακτηριστικά της οδού επί της οποίας ευρίσκεται το πρατήριο, τον τύπο εξυπηρέτησεως που παρέχει, το μέγεθος και την παλαιότητά του, την απόστασή του από ανταγωνιστές ή/και το διυλιστήριο, την ημέρα της εβδομάδος (καθημερινή, σχόλη, εορτή ή διακοπές) και την ώρα της ημέρας. (Βλ. ενδεικτικά Slade, 1992· Eckert και West, 2004· Pennerstorfer, 2009· Davis, 2010).

Εν προκειμένω τα διαθέσιμα στοιχεία επιτρέπουν την εμπειρική ανάλυση των μέσων τιμών αμόλυβδης βενζίνης σε επίπεδο δήμου με την μέθοδο των ελαχίστων τετραγώνων και, κατ' επέκτασιν, τον υπολογισμό (ή την απομόνωση) των επιδράσεων:

- της τάσεως, της εποχικότητας και της ημέρας,
- της (μετά φόρων) τιμής του διυλιστηρίου, του αριθμού των πρατηρίων των εταιρειών που δραστηριοποιούνται σε μια περιοχή,
- των χωρικών ιδιαιτεροτήτων, των απεργιών σε άλλα μέσα μεταφοράς κ.ά.

Ειδικά ως προς την χωρική διάσταση, διαπιστώνεται ότι το υπόδειγμα που βασίζεται στην νομαρχιακή οργάνωση των στοιχείων (οργάνωση βάσει της οποίας, σύμφωνα με το ΠΥΚ, δομείται η εξαρτημένη μεταβλητή) ερμηνεύει μικρότερο μέρος της συνολικής μεταβλητότητας του υποδείγματος ($R^2=81,4\%$, με 53 χωρικές ψευδομεταβλητές) σε σχέση με την παραλλαγή που αποδέχεται/επιτρέπει μεγαλύτερη ετερογένεια στον χώρο ($R^2=92,6\%$, με 25 χωρικές ψευδομεταβλητές). Η ερμηνευτική ικανότης της δευτέρας εκδοχής είναι μάλλον υψηλή. Βλ. Πίνακα 1. Την εκδοχή αυτή παραθέτουμε και συζητούμε στην συνέχεια. Σημειώνεται ότι η εκτίμηση είναι με ανθεκτικούς (*robust*) εκτιμητές.

Συνάγεται ότι η αύξηση της τιμής του διυλιστηρίου μετακυλύεται σχεδόν ολόκληρη (94,7%) στον τελικό καταναλωτή. Επιπλέον, το μέσο περιθώριο εμπορίας στην Αθήνα είναι της τάξεως των 18 λεπτών ανά λίτρο, ενώ η τιμή αυξάνει διαχρονικά με φθίνοντα ρυθμό. Υπάρχει εποχικότητα (η τιμή είναι χαμηλότερη από τα μέσα Δεκεμβρίου ως τις αρχές Απριλίου και από τις αρχές Ιουλίου ως τα μέσα Σεπτεμβρίου) και

Πίνακας 1: Εκτίμηση με ανθεκτικούς (robust) εκτιμητές των παραγόντων που επηρεάζουν την τιμή της βενζίνης (σε λεπτά του ευρώ ανά λίτρο)

Ερμηνευτικές μεταβλητές	εκτιμώμενοι συντελεστές	p values
1. Σταθερός όρος	17,74	0,000
2. Τιμή των δύο διυλιστηρίων (μετά φόρων)	94,69	0,000
3. Χρόνος (τάση)	0,00	0,000 ¹
4. Χρόνος στο τετράγωνο (για υπολογισμό ρυθμού μεταβολής) ^α	-0,00	0,000
Εποχικότητα		
5. Μέσα Δεκεμβρίου – αρχές Απριλίου (περίοδος αναφοράς)		
6. Μέσα Απριλίου – τέλη Ιουνίου	1,97	0,000
7. Αρχές Ιουλίου – μέσα Σεπτεμβρίου	0,34	0,000
8. Μέσα Σεπτεμβρίου – μέσα Δεκεμβρίου	1,47	0,000
Επίδραση της ημέρας		
9. Τετάρτη ή Πέμπτη ημέρα της εβδομάδος	-0,02	0,078
Χωρική διάσταση		
10. Αθήνα, Βύρων, Καισαριανή, Ζωγράφου (περιοχή αναφοράς)		
11. Θέρμη (Θεσσαλονίκης)	-6,87	0,000
12. Πέραμα, Χαϊδάρι, Αγ.Βαρβάρα	-6,28	0,000
13. Άλλες περιοχές γύρω από Αθήνα, Θεσσαλονίκη, άλλες πόλεις και κόμβοι ηπειρωτικής χώρας ^β	-4,18	0,000
14. Απόμακρες περιοχές ηπειρωτικής χώρας και κοντινών νήσων ^γ	1,93	0,000
15. Λοιπή ηπειρωτική χώρα και κοντινοί νήσοι (Σαλαμίς, Λευκάς, Ζάκυνθος, Ελαφώνησος)	-1,26	0,000
16. Πόλεις Ηρακλείου και Χανίων	2,02	0,000
17. Κεντρικές περιοχές της Κρήτης ^δ	4,71	0,000
18. Ανατολικές και δυτικές περιοχές της Κρήτης ^ε	7,14	0,000

Σημειώσεις

^α Άρα η συνάρτηση είναι αόζουσα στρέφουσα τα κοίλα προς τα κάτω. ^β Η Αττική πλην Μεγάρων, ο ν. Θεσσαλονίκης πλην Βόλβης, οι ν.Ημαθίας και Κιλκίς, οι δήμοι Αλμωπί-ας, Κατερίνης, Πέλλης, Πύδνας-Κολινδρόν, Πατραίων και Δ.Αχαΐας, Βόλου, Ρ.Φεραι-ον, Λαρίσης και Τυρνάβου, Τρικαίων, Καρδίτσας, Ιωαννιτών, Πρεβέζης, Σερρών και Ε. Παππά, Δράμας, Ξάνθης, Λαμίας και Μακρακώμης, Καλαμών, Ναυπλίου, Βέλους-Βό-χας. ^γ Οι δήμοι Ορεστιάδος, Διδυμοτειχείου, Σουφλίου, Αρριανών, Μύκης, Κ.Νευροκοπίου, Πωγωνίου, Δωδώνης, Μετσόβου, Δεσκάτης, Α. Πλαστήρα, Αγράφων, Αμφιλοχίας, Θέρμου, Καρπενησίας, Δωρίδος, Μεγανησίου, Καλαβρύτων, Πύλου-Νέστορος, Μάνης (ανατολικής και δυτικής), Ελαφονήσου, Κυνουρίας (βορείου και νοτίου), Τροιζηνίας, Καρύστου, Ν.Πηλίου, Ζαγοράς-Μουρεσίου, Αγιάς. ^δ Οι δήμοι Αποκορώνου, Πλατανιά, Αγ.Βασιλείου, Ανωγειών, Αμαρίου, Μυλοποτάμου, Ρεθύμνης, Αρχάνων-Αστερουσίου, Γόρτυνας, Μαλεβιζίου, Μίνωα-Πεδιάδος, Φαιστού, Χερσονήσου. ^ε Οι δήμοι Κανδάνου-Σελίνου, Κισσάμου, Σφακίων, Αγ.Νικολάου, Ιεραπέτρης, Οροπεδίου, Σητείας.

Πίνακας 1 (συνέχεια)

<i>Ερμηνευτικές μεταβλητές</i>	<i>εκτιμώμενοι συντελεστές</i>	<i>p values</i>
Χωρική διάσταση		
19. Βιάννος (Κρήτης)	11,75	0,000
20. Νήσος Κέρκυρα	3,09	0,000
21. Νήσος Κεφαλληνία	6,38	0,000
22. Νήσοι Ιθάκη και Παξοί	11,46	0,000
23. Νήσοι Αίγινα, Αγκίστρι, Σπέτσες, Κέα, Άνδρος, Τήνος, Σύρος	6,67	0,000
24. Νήσοι Πάρος, Νάξος, Αντίπαρος	9,29	0,000
25. Ζώνη νήσων νοτίου Αιγαίου (Κύθηρα, λοιπές Κυκλάδες πλην Σικίνου και Αμοργού, Αστυπάλαια, Κάλυμνος, Λέρος)	14,55	0,000
26. Νήσοι βορείως της ζώνης (Αμοργός, Πάτμος, Ικαρία)	17,16	0,000
27. Νήσοι νοτίως της ζώνης (Κάρπαθος, Τήλος)	18,45	0,000
28. Νήσοι Σίκινος και Φούρνοι	21,38	0,000
29. Νήσοι Λέσβος, Ρόδος, Χίος, Θάσος	7,37	0,000
30. Μεσαίου μεγέθους ακριτικοί νήσοι κεντρικού και νοτίου Αιγαίου (Σάμος, Κως)	10,91	0,000
31. Μικροί ακριτικοί νήσοι κεντρικού και νοτίου Αιγαίου (Λειψοί, Σύμη)	13,05	0,000
32. Μεσαίου μεγέθους ακριτικοί νήσοι βορείου Αιγαίου (Λήμνος, Σαμοθράκη)	15,02	0,000
33. Νήσος Σκόπελος	16,12	0,000
34. Νήσοι Αλόνησος, Σκιάθος, Σκύρος	20,09	0,000
35. Νήσος Αγ. Ευστράτιος	26,95	0,000
Εμπορική – ανταγωνιστική διάσταση: αριθμός πρατηρίων των εταιρειών διανομής στην περιοχή		
36. Sunoil	-0,91	0,000
37. Medoil	-0,16	0,000
38. Aegean	-0,02	0,000
39. Elinoil	-0,02	0,000
40. EKO	-0,01	0,000
41. ETEKA	0,00	0,668
42. Ανεξάρτητοι πρατηριούχοι	0,01	0,002
43. Shell	0,01	0,000
44. Silkoil	0,02	0,000
45. Jetoil	0,02	0,000
46. Revoil	0,03	0,000
47. Αργώ	0,03	0,000
48. BP	0,03	0,000
49. Avin	0,04	0,000
50. Cyclon	0,05	0,000
51. Kaoil	0,05	0,000
52. Galonoil	0,14	0,066

Πίνακας 1 (συνέχεια)

<i>Ερμηνευτικές μεταβλητές</i>	<i>εκτιμώμενοι συντελεστές</i>	<i>p values</i>
Εμπορική – ανταγωνιστική διάσταση		
53. Dracoil	0,15	0,000
54. KMoil	0,21	0,000
55. El Petroil	0,38	0,000
Αριθμός ημερών απεργίας σε		
56. Μισθωμένα οχήματα (ταξί) ^ς (34 ισοδύναμα 24ώρου)	0,26	0,000
57. Ακτοπλοία ^{ς,ζ} (23 ισοδύναμα 24ώρου)	-0,21	0,000
58. Προαστιακό σιδηρόδρομο της Αττικής ^{ς,ζ} (23 ισοδ. 24ώρου)	0,39	0,000
59. Υπόγειο σιδηρόδρομο της Αττικής (μετρό) ^{ς,ζ} (25 ισοδ. 24ώρου)	0,11	0,000
60. Χρονοσειρά των καταλοίπων με μια χρονική υστέρηση ^η	1,99	0,000

Αριθμός παρατηρήσεων: 193.656. Ερμηνευτική ικανότης υποδείγματος: $R^2 = 92,55\%$.

Σημειώσεις

^ς Άνευ των επιδράσεων των ερμηνευτικών μεταβλητών υπ' αριθμ. 2-9 που ενεργάνιζαν συσχέτιση της τάξεως του 15-25%.

^ζ Άνευ της επιδράσεως των απεργιών στα προαναφερόμενα μέσα μεταφοράς.

^η Χρησιμοποιήθηκε προκειμένου να διορθωθεί η αυτοσυσχέτιση στην σειρά της εξαρτη-μένης μεταβλητής.

ενδείξεις συστηματικών αυξομειώσεων μέσα στην εβδομάδα: ενδέχεται οι τιμές να είναι χαμηλότερες Τετάρτη ή Πέμπτη. Οι διαφοροποιήσεις στον χώρο δεν ακολουθούν την διοικητική οργάνωση της χώρας καθότι εντός των νομών και περιφερειών υπάρχει αισθητή, στατιστικά σημαντική ετερογένεια. Οι νησιωτικές περιοχές (παρά τον μειωμένο ΦΠΑ), καθώς και οι δύσβατες/απόμακρες ηπειρωτικές περιοχές είναι ακριβότερες. Οι απεργίες σε μέσα μεταφοράς όπως το ταξί, ο προαστιακός, το μετρό αυξάνει την ανάγκη χρήσεως οχημάτων ιδιωτικής χρήσεως (ΙΧ) ωθώντας την τιμή της βενζίνης προς τα άνω. Από την άλλη, οι απεργίες στην ακτοπλοία φαίνεται να αποθαρρύνουν την χρήση οχημάτων ΙΧ επιφέροντας μείωση της τιμής. Η μεταβολή της τιμής δεν φαίνεται να εξαρτάται τόσο από τον αριθμό των πρατηρίων σε κάθε περιοχή όσο από την εταιρεία εμπορίας. Μεταξύ των τριών μεγάλων εταιρειών τα πρατήρια της ΕΚΟ είναι εν γένει φθηνότερα, της Shell ακριβότερα και της BP ακόμη πιο ακριβά.

4. ΕΝΔΕΙΞΕΙΣ ΣΥΣΤΗΜΑΤΙΚΩΝ ΠΡΩΤΟΒΟΥΛΙΩΝ ΣΤΙΣ ΑΛΛΑΓΕΣ ΤΙΜΩΝ ΑΠΟ ΚΑΠΟΙΕΣ ΕΤΑΙΡΕΙΕΣ

Εν συνεχεία επιχειρούμε να διαμορφώσουμε άποψη για την λειτουργία του ανταγωνισμού. Από τα οικονομετρικά εργαλεία που χρησιμοποιούνται σε τέτοιες αναλύσεις (Bishop and Walker, 2002), οι διαθέσιμες μεταβλητές επιτρέπουν τον έλεγχο αιτιό-

τητας κατά Granger (*Granger causality test*). Μέσω αυτού ερευνούμε αν η πορεία των καταχωρημένων τιμών υποδηλώνει συστηματικής κατευθύνσεως επιδράσεις μεταξύ των εταιρειών εμπορίας (ή αλυσίδων πρατηρίων) και, κατ' επέκτασιν, την ύπαρξη εναρμονισμένων πρακτικών. Με άλλα λόγια, μπορούμε να εξετάσουμε αν η σημερινή τιμή, Y , ή μεταβολή τιμής, ΔY , του πωλητού (δηλ της εταιρείας εμπορίας), εξαρτάται —πέραν της παρελθούσης τιμής (ή της μεταβολής τιμής) του ιδίου πωλητού— από την παρελθούσα τιμή, X , ή την μεταβολή τιμής, ΔX , κάποιου άλλου πωλητού ή το αντίστροφο. Προς τούτο προβαίνουμε σε μια παλινδρόμηση για κάθε πιθανό ζεύγος πωλητών. Διευκρινίζουμε ότι επειδή η ανάλυση βάσει της τιμής προσκρούει στην προϋπόθεση των στασίμων χρονοσειρών, θα βασιστούμε στην μεταβολή της τιμής (δηλ. τις πρώτες διαφορές μεταξύ τιμών). Σε όρους της συντομογραφικής γραφής που συνηθίζεται σε τέτοιες περιπτώσεις, για τον έλεγχο κάθε ζεύγους ορίζουμε δύο εξισώσεις. Στην πρώτη ελέγχουμε εάν η υστέρηση (ή υστερήσεις) του στοιχείου ΔX επηρεάζει (επηρεάζουν) το στοιχείο ΔY και στην δεύτερη εάν η υστέρηση (ή υστερήσεις) του ΔY επηρεάζει (επηρεάζουν) το ΔX . Στην περίπτωση δύο υστερήσεων οι εξισώσεις περιγράφονται ως εξής :

$$\Delta Y_t = b_0 + b_1 * \Delta Y_{t-1} + b_2 * \Delta Y_{t-2} + \theta_1 * \Delta X_{t-1} + \theta_2 * \Delta X_{t-2} + e_t \quad \text{και} \quad (1)$$

$$\Delta X_t = \beta_0 + \beta_1 * \Delta X_{t-1} + \beta_2 * \Delta X_{t-2} + \varphi_1 * \Delta Y_{t-1} + \varphi_2 * \Delta Y_{t-2} + u_t, \quad (2)$$

όπου οι όροι b , β , θ και φ αντιστοιχούν σε συντελεστές, ο όρος t στην μονάδα του χρόνου (συγκεκριμένα, τις ημέρες), οι μεταβλητές e και u σε διαταρακτικούς όρους. Η διαπίστωση της στατιστικής σημαντικότητας των μεγεθών θ και φ γίνεται με τον έλεγχο Wald F.

Προκειμένου να βελτιώσουμε την ποιότητα των αποτελεσμάτων που προκύπτουν από τις παλινδρομήσεις ανά ζεύγη, προσφεύγουμε και στην πολυμεταβλητή εκδοχή, το λεγόμενο *Vector Autoregressive (VAR) model* του ελέγχου αιτιότητας κατά Granger. Τουτέστιν, αντί του στοιχείου ενός μόνον άλλου πωλητού (ή μιας άλλης μόνον αλυσίδας πρατηρίων) σε κάθε παλινδρόμηση, στην παλινδρόμηση παρατίθενται (ενσωματώνονται) και τα αντίστοιχα στοιχεία όλων των άλλων πωλητών. Στις περιπτώσεις των αναλύσεων των τιμών που παρατηρούνται στους δήμους Αθηναίων και Πειραιώς, που θα παραθέσουμε στην συνέχεια, ο έλεγχος για την επιλογή του αριθμού των χρονικών υστερήσεων (*VAR lag order selection criteria*) υποδεικνύει την επιλογή μίας υστερήσεως σύμφωνα με τα κριτήρια κατά LR (*sequential modified Likelihood Ratio test statistic* με επίπεδο σημαντικότητας 5%), FPE (*Final prediction error*) και AIC (*Akaike information criterion*). Στις περιπτώσεις των δήμων Θεσσαλονίκης και Ηρακλείου υποδεικνύει δύο υστερήσεις. Στην περίπτωση του δήμου Πατραίων τρεις. Στην περίπτωση του δήμου Λαρίσης πέντε. Κατ' αναλογία προς τις εξισώσεις (1) και (2), ο VAR έλεγχος κατά Granger στις πρώτες διαφορές (ή μεταβολές των τιμών) μπορεί να διατυπωθεί ως εξής:

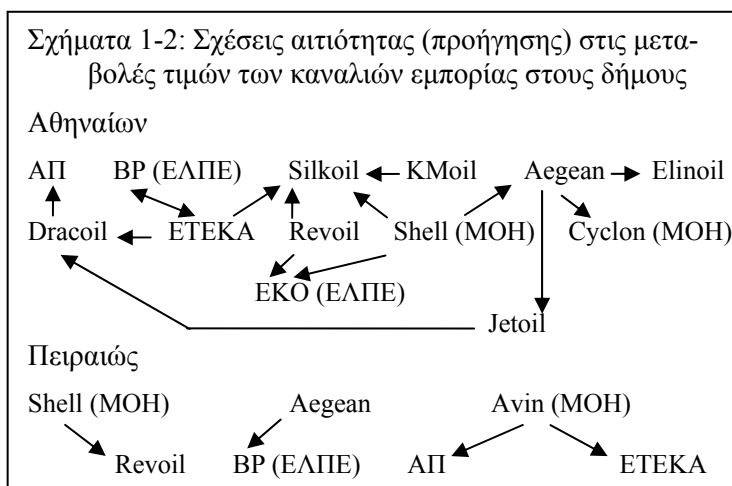
$$\Delta Y_t = b_0 + b_1 \Delta Y_{t-1} + b_2 \Delta Y_{t-2} + \sum_{j=1}^k \theta_{1j} \Delta X_{j,t-1} + \sum_{j=1}^k \theta_{2j} \Delta X_{j,t-2} + e_t \text{ και} \quad (3)$$

$$\Delta X_t = \beta_0 + \beta_1 \Delta X_{t-1} + \beta_2 \Delta Y_{t-2} + \sum_{j=1}^k \phi_{1j} \Delta Y_{j,t-1} + \sum_{j=1}^k \phi_{2j} \Delta Y_{j,t-2} + u_t. \quad (4)$$

όπου k ο αριθμός των εταιρειών εμπορίας. Εν προκειμένω η διαπίστωση της στατιστικής σημαντικότητας της αλληλεξαρτήσεως των εταιρειών γίνεται με τον έλεγχο Wald F όπου εξετάζεται αν $\theta_{11} = \theta_{12} = \dots = \theta_{1j} = \phi_{11} = \phi_{12} = \dots = \phi_{1j} = 0$ για $j=1$ έως k .

Ας περάσουμε στα ευρήματα. Στην πρωτεύουσα δραστηριοποιούνται δώδεκα εταιρείες (κανάλια) εμπορίας βενζίνης, καθώς και ΑΠ που για τις ανάγκες της ανάλυσεως ομαδοποιούνται σε επιπρόσθετο κανάλι εμπορίας. Από την πρώτη εκδοχή του ελέγχου αιτιότητας (δηλ. τους ελέγχους κατά ζεύγη) προκύπτουν 17 μονόπλευρες (μονής κατευθύνσεως) σχέσεις, καθώς και μία αμφίπλευρη (μεταξύ των εταιρειών ΕΤΕΚΑ και BP) σε επίπεδο στατιστικής σημαντικότητας 1%· ενώ από τα αποτελέσματα της δευτέρας εκδοχής του ελέγχου στο ίδιο επίπεδο σημαντικότητας, προκύπτει η κατάσταση στις τιμολογιακές κινήσεις των καναλιών εμπορίας που αποτυπώνεται στο Σχήμα 1. (Δεν υφίσταται σχέση του πολυμεταβλητού υποδείγματος που δεν επιβεβαιώνεται από το απλό υπόδειγμα.) Προκύπτει ότι οι εταιρείες που (α) προηγούνται στην διαμόρφωση τιμών είναι οι Shell, Revoil, Kmoil· (β) κάποτε προηγούνται και άλλοτε δέχονται επιδράσεις είναι οι BP, Jetoil, Aegean, ΕΤΕΚΑ και Dracoil· (γ) έπονται είναι οι ΕΚΟ, Elinoil, Silkoil, Cyclon, καθώς και οι ΑΠ.

Στον Πειραιά δραστηριοποιούνται επτά εταιρείες εμπορίας, καθώς και ΑΠ. Από τον κοινό χώρο αποτελεσμάτων των δύο ελέγχων αναδύεται η εικόνα που αποτυπώνεται στο Σχήμα 2. Προκύπτει ότι: (α) τρία κανάλια εμπορίας προηγούνται στην διαμόρφωση τιμών (Shell, Aegean, Avin)· (β) τέσσερα έπονται (BP, Revoil, ΕΤΕΚΑ, ΑΠ)· (γ) ένα κανάλι κινείται ανεξάρτητα (ΕΚΟ).



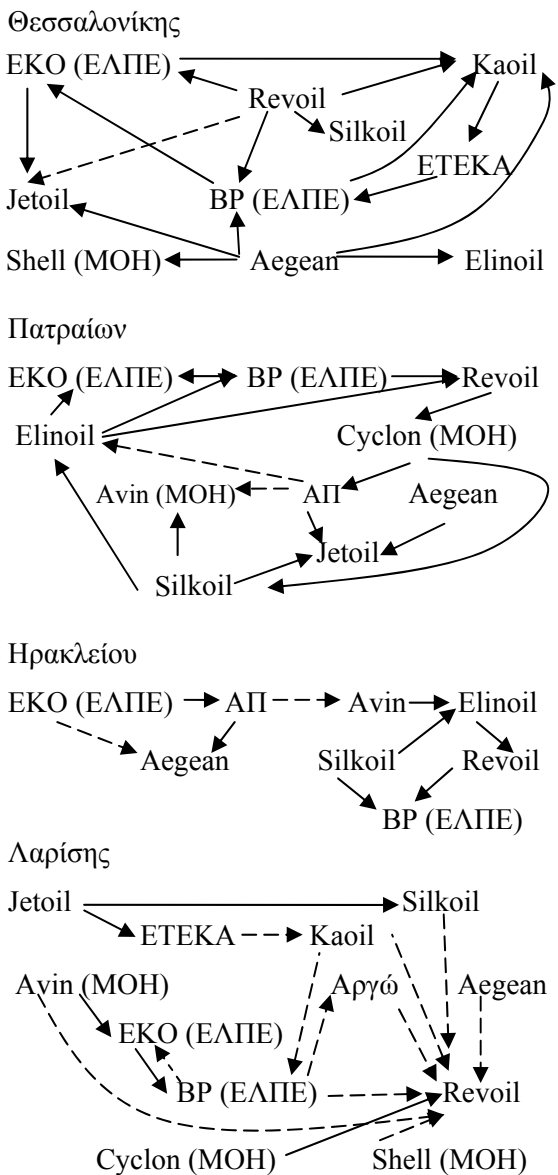
Στον δήμο Θεσσαλονίκης δραστηριοποιούνται δώδεκα εταιρείες εμπορίας, καθώς και ΑΠ. Από τον κοινό χώρο αποτελεσμάτων των δύο ελέγχων αναδύονται οι αιτιότητες που αποτυπώνονται στο Σχήμα 3 με συνεχόμενα βέλη, ενώ μια ακόμη (η εναπομένουσα) στατιστικώς σημαντική αιτιότητα του πολυμεταβλητού υποδείγματος που δεν επιβεβαιώνεται από τους ελέγχους του απλού υποδείγματος δηλώνεται με διακεκομμένο

βέλος. Προκύπτει ότι: (α) δύο κανάλια προηγούνται στην διαμόρφωση τιμών (Aegean, Revoil)· (β) τέσσερα πότε προηγούνται και πότε έπονται (EKO, BP, ETEKA, Kaoil)· (γ) τέσσερα έπονται (Shell, Jetoil, Elinoil, Silkoil)· (δ) τρία κινούνται ανεξάρτητα (Avin, Cyclon, ΑΠ).

Στον δήμο Πατρέων δραστηριοποιούνται δέκα εταιρείες εμπορίας, καθώς και ΑΠ. Από τον κοινό χώρο στατιστικώς σημαντικών απότελεσμάτων των δύο ελέγχων αιτιότητας αναδύονται οι σχέσεις που αποτυπώνονται στο Σχήμα 4 με συνεχόμενα βέλη, ενώ με διακεκομμένα βέλη σημειώνονται οι στατιστικώς σημαντικές αιτιότητες που εμφανίζονται μόνον στο πολυμεταβλητό υπόδειγμα. Προκύπτει ότι: (α) ένα κανάλι προηγείται στην διαμόρφωση τιμών (Aegean)· (β) επτά πότε προηγούνται και πότε έπονται (EKO, BP, Elinoil, Revoil, Silkoil, Cyclon, ΑΠ)· (γ) δύο έπονται (Jetoil, Avin)· (δ) ένα κινείται ανεξάρτητα (Shell).

Στον δήμο Ηρακλείου δραστηριοποιούνται οκτώ εταιρείες εμπορίας, καθώς και ΑΠ. Από τον κοινό χώρο στατιστικώς σημαντικών απότελεσμάτων των δύο ελέγχων αναδύονται οι σχέσεις που αποτυπώνονται στο Σχήμα 5 με συνεχόμενα βέλη, ενώ με διακεκομμένα βέλη σημειώνονται οι στατιστικώς σημαντικές αιτιότητες που εμφανίζονται μόνον στο πολυμεταβλητό υπόδειγμα. Προκύπτει ότι: (α) δύο κανάλια προηγούνται στην διαμόρφωση τιμών (EKO, Silkoil)· (β) τρία πότε προηγούνται και πότε έπονται (Elinoil, Revoil, ΑΠ)· (γ) δύο έπονται (BP, Aegean)· (δ) ένα κινείται ανεξάρτητα (Shell)· (ε) ένα ευρίσκεται σε αμφιλεγό-

Σχήματα 3-6: Σχέσεις αιτιότητας (προήγησης) στις μεταβολές τιμών των καναλιών εμπορίας στους δήμους



Με συνεχή βέλη σημειώνονται οι στατιστικώς σημαντικές κοινές αιτιώδεις σχέσεις των δύο υποδειγμάτων και με διακεκομμένα βέλη οι αιτιώδεις σχέσεις του πολυμεταβλητού υποδείγματος.

μενη κατάσταση όσον αφορά στο αν προηγείται ή αν κάποτε προηγείται και άλλοτε έπεται (Avin).

Στον δήμο Λαρίσης δραστηριοποιούνται 13 εταιρείες εμπορίας, καθώς και ΑΠ. Από τον κοινό χώρο στατιστικώς σημαντικών αποτελεσμάτων των δύο ελέγχων αναδύονται οι σχέσεις που αποτυπώνονται στο Σχήμα 6 με συνεχόμενα βέλη, ενώ με διακεκομμένα βέλη σημειώνονται οι στατιστικώς σημαντικές αιτιότητες που εμφανίζονται μόνον στο πολυμεταβλητό υπόδειγμα. Προκύπτει ότι: (α) τρία κανάλια προηγούνται στην διαμόρφωση τιμών (Jetoil, Avin, Cyclon)· (β) ένα άλλοτε προηγείται και άλλοτε έπεται (ΕΚΟ)· (γ) ένα έπεται (Revoil)· (δ) δύο κινούνται ανεξάρτητα (Elinoil, ΑΠ)· (ε) δύο είτε προηγούνται είτε κινούνται ανεξάρτητα (Shell, Aegean)· (ς) δύο είτε κινούνται ανεξάρτητα είτε κάποιες φορές προηγούνται και άλλες φορές δέχονται επιδράσεις (Kaoil, Αργώ)· (ζ) τρία είτε έπονται είτε περιστασιακά προηγούνται και περιστασιακά δέχονται επιδράσεις (BP, Silkoil, ΕΤΕΚΑ).

Συνάγεται ότι εκ των τριών μεγάλων εταιρειών: (α) η Shell ηγείται σε Αθήνα, Πειραιά (ίσως και Λάρισα), έπεται σε Θεσσαλονίκη και κινείται ανεξάρτητα από τις υπόλοιπες του κλάδου σε Πάτρα και Ηράκλειο· (β) η ΕΚΟ ηγείται στο Ηράκλειο, έπεται στην Αθήνα και κινείται ανεξάρτητα στον Πειραιά· (γ) η BP έπεται στον Πειραιά και το Ηράκλειο. Από την άλλη, η μεσαίου μεγέθους Aegean φαίνεται να έχει την δυνατότητα να διαβάζει τα «βαρομετρικά» στοιχεία των αγορών της Θεσσαλονίκης, του Πειραιώς, των Πατρών, ίσως και της Λαρίσης· ενώ αντίστοιχες δυνατότητες φαίνεται να διαθέτουν οι μεσαίου μεγέθους Revoil στις αγορές Αθηνών και Πειραιώς και Avin στις αγορές Πειραιώς, Λαρίσης, ίσως και Ηρακλείου.

3. ΕΠΙΛΟΓΟΣ

Οι λιανικές τιμές επηρεάζονται στατιστικώς σημαντικά από την τάση, εποχικότητα, τιμή διυλιστηρίου και τους φόρους, τις απεργίες στα μέσα μεταφοράς, τιμολογιακές πολιτικές των εταιρειών και, όπως οι συνθήκες ανταγωνισμού, διαφοροποιούνται στον χώρο. Συνεπώς ενδέχεται να υπάρχουν περιθώρια αύξησης της ευημερίας των καταναλωτών. Επί παραδείγματι, στον βαθμό που το περιθώριο εμπορίας στην Αθήνα ανέρχεται στο (ή υπερβαίνει κατά τι το) 18,7% της μετά φόρων τιμής διυλιστηρίου), ενδέχεται η κοινωνική ευημερία να εξυπηρετείται από την διεύρυνση του ανταγωνισμού στην εμπορία, την μείωση του ειδικού φόρου ή την υποκατάσταση του από άλλους φόρους (κατ' αποκοπήν, επί των κερδών) στα υγρά καύσιμα με ισόποσο αποτέλεσμα, πέραν της (κατά την γνώμη μας προφανή) διευρύνσεως του ανταγωνισμού στην διύλιση.

ABSTRACT

We study unleaded gasoline prices from stations across Greece during April 2011-December 2013. We describe the data and empirically estimate the effects associated with refinery prices, trading companies, geographic areas, the number of competitors in an area, seasonality, strikes in mass-transportation. In addition, we carry out causality tests to explore the state of competition among trading firms in Athens, Thessaloniki and a number of other large municipalities.

Ευχαριστίες: Ευχαριστούμε τους Παναγιώτη Παπαϊωάννου, Αθανάσιο Πετραλιά, έναν ανώνυμο κριτή, καθώς και τους συνέδρους του 26^{ου} Πανελληνίου Συνεδρίου Στατιστικής για εποικοδομητικά σχόλια και υποδείξεις.

ΑΝΑΦΟΡΕΣ

- Bishop S. και M. Walker (2002). *The Economics of EC Competition Law: Concepts, Application and Measurement*. 2^η έκδοση. London: Thomson.
- Davis M.C. (2010). “On Which Days Do Gasoline Stations Raise Prices?” *Atlantic Economic Journal* 38, 113-114.
- Eckert A. και D.S. West (2004). “Retail gasoline price cycles across spatially dispersed gasoline stations.” *Journal of Law and Economics*, 47: 245-273.
- Pennerstorfer D. (2009). “Spatial price competition in retail gasoline markets evidence from Austria.” *Annals of Regional Science*, 43: 133-158.
- Slade M.E. (1992). “Vancouver’s Gasoline-Price Wars: An Empirical Exercise in Uncovering Supergame Strategies.” *Review of Economic Studies*, 59: 257-276.

ΚΑΤΑΝΟΜΕΣ ΧΡΟΝΩΝ ΔΙΑΚΟΠΗΣ ΚΑΙ ΑΘΡΟΙΣΜΑΤΩΝ ΔΙΑΚΟΠΗΣ ΣΕ ΠΛΕΙΟΤΙΜΕΣ ΑΝΤΑΛΛΑΞΙΜΕΣ ΔΟΚΙΜΕΣ

A. X. Ρακιτζής¹, M. B. Μπούτσικας², Δ. Α. Αντζουλάκος²

¹Τμήμα Μαθηματικών και Στατιστικής, Πανεπιστήμιο Κύπρου

²Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης, Πανεπιστήμιο Πειραιώς
arakitz@ucy.ac.cy, mbouts@unipi.gr, dantz@unipi.gr

ΠΕΡΙΛΗΨΗ

Έστω T ένας χρόνος διακοπής (*stopping time*) που βασίζεται σε μια ακολουθία ανεξάρτητων και ισόνομων ή ανταλλάξιμων τ.μ. $X_1, X_2, \dots$, οι οποίες λαμβάνουν τιμές στο σύνολο $\{0, 1, 2, \dots, m\}$. Έστω επίσης $S_{T,i}$ το αντίστοιχο άθροισμα διακοπής (*stopped sum*), το οποίο εκφράζει τον αριθμό των εμφανίσεων του αποτελέσματος « i » στην ακολουθία $X_1, X_2, \dots, X_T$, $0 \leq i \leq m$. Χρησιμοποιώντας μια γενίκευση της μεθοδολογίας των Antzoulakos and Boutsikas (2007) εξάγεται η από κοινού κατανομή των $(T, S_{T,0}, S_{T,1}, \dots, S_{T,m})$ όταν είναι γνωστή η κατανομή του T , αρχικά για την περίπτωση των ανεξάρτητων και ισόνομων τ.μ. και στη συνέχεια στην περίπτωση των ανταλλάξιμων τ.μ. Τέλος, παρουσιάζονται και δύο πρακτικές εφαρμογές των θεωρητικών αποτελεσμάτων.

Λέξεις Κλειδιά: Άθροισμα Διακοπής, Ανταλλάξιμες Δοκιμές, Θεώρημα Αναπαράστασης de Finetti, Πιθανογεννήτρια Συνάρτηση, Πλειότιμες Δοκιμές, Χρόνος Διακοπής.

1. ΕΙΣΑΓΩΓΗ

Η μελέτη των ιδιοτήτων των κατανομών ενός χρόνου διακοπής (*stopping time*) και ενός αθροίσματος διακοπής (*stopped sum*) της μορφής $S_T = X_1 + X_2 + \dots + X_T$ έχουν προσελκύσει το ενδιαφέρον των ερευνητών, κυρίως λόγω των πολλών πρακτικών τους εφαρμογών. Μια ειδική κατηγορία τέτοιων προβλημάτων είναι αυτά τα οποία σχετίζονται με τη μελέτη κατανομών ροών και σχηματισμών (*pattern-waiting time distributions*) με εφαρμογές στα πεδία του ποιοτικού ελέγχου, της αξιοπιστίας συστημάτων και της μοριακής βιολογίας (δείτε, π.χ. Balakrishnan and Koutras (2002)). Πρόσφατα (δείτε Eryilmaz (2008, 2010), Inoue et al. (2011)) η μελέτη των κατανομών ροών και σχηματισμών επεκτάθηκε και στην περίπτωση των πλειότιμων ανταλλάξιμων δοκιμών (*multistate exchangeable trials*).

Σημαντικό ρόλο στη μελέτη της κατανομής του χρόνου διακοπής T σε μια ακολουθία δοκιμών $X_1, X_2, \dots$, παίζει ο συνολικός αριθμός $S_{T,i}$ των αποτελεσμάτων

τύπου «i» τα οποία έχουν εμφανιστεί στην ακολουθία $X_1, X_2, \dots, X_T$, αφού μας παρέχει σημαντικές πληροφορίες για το υπό μελέτη πείραμα. Η μελέτη των κατανομών των T και $S_{T,i}$ ξεκίνησε από τους Aki and Hirano (1994) για ανεξάρτητες και ισόνομες δίτιμες τυχαίες μεταβλητές (τ.μ.) και στη συνέχεια επεκτάθηκε από τους Antzoulakos and Philippou (1996), Balakrishnan (1997), Chadjiconstantinidis et al. (2000) και Inoue (2004).

Πρόσφατα οι Antzoulakos and Boutsikas (2007) πρότειναν μια μέθοδο για την εύρεση της από κοινού πιθανογεννήτριας συνάρτησης των (T, S_T, F_T) μέσω της πιθανογεννήτριας του T , όπου T είναι ο χρόνος αναμονής μέχρι την r -οστή εμφάνιση ενός σύνθετου σχηματισμού E σε μια ακολουθία ανεξάρτητων και ισόνομων δίτιμων δοκιμών. Στην παρούσα εργασία γενικεύεται η μεθοδολογία των Antzoulakos and Boutsikas (2007) για την περίπτωση των ανεξάρτητων και ισόνομων πλειότιμων δοκιμών ενώ παράλληλα επεκτείνεται και στην περίπτωση των ανταλλάξιμων πλειότιμων δοκιμών. Στην Ενότητα 2 εξάγεται αρχικά η από κοινού κατανομή των $(T, \mathbf{S}_T) = (T, S_{T,0}, S_{T,1}, \dots, S_{T,m})$ όταν είναι γνωστή η κατανομή του χρόνου διακοπής T ο οποίος σχετίζεται με μια ακολουθία ανεξάρτητων και ισόνομων τ.μ. ή με μια ακολουθία ανταλλάξιμων τ.μ. με τιμές στο σύνολο $\{0, 1, 2, \dots, m\}$. Στην Ενότητα 3 παρουσιάζονται δύο πρακτικές εφαρμογές των θεωρητικών αποτελεσμάτων.

2. ΘΕΩΡΗΤΙΚΑ ΑΠΟΤΕΛΕΣΜΑΤΑ

2.1. ΑΝΕΞΑΡΤΗΤΕΣ ΚΑΙ ΙΣΟΝΟΜΕΣ ΔΟΚΙΜΕΣ

Έστω $X_1, X_2, \dots$ ακολουθία τ.μ. με τιμές στο σύνολο $\{0, 1, 2, \dots, m\}$. Έστω επίσης T χρόνος διακοπής σχετικός με την ακολουθία $\{X_i\}_{i=1}^{\infty}$, δηλαδή τέτοιος ώστε για κάθε $n \geq 1$ το ενδεχόμενο $[T = n]$ να ανήκει στη σ -άλγεβρα $\sigma(\mathbf{X}_n) = \sigma(X_1, X_2, \dots, X_n)$ η οποία παράγεται από τα $X_1, X_2, \dots, X_n$. Επιπλέον, έστω το άθροισμα διακοπής

$$S_{n,i} = \sum_{j=1}^n I_{[X_j=i]}, \quad i = 0, 1, \dots, m, \quad n \geq 1,$$

το οποίο εκφράζει το συνολικό αριθμό εμφανίσεων του αποτελέσματος «i» στην ακολουθία $X_1, X_2, \dots, X_n$ (I_A είναι δείκτρια συνάρτηση του ενδεχομένου A). Υποθέτουμε ότι όλα τα παραπάνω ορίζονται σε ένα χώρο πιθανότητας $(\Omega, \mathcal{F}, P_{\mathbf{z},\mathbf{p}})$, όπου υπό το μέτρο πιθανότητας $P_{\mathbf{z},\mathbf{p}}$ οι τ.μ. $X_1, X_2, \dots$ είναι ανεξάρτητες και ισόνομες με

$$P_{\mathbf{z},\mathbf{p}}(X_n = i) = p_i z_i / \left(\sum_{j=0}^m p_j z_j \right), \quad i = 0, 1, \dots, m, \quad n \geq 1,$$

με $\mathbf{z} = (z_0, z_1, \dots, z_m)$, $z_i > 0$, $\mathbf{p} = (p_1, \dots, p_m)$ και $p_0 = 1 - \sum_{i=1}^m p_i$, $p_i \in (0, 1)$. Για $\mathbf{z} = \mathbf{1} = (1, 1, \dots, 1)$, οι τ.μ. $X_1, X_2, \dots$ είναι ανεξάρτητες και ισόνομες με

$P_{1,p}(X_n = i) = p_i$, $i = 0, 1, \dots, m$, $n \geq 1$. Θα συμβολίζουμε ως $E_{z,p}(\cdot)$ την αναμενόμενη τιμή όταν αυτή υπολογίζεται από το μέτρο πιθανότητας $P_{z,p}$.

Στη συνέχεια δίδονται (χωρίς απόδειξη) τα βασικά θεωρητικά αποτελέσματα της μελέτης μας, υποθέτοντας ότι ισχύει $P_{z,p}(T < \infty) = 1$, για κάθε $z_0, z_1, \dots, z_m$, σε μια περιοχή του μηδενός. Για τις αποδείξεις των αποτελεσμάτων, δείτε Boutsikas et al. (2013). Όλα τα αποτελέσματα ισχύουν, μετά από κατάλληλες τροποποιήσεις, και για χρόνους διακοπής με μάζα στο άπειρο (δηλ. για T τέτοιους ώστε $P_{z,p}(T = \infty) > 0$). Η επόμενη πρόταση μας δίνει την από κοινού πιθανογεννήτρια συνάρτηση των (T, S_T) υπό το μέτρο πιθανότητας $P_{1,p}$, όταν είναι γνωστή η πιθανογεννήτρια του T υπό το μέτρο πιθανότητας $P_{z,p}$.

Πρόταση 1. *Η από κοινού πιθανογεννήτρια συνάρτηση των (T, S_T) στην περίπτωση των ανεξάρτητων και ισόνομων τ.μ. ικανοποιεί τη σχέση*

$$E_{1,p}\left(u^T z_0^{S_{T,0}} z_1^{S_{T,1}} \dots z_m^{S_{T,m}}\right) = E_{z,p}\left[\left(u \sum_{i=0}^m p_i z_i\right)^T\right], \quad (1)$$

για όλα τα $u, z_0, z_1, \dots, z_m > 0$ για τα οποία υπάρχουν οι παραπάνω μέσες τιμές.

Χρησιμοποιώντας την Πρόταση 1 μπορούμε να καταλήξουμε σε χρήσιμες σχέσεις μεταξύ των κατανομών του T και συναρτήσεων του S_T . Έτσι, αν θεωρήσουμε το άθροισμα διακοπής $S_T = \sum_{i=0}^m i S_{T,i} = \sum_{j=1}^T X_j$ και θέσουμε $z_i = z^i$ στην (1), έπεται ότι

$$E_{1,p}(u^T z^{S_T}) = E_{1,p}(u^T z^{0S_{T,0}} z^{1S_{T,1}} \dots z^{mS_{T,m}}) = E_{\tilde{z},p}\left[\left(u \sum_{i=0}^m p_i z^i\right)^T\right] = E_{\tilde{z},p}\left[\left(u E_{1,p}(z^{X_1})\right)^T\right], \quad (2)$$

όπου $\tilde{z} = (z^0, z^1, \dots, z^m)$. Η σχέση (2) έχει αποδειχθεί από τους Antzoulakos and Boutsikas (2007) στην περίπτωση των ανεξάρτητων και ισόνομων δίτιμων τ.μ. (για συγκεκριμένο χρόνο διακοπής T), ενώ για μια πιο γενική απόδειξη, για οποιεσδήποτε τ.μ. X_i , δείτε Boutsikas et al. (2011).

Επιπλέον, θέτοντας στην (1) $u = \left(\sum_{i=0}^m p_i z_i\right)^{-1}$ προκύπτει

$$E_{1,p}\left[\left(\sum_{i=0}^m p_i z_i\right)^{-T} z_0^{S_{T,0}} z_1^{S_{T,1}} \dots z_m^{S_{T,m}}\right] = 1, \quad (3)$$

η οποία μπορεί να θεωρηθεί ως ειδική περίπτωση της ταυτότητας του Wald (*Wald's Fundamental Identity*, Karlin and Taylor (1975, pp. 264)). Παραγωγίζοντας την (3) ως προς z_i και υπολογίζοντας τις παραγώγους στο $\mathbf{z} = \mathbf{1}$ λαμβάνουμε το παρακάτω σύνολο εξισώσεων

$$E_{1,p}(S_{T,i}) = p_i E_{1,p}(T), \quad i = 0, 1, \dots, m.$$

Από τις παραγώγους 2^{ας} τάξης της (3) ως προς z_i, z_j (για $\mathbf{z} = \mathbf{1}$) έπεται ότι

$\text{Var}_{\mathbf{1},\mathbf{p}}(S_{T,i} - p_i T) = p_i(1 - p_i)E_{\mathbf{1},\mathbf{p}}(T)$, $\text{Cov}_{\mathbf{1},\mathbf{p}}(S_{T,i} - p_i T, S_{T,j} - p_j T) = -p_i p_j E_{\mathbf{1},\mathbf{p}}(T)$,
για $i, j = 0, 1, \dots, m$, $i \neq j$. Οι εξισώσεις για τα $E_{\mathbf{1},\mathbf{p}}(S_{T,i})$ και $\text{Var}_{\mathbf{1},\mathbf{p}}(S_{T,i} - p_i T)$
αποτελούν παραλλαγές της 1^{ης} και της 2^{ης} εξίσωσης του Wald.

2.2. ΑΝΤΑΛΛΑΞΙΜΕΣ ΔΟΚΙΜΕΣ

Έστω χώρος πιθανότητας (Ω, F, P_e) και $X_1, X_2, \dots$ μια άπειρη ακολουθία
ανταλλάξιμων πλειότιμων τ.μ. με τιμές στο σύνολο $\{0, 1, 2, \dots, m\}$. Για την από κοινού
κατανομή των $X_1, X_2, \dots, X_n$, $n \geq 1$, ισχύει ότι

$$P_e(X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) = P_e(X_1 = x_{\pi(1)}, X_2 = x_{\pi(2)}, \dots, X_n = x_{\pi(n)}),$$

για οποιαδήποτε μετάθεση $(\pi(1), \pi(2), \dots, \pi(n))$ των δεικτών $\{1, 2, \dots, n\}$. Επιπλέον,
υπό το μέτρο πιθανότητας P_e (αντ. $P_{z,\mathbf{p}}$) οι τ.μ. $X_1, X_2, \dots$ είναι ανταλλάξιμες (αντ.
ανεξάρτητες και ισόνομες).

Θεωρούμε το σύνολο (*simplex*)

$$D = \{\mathbf{p} = (p_1, p_2, \dots, p_m) \in [0, 1]^m : p_1 + p_2 + \dots + p_m \leq 1\}.$$

Από το Θεώρημα Αναπαράστασης του de Finetti (*De Finetti's Representation Theorem*, Zabell (1982), Mauldin et al. (1992)) υπάρχει ένα μέτρο πιθανότητας στο D
με συνάρτηση κατανομής G , τέτοιο ώστε

$$P_e(\mathbf{X}_n = \mathbf{x}) = \int_D \prod_{i=0}^m p_i^{r_{n,i}(\mathbf{x})} dG(\mathbf{p}),$$

για $\mathbf{x} = (x_1, x_2, \dots, x_n)$ και $r_{n,i}(\mathbf{x}) = \sum_{j=1}^n I_{[x_j=i]}$, $i = 0, 1, \dots, m$. Το μέτρο πιθανότητας G
είναι γνωστό και ως μέτρο του de Finetti ή μέτρο μίξης (*mixing measure*) της
ακολουθίας $X_1, X_2, \dots$. Η επόμενη πρόταση συνδέει την από κοινού πιθανογεννήτρια
συνάρτηση του τυχαίου διανύσματος $(T, \mathbf{S}_T)$ με την πιθανογεννήτρια του χρόνου
διακοπής T , στην περίπτωση των ανταλλάξιμων πλειότιμων δοκιμών.

Πρόταση 2: Έστω χρόνος διακοπής T , τέτοιος ώστε $P_e(T < \infty) = P_{z,\mathbf{p}}(T < \infty) = 1$,
 $\mathbf{p} \in D$. Τότε η από κοινού πιθανογεννήτρια συνάρτηση του τυχαίου διανύσματος
 $(T, \mathbf{S}_T)$ στην περίπτωση των ανταλλάξιμων πλειότιμων δοκιμών ικανοποιεί τη σχέση

$$E_e(u^T z_0^{S_{T,0}} z_1^{S_{T,1}} \dots z_m^{S_{T,m}}) = \int_D E_{z,\mathbf{p}}[(u \sum_{i=0}^m p_i z_i)^T] dG(\mathbf{p}) \quad (4)$$

για όλα τα $u, z_0, z_1, \dots, z_m > 0$ για τα οποία υπάρχουν οι παραπάνω μέσες τιμές.

Χρησιμοποιώντας τη σχέση (4), αποδεικνύεται ότι (βλ. Boutsikas et al. (2013)) αν
 Y είναι τ.μ. τέτοια ώστε η $Y \cdot I_{[T=n]}$ είναι μια $\sigma(\mathbf{X}_n)$ -μετρήσιμη τ.μ., τότε

$$E_e(Y) = \int_D E_{\mathbf{1},\mathbf{p}}(Y) dG(\mathbf{p}).$$

Θέτοντας $Y = S_{T,i}$ και χρησιμοποιώντας τη σχέση $E_{\mathbf{1},\mathbf{p}}(S_{T,i}) = p_i E_{\mathbf{1},\mathbf{p}}(T)$, έπεται ότι

$$E_e(S_{T,i}) = \int_D E_{1,p}(S_{T,i}) dG(\mathbf{p}) = \int_D p_i E_{1,p}(T) dG(\mathbf{p}),$$

ενώ με τρόπο ανάλογο έχουμε τις παρακάτω σχέσεις ($i, j = 0, 1, \dots, m, i \neq j$)

$$\text{Var}_e(S_{T,i} - p_i T) = \int_D p_i(1 - p_i) E_{1,p}(T) dG(\mathbf{p}),$$

$$\text{Cov}_e(S_{T,i} - p_i T, S_{T,j} - p_j T) = - \int_D p_i p_j E_{1,p}(T) dG(\mathbf{p}), \quad E_e(T) = \int_D E_{1,p}(T) dG(\mathbf{p}).$$

Τέλος, για $Y = u^T z^{S_T}$, $S_T = \sum_{j=1}^T X_j$, από τις σχέσεις (2) και (4) λαμβάνουμε

$$E_e(u^T z^{S_T}) = \int_D E_{z,p} \left[\left(u E_{1,p}(z^{X_1}) \right)^T \right] dG(\mathbf{p})$$

όπου $\tilde{\mathbf{z}} = (z^0, z^1, \dots, z^m)$, και

$$E_e(S_T) = \int_D E_{1,p}(X_1) E_{1,p}(T) dG(\mathbf{p}).$$

Στην επόμενη Ενότητα παρουσιάζουμε πρακτικές εφαρμογές των θεωρητικών αποτελεσμάτων της Ενότητας 2.

3. ΕΦΑΡΜΟΓΕΣ

3.1. ΚΑΤΑΝΟΜΗ ΠΛΗΘΟΥΣ ΑΠΟΔΕΚΤΩΝ ΣΩΡΩΝ ΣΕ ΕΝΑ ΣΥΣΤΗΜΑ ΔΕΙΓΜΑΤΟΛΗΨΙΑΣ ΑΠΟΔΟΧΗΣ

Στα συστήματα δειγματοληψίας αποδοχής κατά μεταβλητές, κατά την επιθεώρηση ενός σωρού αντικειμένων λαμβάνεται τυχαίο δείγμα n προϊόντων από κάθε σωρό. Η αποδοχή ή η απόρριψη του σωρού βασίζεται στην πληροφορία η οποία περιέχεται στο δείγμα των n αντικειμένων. Συνήθως, η διαδικασία επιθεώρησης ξεκινάει υπό «κανονικό» επίπεδο και συνεχίζει υπό αυτό για όσο χρονικό διάστημα η παραγωγική διεργασία παράγει αντικείμενα κοντά σε ένα αποδεκτό επίπεδο ποιότητας. Όταν όμως υπάρχουν ενδείξεις ότι η ποιότητα των παραγόμενων προϊόντων έχει χειροτερεύσει (αντ. βελτιωθεί), το επίπεδο επιθεώρησης αλλάζει από «κανονικό» σε «αυστηρό» (αντ. «ελαστικό»). Η εναλλαγή μεταξύ των διαφορετικών επιπέδων γίνεται με χρήση κατάλληλων κανόνων διακοπής (*switching rules*). Μας ενδιαφέρει η εύρεση της κατανομής του αριθμού των αποδεκτών σωρών μέχρι να επέλθει η αλλαγή στο επίπεδο επιθεώρησης (κατανομή αθροίσματος διακοπής).

Έστω ότι ν ανεξάρτητα δείγματα αναφοράς είναι διαθέσιμα από μια διεργασία η οποία παράγει προϊόντα κοντά σε ένα αποδεκτό επίπεδο ελέγχου και έστω Y_t , $1 \leq t \leq \nu$, στατιστική συνάρτηση (σ.σ.) του t -οστού δείγματος (π.χ. δειγματικός μέσος, δειγματική διάμεσος). Έστω $U = g(Y_1, Y_2, \dots, Y_\nu)$ τυχαίο κατώφλι, το οποίο λειτουργεί ως άνω όριο προδιαγραφών για τη διεργασία. Η διαδικασία επιθεώρησης είναι η εξής: Από κάθε (μελλοντικό) σωρό προϊόντων, λαμβάνεται τυχαίο δείγμα μεγέθους n αντικειμένων και έστω Z_i η σ.σ. του i -οστού δείγματος. Τότε, ο σωρός γίνεται αποδεκτός αν και μόνο αν $Z_i \leq U$, $i = 1, 2, \dots$. Θεωρώντας τις τ.μ.

$$X_i = I_{[Z_i \leq U]} = \begin{cases} 1, & Z_i \leq U \\ 0, & Z_i > U \end{cases},$$

μπορούμε να διαπιστώσουμε ότι αυτές αποτελούν μια άπειρη ακολουθία δίτιμων ανταλλάξιμων τ.μ. όπου το αντίστοιχο μέτρο πιθανότητας G έχει αθροιστική συνάρτηση κατανομής (α.σ.κ.) $G(p) = F_U(F_Z^{-1}(p))$, $p \in (0,1)$ (όπου $F_V(\cdot)$, $f_V(\cdot)$ είναι η α.σ.κ. και η συνάρτηση πυκνότητας πιθανότητας (σ.π.π.) της τ.μ. V). Πράγματι, για $(x_1, x_2, \dots, x_n) \in \{0,1\}^n$ και $c = \sum_{j=1}^n x_j$ έπεται ότι

$$P_e(\mathbf{X}_n = \mathbf{x}_n) = \int_R P_e(\mathbf{X}_n = \mathbf{x}_n | U = u) dF_U(u) = \int_R (F_Z(u))^c (1 - F_Z(u))^{n-c} dF_U(u) \Leftrightarrow$$

$$P_e(\mathbf{X}_n = \mathbf{x}_n) = \int_0^1 p^c (1-p)^{n-c} dF_U(F_Z^{-1}(p)) = \int_0^1 p^c (1-p)^{n-c} dG(p).$$

Θεωρώντας ως τυχαίο όριο την s -τάξης ($1 \leq s \leq \nu$) διατεταγμένη παρατήρηση του δείγματος αναφοράς (δηλ., $U = Y_{s:\nu}$) έπεται ότι η σ.π.π. $f_U(y)$ είναι

$$f_U(y) = f_{Y_{s:\nu}}(y) = \frac{\nu!}{(s-1)!(\nu-s)!} (1 - F_Y(y))^{\nu-s} (F_Y(y))^{s-1} f_Y(y), \quad y \in R.$$

Αν οι τ.μ. Z, Y είναι ανεξάρτητες και ισόνομες (δηλ. $F_Z = F_Y$) έπεται ότι

$$dG(p) = d(F_U(F_Y^{-1}(p))) = f_U(F_Y^{-1}(p)) dF_Y^{-1}(p) = \frac{\nu!}{(s-1)!(\nu-s)!} p^{s-1} (1-p)^{\nu-s} dp$$

και άρα το μέτρο του de Finetti σε αυτή την περίπτωση θα είναι η κατανομή Βήτα με παραμέτρους s και $\nu - s + 1$.

Στη συνέχεια θεωρούμε τον παρακάτω κανόνα εναλλαγής μεταξύ των διαφορετικών επιπέδων επιθεώρησης: Αν k (αντ. r) συνεχόμενες τιμές της σ.σ. Z είναι μικρότερες (αντ. μεγαλύτερες) από το τυχαίο όριο U , το επίπεδο της επιθεώρησης αλλάζει από «κανονικό» σε «ελαστικό» (αντ. «αυστηρό»). Ο χρόνος διακοπής T μέχρι την αλλαγή στο επίπεδο της επιθεώρησης είναι $T = \min\{T_0, T_1\}$, όπου

$$T_0 = \min\{j : X_{j-r+1} = \dots = X_j = 0\}, \quad T_1 = \min\{j : X_{j-k+1} = \dots = X_j = 1\}.$$

Τα αθροίσματα διακοπής $S_{T,0} = \sum_{j=1}^T I_{[X_j=0]}$ και $S_{T,1} = \sum_{j=1}^T I_{[X_j=1]}$ εκφράζουν, αντίστοιχα, το συνολικό αριθμό απορριπτέων και αποδεκτών σωρών μέχρι την αλλαγή στο επίπεδο της επιθεώρησης. Η γνώση των κατανομών των $S_{T,0}$ και $S_{T,1}$ μπορεί να δώσει περισσότερες πληροφορίες σχετικά με το επίπεδο ποιότητας των παραγόμενων σωρών. Χρησιμοποιώντας την Πρόταση 2, έπεται ότι η από κοινού πιθανογεννήτρια συνάρτηση των $(T, S_{T,0}, S_{T,1})$ ισούται με

$$E_e(u^T z_0^{S_{T,0}} z_1^{S_{T,1}}) = \int_0^1 E_{z,p} \left[(u((1-p)z_0 + pz_1))^T \right] dG(p),$$

όπου $\mathbf{z} = (z_0, z_1)$ και $\mathbf{p} = p$. Θέτοντας $z_0 = 1$ στην παραπάνω σχέση, λαμβάνουμε την από κοινού πιθανογεννήτρια συνάρτηση των $(T, S_{T,1})$,

$$E_e(u^T z_1^{S_{T,1}}) = \int_0^1 E_{\mathbf{z}, \mathbf{p}} \left[(u(1-p + pz_1))^T \right] dG(p).$$

Υπό το μέτρο πιθανότητας $P_{\mathbf{z}, \mathbf{p}}$ οι X_i είναι ανεξάρτητες και ισόνομες δίτιμες τ.μ. με

$$P_{\mathbf{z}, \mathbf{p}}(X_n = 0) = \frac{1-p}{1-p+pz}, \quad P_{\mathbf{z}, \mathbf{p}}(X_n = 1) = \frac{pz}{1-p+pz}, \quad n \geq 1.$$

Η πιθανογεννήτρια συνάρτηση του χρόνου αναμονής T έχει προσδιοριστεί από τον Feller (1968, p. 327) και δίνεται από τη σχέση

$$E_{1,p}(w^T) = \frac{(1-\rho w)(\rho w)^k (1 - ((1-\rho)w)^r) + (1 - (1-\rho)w)((1-\rho)w)^r (1 - (\rho w)^k)}{1 - w + (1-\rho)w(\rho w)^k + (\rho w)((1-\rho)w)^r - (\rho w)^k ((1-\rho)w)^r}, \quad (5)$$

όπου ρ είναι η πιθανότητα επιτυχίας. Αντικαθιστώντας στη σχέση (5) το ρ με $pz/(1-p+pz)$ και το w με $u(1-p+pz)$, προκύπτει (για $q = 1-p$)

$$h_{p,u}(z) = E_{\mathbf{z}, \mathbf{p}} \left[(u(1-p+pz))^T \right] = \frac{(1-pzu)(pzu)^k (1 - (qu)^r) + (1-qu)(qu)^r (1 - (pzu)^k)}{1 - u(q+pz) + qu(pzu)^k + pzu(qu)^r - (pzu)^k (qu)^r}.$$

Άρα, η από κοινού πιθανογεννήτρια συνάρτηση των $(T, S_{T,1})$ είναι

$$E_e(u^T z^{S_{T,1}}) = \frac{\nu!}{(s-1)!(\nu-s)!} \int_0^1 h_{p,u}(z) p^{s-1} (1-p)^{\nu-s} dp$$

ενώ η πιθανογεννήτρια του $S_{T,1}$ προκύπτει άμεσα από τον παραπάνω τύπο για $u = 1$.

Η συνάρτηση πιθανότητας του $S_{T,1}$ υπολογίζεται αριθμητικά συνήθως μέσω κατάλληλου λογισμικού (για διάφορες τιμές των k, r, ν και s) χρησιμοποιώντας τη σχέση

$$P_e(S_{T,1} = i) = \frac{1}{i!} \frac{d^i}{dz^i} E_e(z^{S_{T,1}}) \Big|_{z=0} = \frac{\nu}{(s-1)!(\nu-s)! i!} \int_0^1 h_{p,1}^{(i)}(0) p^{s-1} (1-p)^{\nu-s} dp, \quad i = 0, 1, \dots$$

όπου $h_{p,1}^{(i)}(0)$ είναι η i -οστής τάξης παράγωγος της $h_{p,1}$ στο $z = 0$.

3.2. ΤΟ ΠΡΟΒΛΗΜΑ ΤΟΥ ΣΥΛΛΕΚΤΗ ΚΟΥΠΟΝΙΩΝ

Θεωρούμε μια κάλπη η οποία περιέχει $c_0 + c_1 + \dots + c_m = c$ κουπόνια, c_i από τα οποία φέρουν τον αριθμό « i », $i = 0, 1, \dots, m$. Ένα κουπόνι επιλέγεται με τυχαίο τρόπο από την κάλπη, ο αριθμός του καταγράφεται και στη συνέχεια τοποθετείται πάλι στην κάλπη μαζί με s ($s = 0, 1, \dots$) κουπόνια τα οποία φέρουν τον ίδιο αριθμό. Η διαδικασία αυτή επαναλαμβάνεται μέχρι να εμφανιστεί κάθε ένας από τους $m+1$ αριθμούς τουλάχιστον 1 φορά. Το πρόβλημα αυτό είναι γνωστό ως *πρόβλημα του συλλέκτη κουπονιών* (*coupon collector's problem*).

Έστω X_n ($n \geq 1$) η τ.μ. η οποία εκφράζει τον αριθμό του κουπονιού το οποίο εξάγεται κατά τη n -οστή δοκιμή. Όταν $s = 0$, η ακολουθία των $X_1, X_2, \dots$ αποτελεί ακολουθία ανεξάρτητων και ισόνομων τ.μ. με αντίστοιχες πιθανότητες $p_i = P_{1,p}(X_n = i) = c_i / c$, $i = 0, 1, \dots, m$. Η πιθανογεννήτρια συνάρτηση του χρόνου διακοπής T (αριθμός που εξάγονται μέχρι τον τερματισμό του πειράματος) είναι γνωστή (Inoue and Aki (2007)) και ισούται με

$$E_{1,p}(w^T) = 1 + (-1)^{m+1}(1-w) + \sum_{j=0}^{m-1} \left(\sum_{0 \leq i_0 < i_1 < \dots < i_j \leq m} \frac{(-1)^{m-j}(1-w)}{1 - (p_{i_0} + \dots + p_{i_j})w} \right).$$

Όταν $s > 0$, η ακολουθία των $X_1, X_2, \dots$ αποτελεί ακολουθία ανταλλάξιμων τ.μ., όπου το μέτρο G του de Finetti είναι η κατανομή Dirichlet (Eryilmaz (2008)),

$$dG(\mathbf{p}) = \frac{\Gamma(\sum_{i=0}^m a_i)}{\prod_{i=0}^m \Gamma(a_i)} \prod_{i=0}^m p_i^{a_i-1} d\mathbf{p},$$

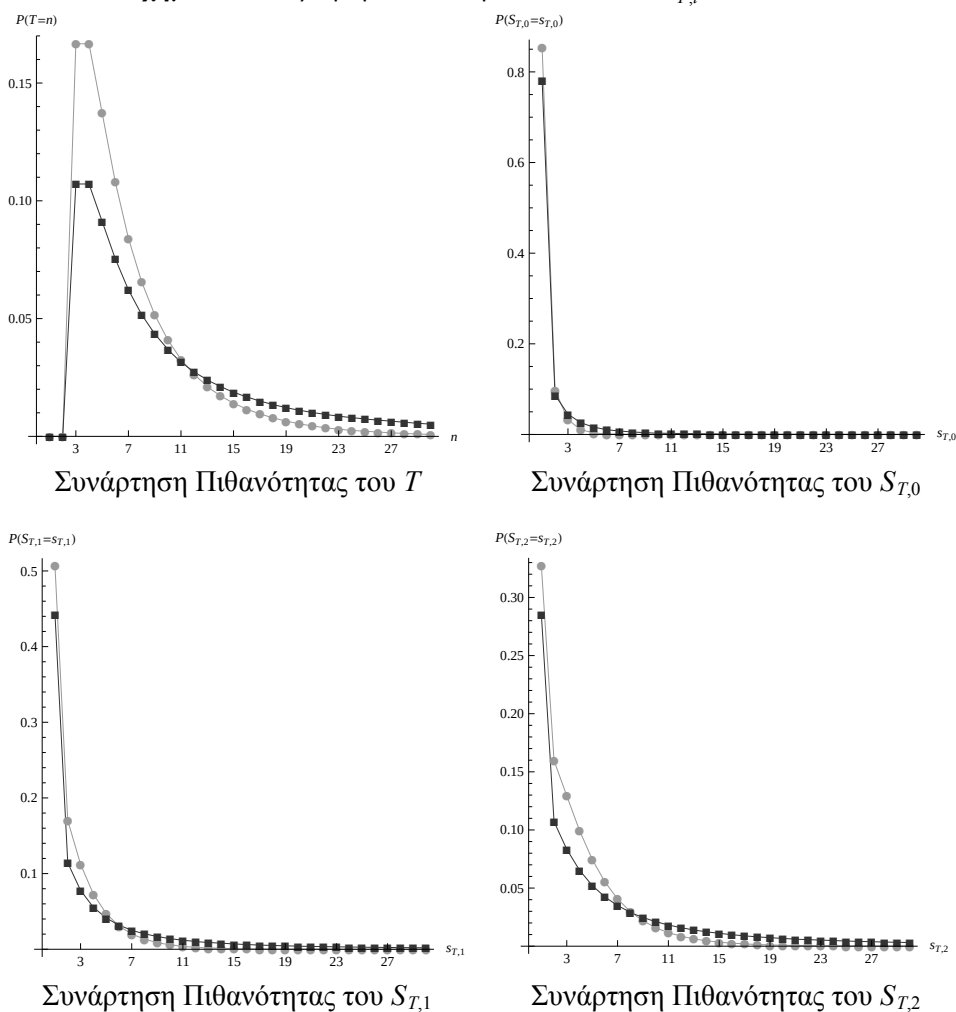
όπου $a_i = c_i / s$, $i = 0, 1, \dots, m$. Επίσης, η πιθανότητα $P_e(X_n = i)$ είναι η ίδια είτε $s > 0$ είτε $s = 0$ και ισούται με $P_e(X_n = i) = c_i / c = a_i / (a_0 + a_1 + \dots + a_m)$, $i = 0, 1, \dots, m$.

Έστω το άθροισμα διακοπής $S_{T,i} = \sum_{j=1}^T I_{[X_j=i]}$, το οποίο εκφράζει το συνολικό αριθμό κουπονιών που φέρουν τον αριθμό « i » και τα οποία έχουν εξαχθεί μέχρι τον τερματισμό του πειράματος. Χρησιμοποιώντας τις Προτάσεις 1 και 2 μπορούμε να προσδιορίσουμε την κατανομή του $S_{T,i}$ στην περίπτωση των ανεξάρτητων και ισόνομων δοκιμών όπως και στην περίπτωση των ανταλλάξιμων δοκιμών. Ειδικότερα, όταν $s > 0$, η από κοινού πιθανογεννήτρια συνάρτηση των $(T, S_{T,0}, S_{T,1}, \dots, S_{T,m})$ δίνεται από τη σχέση

$$E_e(u^T z_0^{S_{T,0}} z_1^{S_{T,1}} \dots z_m^{S_{T,m}}) = \frac{\Gamma(\sum_{i=0}^m a_i)}{\prod_{i=0}^m \Gamma(a_i)} \int_0^1 \dots \int_0^1 E_{z,p} \left[\left(u \left(\sum_{i=0}^m p_i z_i \right) \right)^T \right] \prod_{i=0}^m p_i^{a_i-1} dp_1 \dots dp_m,$$

όπου $p_0 = 1 - \sum_{i=1}^m p_i$ ενώ η $E_{z,p} \left[\left(u \left(\sum_{i=0}^m p_i z_i \right) \right)^T \right]$ προκύπτει άμεσα από τη $E_{1,p}(w^T)$ αντικαθιστώντας το p_i με $p_i z_i / \sum_{i=0}^m p_i z_i$, $i = 0, 1, \dots, m$, και το w με $u \left(\sum_{i=0}^m p_i z_i \right)$. Αξίζει να αναφέρουμε ότι η προς ολοκλήρωση ποσότητα είναι η από κοινού κατανομή των (T, S_T) υπό το μοντέλο των ανεξάρτητων και ισόνομων τ.μ. (δείτε Πρόταση 1). Στο Σχήμα 1 δίνεται η κατανομή των T και $S_{T,i}$ υπό το μοντέλο των ανεξάρτητων και ισόνομων τ.μ. (■) και των ανταλλάξιμων τ.μ. (●). Για το μοντέλο των ανταλλάξιμων τ.μ. επιλέξαμε $m = 2$, $c_0 = 1$, $c_1 = 2$, $c_2 = 3$, $s = 1$ ενώ για το μοντέλο των ανεξάρτητων και ισόνομων τ.μ. ($s = 0$) επιλέξαμε $p_0 = 1/6$, $p_1 = 1/3$, $p_2 = 1/2$.

Σχήμα 1. Συνάρτηση Πιθανότητας των T και $S_{T,i}$, $i = 0, 1, 2$.



ABSTRACT

Let T be a stopping time associated with a sequence of independent and identically distributed (iid) or exchangeable random variables (r.v.) $X_1, X_2, \dots$, taking values in $\{0, 1, 2, \dots, m\}$. Let also $S_{T,i}$ be the respective stopped sum which denotes the number of appearances of outcome "i" in the sequence $X_1, X_2, \dots, X_T$, $0 \leq i \leq m$. Using a generalization of the methodology of Antzoulakos and Boutsikas (2007) we derive the joint distribution of $(T, S_{T,0}, S_{T,1}, \dots, S_{T,m})$ via the distribution of T , in the iid case as well as in the exchangeable case. Two practical applications are also offered in order to illustrate the applicability and the usefulness of the theoretical results.

ΑΝΑΦΟΡΕΣ

Aki, S. and Hirano, K. (1994). Distributions of numbers of failures and successes until the first consecutive k successes. *Annals of the Institute of Statistical Mathematics*, **46**, 193-202.

- 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. and Philippou, A. N. (1996). Derivation of the probability distribution functions for succession quota random variables. *Annals of the Institute of Statistical Mathematics*, **48**, 551-561.
- Balakrishnan, N. (1997). Joint distributions of numbers of success-runs and failures until the first consecutive k successes in a binary sequence. *Annals of the Institute of Statistical Mathematics*, **49**, 519-529.
- Balakrishnan, N. and Koutras, M. V. (2002). *Runs, Scans and Applications*, John Wiley and Sons.
- Boutsikas, M., Antzoulakos, D. L. and Rakitzis, A.C. (2013). On the joint distribution of stopping times and stopped sums in multistate exchangeable trials. *Journal of Applied Probability*. To appear.
- Boutsikas, M., Rakitzis, A.C. and Antzoulakos, D. L. (2011). On the relation between the distributions of stopping time and stopped sum with applications, *arXiv:1008.0116v2*.
- Chadjiconstantinidis S., Antzoulakos, D. L. and Koutras, M. V. (2000). Joint distributions of successes, failures and patterns in enumeration problems. *Advances in Applied Probability*, **32**, 866-884.
- Eryilmaz, S. (2008) Distribution of runs in a sequence of exchangeable multistate trials. *Statistics and Probability Letters*, **78**, 1505-1513.
- Eryilmaz, S. (2010). Discrete scan statistics generated by exchangeable binary trials. *Journal of Applied Probability*, **45**, 1007-1023.
- Feller, W. (1968) *An Introduction to Probability Theory and Its Applications, Vol. 1*, 3rd Edn, John Wiley, NY.
- Inoue, K. (2004). Joint distributions associated with patterns, successes and failures in a sequence of multi-state trials. *Annals of the Institute of Statistical Mathematics*, **56**, 143-168.
- Inoue, K. and Aki, S. (2007) On generating functions of waiting times and number of occurrences of compound patterns in a sequence of multistate trials. *Journal of Applied Probability*, **44**, 71-81.
- Inoue, K., Aki, S. and Hirano, K. (2011). Distributions of simple patterns in some kinds of exchangeable sequences. *Journal of Statistical Planning and Inference*, **141**, 2532-2544.
- Karlin, S. and Taylor, H. M. (1975). *A First Course in Stochastic Processes*, 2nd Edn, Academic Press.
- Mauldin, R., Sudderth, W. D. and Williams, S. C. (1992). Polya trees and random distributions. *Annals of Statistics*, **20**, 1203-1221.
- Zabell, S. L. (1982). W. E. Johnson's "Sufficientness" postulate. *Annals of Statistics*, **10**, 1091-1099.

ΚΑΝΟΝΕΣ ΔΙΑΚΟΠΗΣ ΓΙΑ ΠΟΛΥΣΤΑΔΙΑΚΕΣ ΚΛΙΝΙΚΕΣ ΔΟΚΙΜΕΣ ΦΑΣΗΣ II ΜΕ ΔΥΟ ΜΕΤΑΒΛΗΤΕΣ ΑΠΟΚΡΙΣΗΣ

Α. Σαχλάς¹, Σ. Μπερσίμης², Τ. Παπαϊωάννου³

¹Τμήμα Νοσηλευτικής, Πανεπιστήμιο Πελοποννήσου
asachlas@uop.gr

²Τμήμα Στατιστικής & Ασφαλιστικής Επιστήμης, Πανεπιστήμιο Πειραιώς
sbarsim@unipi.gr

³ Τμήμα Στατιστικής & Ασφαλιστικής Επιστήμης, Πανεπιστήμιο Πειραιώς και Τμήμα Μαθηματικών, Πανεπιστήμιο Ιωαννίνων
takpap@unipi.gr

ABSTRACT

Στην εργασία αυτή, και στη συνέχεια προηγούμενης εργασίας μας, προτείνονται κανόνες διακοπής που μπορούν να χρησιμοποιηθούν για τον πρόωρο τερματισμό κλινικών δοκιμών Φάσης II κάνοντας χρήση δύο εξαρτημένων μεταβλητών απόκρισης. Οι κανόνες αυτοί εστιάζουν στον τερματισμό των κλινικών δοκιμών όταν παρατηρούνται, σχετικά νωρίς, υψηλά ποσοστά επιθυμητών ή ανεπιθύμητων αποκρίσεων σε σχέση με το αδρανές φάρμακο (placebo) ή μια ανταγωνιστική θεραπεία. Οι νέοι κανόνες ορίζονται σε μια διμεταβλητή σειρά αποκρίσεων που προκύπτουν από μια πολυσταδιακή δοκιμή. Σε κάθε στάδιο της κλινικής δοκιμής συμμετέχουν δύο ασθενείς (ο ένας λαμβάνει τη νέα θεραπεία, ενώ ο άλλος την τυπική θεραπεία), οι οποίοι μετά τη θεραπεία εξετάζονται ως προς δύο χαρακτηριστικά. Η απόκριση των ασθενών αναμένεται να είναι διαθέσιμη χωρίς καθυστέρηση. Στην ουσία, πρόκειται για έναν ακολουθιακό σχεδιασμό τυχαιοποιημένων μπλοκ μεγέθους δύο. Εξετάζονται οι περιπτώσεις καθαρής ανωτερότητας, ως προς το ένα ή το άλλο χαρακτηριστικό, της μιας θεραπευτικής αγωγής ως προς την άλλη και παρουσίας σοβαρών (ακράιων) ανεπιθύμητων παρενεργειών. Για τη στατιστική αιτιολόγηση των κανόνων, η κατανομή του αριθμού των σταδίων μέχρι να εμφανιστούν στη σειρά συγκεκριμένοι τύποι αποκρίσεων μελετάται χρησιμοποιώντας την τεχνική της Μαρκοβιανής εμφύτευσης. Τα αριθμητικά αποτελέσματα έδειξαν πολύ καλή απόδοση των νέων κανόνων.

Keywords: Κλινικές Δοκιμές Φάσης II, Κανόνες Διακοπής, Πολυσταδιακοί Σχεδιασμοί, Διμεταβλητές Ακολουθίες Πολυσταδιακών Δοκιμών, Μαρκοβιανή Εμφύτευση.

1. ΕΙΣΑΓΩΓΗ

Στην εργασία αυτή, η οποία αποτελεί συνέχεια παλαιότερης εργασίας (Σαχλάς και συν. 2012), όπου παραπέμπουμε για μια γενικότερη εισαγωγή στο θέμα, προτείνονται κανόνες διακοπής που μπορούν να χρησιμοποιηθούν για τον πρόωρο τερματισμό κλι-

νικών δοκιμών Φάσης II κάνοντας χρήση δύο εξαρτημένων μεταβλητών απόκρισης. Οι κανόνες αυτοί εστιάζουν στον τερματισμό των κλινικών δοκιμών όταν παρατηρούνται, σχετικά νωρίς, υψηλά ποσοστά επιθυμητών ή ανεπιθύμητων αποκρίσεων σε σχέση με το αδρανές φάρμακο (placebo) ή μια ανταγωνιστική θεραπεία. Αρκετά συχνά, προκειμένου η απόκριση των ασθενών να είναι άμεσα διαθέσιμη, ακόμη και σε ασθένειες όπου η επιβίωση των ασθενών είναι το κύριο τελικό σημείο, χρησιμοποιούνται εναλλακτικές παράμετροι αξιολόγησης (surrogates) με άμεση ή γρήγορη απόκριση, όπως η έκταση της συρρίκνωσης του όγκου, η ύφεση των συμπτωμάτων, η πλήρης και μερική απόκριση, κλπ. Οι νέοι κανόνες ορίζονται σε μια διμεταβλητή σειρά αποκρίσεων που προκύπτουν από μια πολυσταδιακή δοκιμή. Σε κάθε στάδιο της κλινικής δοκιμής συμμετέχουν δύο ασθενείς (ο ένας λαμβάνει τη νέα θεραπεία, ενώ ο άλλος την τυπική θεραπεία), οι οποίοι μετά τη θεραπεία εξετάζονται ως προς δύο χαρακτηριστικά. Στην ουσία, πρόκειται για έναν ακολουθιακό σχεδιασμό τυχαιοποιημένων μπλοκ μεγέθους δύο. Εξετάζονται οι περιπτώσεις καθαρής ανωτερότητας, ως προς το ένα ή το άλλο χαρακτηριστικό, της μιας θεραπείας ως προς την άλλη και παρουσίας σοβαρών ανεπιθύμητων παρενεργειών. Για τη στατιστική αιτιολόγηση των κανόνων, η κατανομή του αριθμού των σταδίων μέχρι να εμφανιστούν στη σειρά συγκεκριμένοι τύποι αποκρίσεων μελετάται χρησιμοποιώντας την τεχνική της Μαρκοβιανής εμφύτευσης (Balakrishnan et al., 2009).

Η εργασία διαρθρώνεται ως εξής: στην Ενότητα 2 παρουσιάζεται συνοπτικά και βελτιωμένα ο νέος κανόνας διακοπής, στην Ενότητα 3 παρουσιάζονται 3 επεκτάσεις αυτού του κανόνα, ενώ στην τελευταία ενότητα αναφέρονται τα συμπεράσματα και προτάσεις για μελλοντική έρευνα.

2. Ο ΝΕΟΣ ΚΑΝΟΝΑΣ ΔΙΑΚΟΠΗΣ

Ο Σαχλάς και συν. (2012) παρουσίασαν έναν κανόνα διακοπής που μπορεί να χρησιμοποιηθεί σε συγκριτικές, πολυσταδιακές κλινικές δοκιμές με μικρό αριθμό ασθενών, όπου ο κίνδυνος από την υπό διερεύνηση θεραπεία είναι υψηλός και η θεραπεία αξιολογείται μέσω δυο μετρήσιμων χαρακτηριστικών.

Οι ασθενείς λαμβάνουν με τυχαίο τρόπο μια από δυο θεραπείες, την πειραματική ($g = E$) και τη θεραπεία αναφοράς ($g = R$). Επειδή απαιτείται μικρός αριθμός ασθενών, δυο μόνο ασθενείς εμπλέκονται σε κάθε στάδιο. Ο πρώτος ασθενής λαμβάνει με τυχαίο τρόπο είτε την πειραματική είτε τη θεραπεία αναφοράς ενώ ο δεύτερος λαμβάνει τη θεραπεία που απομένει. Σε κάθε στάδιο, η απόκριση των δυο ασθενών μετράται μετά από ένα εύλογο χρονικό διάστημα (πριν από το επόμενο στάδιο) ως προς τα δυο χαρακτηριστικά. Εάν τα χαρακτηριστικά είναι συνεχή, μπορούμε να τα μετατρέψουμε, χρησιμοποιώντας ένα σημείο αποκοπής για κάθε χαρακτηριστικό, σε δυο τυχαίες μεταβλητές Y_1^g, Y_2^g , η κάθε μια από τις οποίες αντιπροσωπεύει εάν η θεραπεία g είναι αποτελεσματική (1) ή όχι (0) όσον αφορά τα δυο χαρακτηριστικά.

Η πιθανότητα ότι η θεραπεία g είναι αποτελεσματική ως προς το Y_i θα συμβολίζεται με $\pi_{1,i}^g = P(Y_i^g = 1)$ ενώ η πιθανότητα ότι η θεραπεία g είναι αναποτελεσματική ως προς το Y_i θα συμβολίζεται με $\pi_{0,i}^g = P(Y_i^g = 0) = 1 - \pi_{1,i}^g$, $i = 1, 2$. Για κάθε

Πίνακας 1: Υλοποίηση των 12 πρώτων σταδίων μιας υποθετικής κλινικής δοκιμής

Στάδιο (s)		1	2	3	4	5	6	7	8	9	10	11	12	...
Θεραπεία R	Y_1^R	0	1	1	1	1	0	0	0	1	0	0	0	...
	Y_2^R	0	0	1	0	1	0	1	1	0	1	0	1	...
Θεραπεία E	Y_1^E	1	1	0	1	1	0	0	1	0	1	1	1	...
	Y_2^E	1	1	1	0	1	1	1	0	0	1	0	0	...
Αριθμός ασθενών		2	4	6	8	10	12	14	16	18	20	22	24	...

στάδιο s , $s = 1, 2, 3, \dots$ υπάρχουν τέσσερα αποτελέσματα που αφορούν την επιτυχία ή αποτυχία των δυο θεραπειών ως προς τα δυο χαρακτηριστικά. Ένα υποθετικό παράδειγμα ενός τέτοιου σχεδιασμού παρουσιάζεται στον Πίνακα 1. Παρατηρούμε μια τετραδιάστατη ακολουθία, όπου τα δυο πρώτα στοιχεία της αναφέρονται στις τιμές των δυο χαρακτηριστικών για τη «Θεραπεία E » ενώ τα επόμενα δυο στοιχεία αναφέρονται στις τιμές των δυο χαρακτηριστικών για τη «Θεραπεία R ». Τα αποτελέσματα σε κάθε στάδιο για όλα τα στοιχεία είναι ανεξάρτητα του χρόνου καθώς σε κάθε στάδιο χρησιμοποιείται διαφορετικό ζεύγος ασθενών. Επιπλέον, τα αποτελέσματα μεταξύ των ασθενών είναι ανεξάρτητα.

Στο πρώτο στάδιο, η «Θεραπεία R » είναι αναποτελεσματική και ως προς τα δυο χαρακτηριστικά, ενώ η «Θεραπεία E » είναι αποτελεσματική και ως προς τα δυο χαρακτηριστικά. Ανάλογα, στο έβδομο στάδιο οι δυο θεραπείες είναι αποτελεσματικές μόνο ως προς το δεύτερο χαρακτηριστικό. Είναι προφανές ότι η «Θεραπεία R » θεωρείται καλύτερη από τη «Θεραπεία E » εάν ένας μεγάλος αριθμός από 1 σε ένα τουλάχιστον από τα χαρακτηριστικά εμφανιστεί σχετικά γρήγορα στη σειρά των Y_1^R , Y_2^R . Αντίστοιχα, η «Θεραπεία E » θεωρείται καλύτερη από τη «Θεραπεία R » εάν ένας μεγάλος αριθμός από 1 σε ένα τουλάχιστον από τα χαρακτηριστικά εμφανιστεί σχετικά γρήγορα στη σειρά των Y_1^E , Y_2^E .

Συνεπώς, ο κανόνας διακοπής που μπορεί να χρησιμοποιηθεί για τον πρόωρο τερματισμό της κλινικής δοκιμής χρησιμοποιώντας τον παραπάνω σχεδιασμό απαρτίζεται από τους ακόλουθους υποκανόνες (περιπτώσεις):

- $(sr)_1$: Η μελέτη διακόπτεται υπέρ της «Θεραπείας E » όταν παρατηρηθούν συνολικά k περιπτώσεις που δείχνουν βελτίωση λόγω της «Θεραπείας E » σε τουλάχιστον ένα από τα χαρακτηριστικά **σχετικά γρήγορα, έστω πριν το στάδιο c ή**
- $(sr)_2$: Η μελέτη διακόπτεται υπέρ της «Θεραπείας R » όταν παρατηρηθούν συνολικά k περιπτώσεις που δείχνουν βελτίωση λόγω της «Θεραπείας R » σε τουλάχιστον ένα από τα χαρακτηριστικά **σχετικά γρήγορα, έστω πριν το στάδιο c ή**
- $(sr)_3$: Η μελέτη διακόπτεται με το συμπέρασμα ότι οι δυο θεραπείες είναι ισοδύναμες όταν δεν παρατηρηθούν k περιπτώσεις που δείχνουν βελτίωση λόγω της «Θεραπείας R » ή λόγω της «Θεραπείας E » **μέχρι το στάδιο c .**

Για να προσδιορίσουμε τις τιμές των c και k , ορίζουμε 2 τυχαίες μεταβλητές που μετρούν τις περιπτώσεις που δείχνουν βελτίωση λόγω της «Θεραπείας R » και της «Θεραπείας E », αντίστοιχα. Συνεπώς, στη σειρά των ενδεχομένων $\{Y_1^R, Y_2^R, Y_1^E, Y_2^E\}$, ορίζουμε μια διδιάστατη τυχαία μεταβλητή (L_s^R, L_s^E) ως εξής:

$$L_s^R = Y_1^R + Y_2^R, \quad L_s^E = Y_1^E + Y_2^E.$$

Οι L_s^R και L_s^E είναι αθροίσματα δυο μεταβλητών Bernoulli με πιθανότητες επιτυχίας $(\pi_{1,1}^R, \pi_{1,2}^R)$ και $(\pi_{1,1}^E, \pi_{1,2}^E)$, αντίστοιχα, και η (L_s^R, L_s^E) παίρνει τιμές στο $\{0, 1, 2\} \times \{0, 1, 2\}$. Οι L_s^R και L_s^E είναι ανεξάρτητες για κάθε s .

Οι πιθανότητες που σχετίζονται με τις L_s^R, L_s^E , δηλαδή οι $p_{uv} = P(L_s^R = u, L_s^E = v)$, $u, v = 0, 1, 2$, είναι οι

$$\begin{aligned} p_{00} &= \pi_{00}^R \cdot \pi_{00}^E, & p_{10} &= \pi_{10}^R \cdot \pi_{00}^E + \pi_{01}^R \cdot \pi_{00}^E, & p_{20} &= \pi_{11}^R \cdot \pi_{00}^E, \\ p_{01} &= \pi_{00}^R \cdot \pi_{01}^E + \pi_{00}^R \cdot \pi_{01}^E, & p_{11} &= \pi_{10}^R \cdot \pi_{01}^E + \pi_{01}^R \cdot \pi_{10}^E + \pi_{01}^R \cdot \pi_{01}^E + \pi_{10}^R \cdot \pi_{10}^E, & p_{21} &= \pi_{11}^R \cdot \pi_{01}^E + \pi_{11}^R \cdot \pi_{10}^E, \\ p_{02} &= \pi_{00}^R \cdot \pi_{11}^E, & p_{12} &= \pi_{10}^R \cdot \pi_{11}^E + \pi_{01}^R \cdot \pi_{11}^E, & p_{22} &= \pi_{11}^R \cdot \pi_{11}^E, \end{aligned} \quad (1)$$

με

$$\pi_{ij}^R = P(Y_1^R = i, Y_2^R = j), i, j = 0, 1 \text{ και } \pi_{i'j'}^E = P(Y_1^E = i', Y_2^E = j'), i', j' = 0, 1.$$

Επάνω στην ακολουθία των L_s^R, L_s^E ορίζουμε μια τυχαία μεταβλητή S που καταμετρά τον αριθμό των σταδίων μέχρι να πραγματοποιηθεί ένας από τους υπό-κανόνες. Στην πράξη, η S μετρά τον αριθμό των σταδίων μέχρι την πρώτη εμφάνιση είτε k περιπτώσεων βελτίωσης λόγω της «Θεραπείας R » τουλάχιστον σε ένα από τα 2 χαρακτηριστικά, είτε k περιπτώσεων βελτίωσης λόγω της «Θεραπείας E » σε οποιοδήποτε από τα χαρακτηριστικά. Είναι μια διακριτή μεταβλητή με τιμές $1, 2, \dots$ και ισούται με n εάν και μόνο εάν με τη συμπλήρωση του n -οστού σταδίου, έχουμε

$$\sum_{s=1}^n L_s^R \geq k \quad \text{ή} \quad \sum_{s=1}^n L_s^E \geq k$$

και κανένα από τα ενδεχόμενα δεν έχει εμφανιστεί πριν το n -οστό στάδιο.

Η διαδικασία πρόωρου τερματισμού της κλινικής δοκιμής μπορεί να αναπαρασταθεί ως ένας έλεγχος υπόθεσης που βασίζεται στην κατανομή της S . Η κρίσιμη τιμή k και οι πιθανότητες $\pi_{1,1}^R, \pi_{1,2}^R, \pi_{1,1}^E, \pi_{1,2}^E$ είναι παράμετροι της κατανομής της S , ενώ το c είναι ένα κατάλληλο ποσοστημόριο της κατανομής της S . Η μηδενική υπόθεση του ελέγχου είναι

$$H_0 : \pi_{1,1}^R = \pi_{1,1}^E \text{ και } \pi_{1,2}^R = \pi_{1,2}^E,$$

δηλαδή οι πιθανότητες επιτυχίας των δυο θεραπειών είναι ίσες και στα δυο χαρακτηριστικά. Η κύρια εναλλακτική υπόθεση είναι η

$$H_1^1 : \pi_{1,1}^R < \pi_{1,1}^E \text{ και } \pi_{1,2}^R < \pi_{1,2}^E \quad \text{ή} \quad \pi_{1,1}^R = \pi_{1,1}^E \text{ και } \pi_{1,2}^R < \pi_{1,2}^E \quad \text{ή} \quad \pi_{1,1}^R < \pi_{1,1}^E \text{ και } \pi_{1,2}^R = \pi_{1,2}^E,$$

ενώ μια δευτερεύουσα εναλλακτική υπόθεση είναι η

$$H_1^2 : \pi_{1,1}^R > \pi_{1,1}^E \text{ και } \pi_{1,2}^R > \pi_{1,2}^E \quad \text{ή} \quad \pi_{1,1}^R = \pi_{1,1}^E \text{ και } \pi_{1,2}^R > \pi_{1,2}^E \quad \text{ή} \quad \pi_{1,1}^R > \pi_{1,1}^E \text{ και } \pi_{1,2}^R = \pi_{1,2}^E.$$

Ο αλγόριθμος για να εφαρμόσουμε τον παραπάνω έλεγχο είναι ο εξής:

1. Όρισε τις πιθανότητες επιτυχίας της H_0 .
2. Όρισε την επιθυμητή πιθανότητα σφάλματος τύπου I, α .
3. Υπολόγισε διάφορες υποψήφιες τιμές του c μέσω της $P(S \leq c | k, H_0) = \alpha$, για διάφορες τιμές του k και συγκεκριμένες τιμές των $\pi_{1,j}^R$ και $\pi_{1,j}^E$, $j = 1, 2$ στην H_0 .
4. Επέλεξε το k έτσι ώστε να έχεις ένα επιθυμητό επίπεδο ισχύος γ , για την απαιτούμενη εναλλακτική υπόθεση. Αυτό θα καθορίσει το c .
5. Τρέξε την κλινική δοκιμή, παρακολουθώντας τα $\sum_{s=1}^n L_s^R$, $\sum_{s=1}^n L_s^E$.
 - Εάν για πρώτη φορά στο n -οστό στάδιο με $n \leq c$, ισχύει $\sum_{s=1}^n L_s^R \geq k$ τότε η δοκιμή διακόπτεται υπέρ της “Θεραπείας R” ή εάν $\sum_{s=1}^n L_s^E \geq k$ τότε η δοκιμή διακόπτεται υπέρ της “Θεραπείας E”.
 - Εάν $n > c$ και τα δυο αθροίσματα είναι μικρότερα του k τότε η δοκιμή διακόπτεται συμπεραίνοντας ότι οι δυο θεραπείες είναι ισοδύναμες.

Τα αθροίσματα $\sum_{s=1}^n L_s^R$, $\sum_{s=1}^n L_s^E$, $n = 1, 2, \dots$ αποτελούν έναν τυχαίο περίπατο στο διακριτό επίπεδο και η κατανομή της S μπορεί να μελετηθεί μέσω της τεχνικής της Μαρκοβιανής εμφύτευσης, όπου ένα απορροφητικό φράγμα που εξαρτάται από το k εισάγεται στον τυχαίο περίπατο.

Η τεχνική της Μαρκοβιανής εμφύτευσης απαιτεί τον ορισμό ενός κατάλληλου χώρου καταστάσεων $\Omega = \{\alpha_1, \alpha_2, \dots, \alpha_d\}$ για την Μαρκοβιανή αλυσίδα $\{V_s, s = 0, 1, 2, \dots\}$. Για να ορίσουμε την ακριβή συνάρτηση κατανομής της S χρησιμοποιούμε τη σχέση

$$P(S \leq n) = P(V_n = \alpha_d) = \pi'_0 \Lambda_0^n \mathbf{e}_d,$$

όπου α_d είναι μια κατάλληλη απορροφητική κατάσταση, π'_0 είναι το διάνυσμα (γραμμή) των αρχικών πιθανοτήτων της Μαρκοβιανής αλυσίδας, Λ_0 είναι ο πίνακας πιθανοτήτων μετάβασης και $\mathbf{e}'_d$ είναι ένα $1 \times d$ διάνυσμα της μορφής $(0, 0, \dots, 1)$. Η κατανομή πιθανότητας της S υπολογίζεται μέσω της

$$P(S = n) = \pi'_1 (\Lambda_1)^{n-1} \mathbf{h},$$

όπου π_1 είναι ένα $(d-1) \times 1$ διάνυσμα (στήλη) που περιέχει όλα τα στοιχεία του διανύσματος των αρχικών πιθανοτήτων π_0 εκτός από το τελευταίο και Λ_1 είναι ο πίνακας Λ_0 χωρίς την τελευταία γραμμή και στήλη. Το διάνυσμα $\mathbf{h}$ υπολογίζεται εύκολα ως $\mathbf{h} = \mathbf{1} - \Lambda_1 \mathbf{1}$, όπου $\mathbf{1}$ είναι ένα $(d-1) \times 1$ διάνυσμα (στήλη) μονάδων (ο Λ_0 είναι στοχαστικός πίνακας). Το διάνυσμα $\mathbf{h}$ περιέχει τις πιθανότητες μετάβασης από την αντίστοιχη κατάσταση στην απορροφητική.

Ο κατάλληλος χώρος καταστάσεων αποτελείται από $d = k^2 + 1$ στοιχεία και είναι της μορφής $\Omega = \{(j_1, j_2) : j_1, j_2 = 0, 1, 2, \dots, k-1\} \cup \{\alpha_d\}$, όπου α_d είναι η απορροφητική κατάσταση. Το j_1 παρακολουθεί το $\sum_{s=1}^n L_s^R$ στη διμεταβλητή σειρά (L_s^R, L_s^E) , και το j_2 παρακολουθεί το $\sum_{s=1}^n L_s^E$ στη διμεταβλητή σειρά (L_s^R, L_s^E) .

Η αλυσίδα $\{V_s, s = 0, 1, 2, \dots\}$ ορίζεται ως εξής:

$$V_n = \begin{cases} (j_1, j_2), & \text{εάν } \sum_{s=1}^n L_s^R = j_1, \sum_{s=1}^n L_s^E = j_2 \\ \alpha_d, & \text{εάν } \sum_{s=1}^n L_s^R \geq k \text{ ή } \sum_{s=1}^n L_s^E \geq k, \end{cases}$$

με $V_0 = (0,0)$, από σύμβαση, $n = 1, 2, \dots$. Η κατάσταση α_d αναπαριστά όλες τις πιθανές απορροφητικές καταστάσεις της αλυσίδας.

Ο πίνακας των πιθανοτήτων μετάβασης ορίζεται ως

$$\mathbf{\Lambda}_0 = [P(V_s = \alpha_j | V_{s-1} = \alpha_i)]_{(k^2+1) \times (k^2+1)}.$$

Είναι ένας $(k^2 + 1) \times (k^2 + 1)$ πίνακας της μορφής

$$\mathbf{\Lambda}_0 = \begin{bmatrix} \mathbf{\Lambda}_1 & \mathbf{h} \\ \mathbf{0} & 1 \end{bmatrix}$$

και μπορεί επίσης να γραφεί ως

$$\mathbf{\Lambda}_0 = \left[\begin{array}{cccccc|c} \mathbf{A}_0 & \mathbf{B}_0 & \mathbf{\Gamma}_0 & & & & \\ & \mathbf{A}_1 & \mathbf{B}_1 & \mathbf{\Gamma}_1 & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & \mathbf{A}_{k-2} & \mathbf{B}_{k-2} & \mathbf{\Gamma}_{k-2} & \\ & & & & \mathbf{A}_{k-1} & \mathbf{B}_{k-1} & \\ & & & & & \mathbf{A}_k & \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{array} \right]$$

όπου $\mathbf{A}_i = \mathbf{A}$, $i = 0, 1, \dots, k$, $\mathbf{B}_i = \mathbf{B}$, $i = 0, 1, \dots, k-1$, και $\mathbf{\Gamma}_i = \mathbf{\Gamma}$, $i = 0, 1, \dots, k-2$ είναι $k \times k$ πίνακες της μορφής

$$\mathbf{A}_i = \mathbf{A} = \begin{bmatrix} (j_i, 0) & (j_i, 1) & (j_i, 2) & \dots & (j_i, k-3) & (j_i, k-2) & (j_i, k-1) \\ (j_i, 0) & p_{00} & p_{01} & p_{02} & & & \\ (j_i, 1) & & p_{00} & p_{01} & \ddots & & \\ (j_i, 2) & & & p_{00} & \ddots & & \\ \vdots & & & & \ddots & & \\ (j_i, k-3) & & & & & p_{00} & p_{01} & p_{02} \\ (j_i, k-2) & & & & & & p_{00} & p_{01} \\ (j_i, k-1) & & & & & & & p_{00} \end{bmatrix},$$

$$\mathbf{B}_i = \mathbf{B} = \begin{bmatrix} (j_i, 0) & (j_i+1, 0) & (j_i+1, 1) & (j_i+1, 2) & \dots & (j_i+1, k-3) & (j_i+1, k-2) & (j_i+1, k-1) \\ (j_i, 0) & & p_{10} & p_{11} & p_{12} & & & \\ (j_i, 1) & & & p_{10} & p_{11} & \ddots & & \\ (j_i, 2) & & & & p_{10} & \ddots & & \\ \vdots & & & & & \ddots & & \\ (j_i, k-3) & & & & & & p_{10} & p_{11} & p_{12} \\ (j_i, k-2) & & & & & & & p_{10} & p_{11} \\ (j_i, k-1) & & & & & & & & p_{10} \end{bmatrix},$$

$$\mathbf{\Gamma}_i = \mathbf{\Gamma} = \begin{bmatrix} (j_i, 0) & (j_i+2, 0) & (j_i+2, 1) & (j_i+2, 2) & \dots & (j_i+2, k-3) & (j_i+2, k-2) & (j_i+2, k-1) \\ (j_i, 0) & & p_{20} & p_{21} & p_{22} & & & \\ (j_i, 1) & & & p_{20} & p_{21} & \ddots & & \\ (j_i, 2) & & & & p_{20} & \ddots & & \\ \vdots & & & & & \ddots & & \\ (j_i, k-3) & & & & & & p_{20} & p_{21} & p_{22} \\ (j_i, k-2) & & & & & & & p_{20} & p_{21} \\ (j_i, k-1) & & & & & & & & p_{20} \end{bmatrix},$$

$j_i = 0, 1, \dots, k-1$ για $i = 1, 2$.

Η αριθμητική διερεύνηση έδειξε πολύ καλή επίδοση του προτεινόμενου κανόνα (Σαχλάς και συν., 2012). Ο κανόνας διακόπτει την κλινική δοκιμή αρκετά γρήγορα

με υψηλή ισχύ. Ο αναμενόμενος αριθμός σταδίων (η μέση τιμή της κατανομής S που προκύπτει μέσω της τεχνικής της Μαρκοβιανής εμφύτευσης) είναι μικρός και δεδομένου ότι σε κάθε στάδιο συμμετέχουν μόνο 2 ασθενείς, το μέγεθος του δείγματος που απαιτείται είναι μικρό. Η μέθοδος είναι ευέλικτη καθώς ο ερευνητής μπορεί να επιλέξει μεταξύ διαφόρων τιμών του k και της ισχύος που επιθυμεί. Αυτό σημαίνει ότι ο ερευνητής μπορεί να αυξήσει την τιμή του k με σκοπό να αυξήσει την τιμή του ελέγχου.

3. ΕΠΕΚΤΑΣΕΙΣ ΤΟΥ ΚΑΝΟΝΑ

Στην ενότητα αυτή θα παρουσιάσουμε 3 επεκτάσεις του παραπάνω κανόνα. Συγκεκριμένα, θα παρουσιάσουμε την περίπτωση των εξαρτημένων μεταβλητών απόκρισης, της ανωτερότητας της μιας θεραπείας και του τερματισμού της κλινικής δοκιμής όταν παρουσιαστεί μια σοβαρή ανεπιθύμητη παρενέργεια.

3.1 Ο ΚΑΝΟΝΑΣ ΜΕ ΕΞΑΡΤΗΜΕΝΕΣ ΜΕΤΑΒΛΗΤΕΣ ΑΠΟΚΡΙΣΗΣ

Υποθέτουμε ότι τα ζεύγη των τυχαίων μεταβλητών Y_1^g, Y_2^g είναι εξαρτημένα για $g = R, E$ με από κοινού πιθανότητες

$$P(Y_1^g = 0, Y_2^g = 0), P(Y_1^g = 1, Y_2^g = 0), P(Y_1^g = 0, Y_2^g = 1), P(Y_1^g = 1, Y_2^g = 1).$$

Μπορούμε να εργασθούμε ακριβώς με τον ίδιο τρόπο, όπως στην περίπτωση των ανεξάρτητων μεταβλητών με τη μόνη διαφορά ότι, τώρα, οι πιθανότητες $\pi_{ij}^R, \pi_{ij}^E, i = 0, 1, j = 0, 1$ θεωρούνται γνωστές και δεν είναι γινόμενα περιθώριων πιθανοτήτων.

3.2 ΑΝΩΤΕΡΟΤΗΤΑ ΤΗΣ ΜΙΑΣ ΘΕΡΑΠΕΙΑΣ

Στην περίπτωση αυτή λαμβάνουμε υπόψιν μόνο τις περιπτώσεις όπου η μια θεραπεία είναι ανώτερη της άλλης σε ένα χαρακτηριστικό. Ανώτερη ως προς ένα χαρακτηριστικό, θεωρείται μια θεραπεία όταν αυτή είναι αποτελεσματική ως προς το συγκεκριμένο χαρακτηριστικό, ενώ η άλλη είναι αναποτελεσματική. Εάν η νέα θεραπεία είναι ανώτερη της τυπικής ως προς το ένα χαρακτηριστικό τότε η νέα θεραπεία σκοράρει έναντι της τυπικής. Εάν οι δυο θεραπείες είναι εξίσου αποτελεσματικές ως προς το ίδιο χαρακτηριστικό, καμιά δεν είναι ανώτερη και συνεπώς καμιά θεραπεία δεν σκοράρει. Γενικά, ο όρος ανωτερότητα στην εργασία αυτή σχετίζεται με τον τρόπο που αντιμετωπίζεται η ισοπαλία (tie).

Από τον Πίνακα 2 παρατηρούμε ότι στο 2^ο στάδιο και οι δυο θεραπείες είναι επιτυχείς και στα δυο χαρακτηριστικά ($Y_1^R = Y_1^E = 1$ και $Y_2^R = Y_2^E = 1$), συνεπώς καμιά θεραπεία δεν σκοράρει. Το ενδεχόμενο ($Y_1^R = Y_1^E = 1$ και $Y_2^R = Y_2^E = 1$) είναι ισοδύναμο με το ενδεχόμενο ($Y_1^R = Y_1^E = 0$ και $Y_2^R = Y_2^E = 0$) που εμφανίζεται στο 1^ο στάδιο. Κανένα από αυτά τα ενδεχόμενα δεν προκαλεί μετάβαση της Μαρκοβιανής αλυσίδας ($\sum_{s=1}^n L_s^R = 0, \sum_{s=1}^n L_s^E = 0$). Επίσης, η Μαρκοβιανή αλυσίδα δεν κάνει μετάβαση στο 3^ο στάδιο ($Y_1^R = Y_1^E = 1$ και $Y_2^R = Y_2^E = 0$)

Πίνακας 2: Αριθμός επιτυχιών και αθροιστικός αριθμός επιτυχιών για κάθε θεραπεία σε κάθε στάδιο της πραγματοποίησης της κλινικής δοκιμής με την Επέκταση 2

Στάδιο (s)	1	2	3	4	5	6	7	8	9	10	...
Y_1^R	0	1	1	0	0	1	0	1	1	0	...
Y_2^R	0	1	0	0	0	0	1	0	0	1	...
Y_1^E	0	1	1	0	1	0	1	0	1	0	...
Y_2^E	0	1	0	1	1	0	0	1	1	1	...
L_s^R	0	0	0	0	0	1	1	1	0	0	...
L_s^E	0	0	0	1	2	0	1	1	1	0	...
$\sum_{s=1}^n L_s^R$	0	0	0	0	0	1	2	3	3	3	...
$\sum_{s=1}^n L_s^E$	0	0	0	1	3	3	4	5	6	6	...

Πίνακας 3: Ισχύς, αναμενόμενος αριθμός σταδίων και c για $k = 5$ έως $k = 8$ για συγκριμένες H_0 και H_1

		Επέκταση 1				Επέκταση 2			
		$\pi_{1,1}^E$	$\pi_{1,2}^E$	5.2%	40.6%	c	5.2%	40.6%	c
$\pi_{1,1}^R$	$\pi_{1,2}^R$			5.2%	40.6%		5.2%	40.6%	
$k = 5$	5.2%	5.2%		0.040 (37.671)	0.916 (6.984)	11	0.044 (39.396)	0.922 (7.404)	12
$k = 6$	5.2%	5.2%		0.042 (42.694)	0.939 (7.676)	13	0.045 (44.727)	0.940 (8.141)	14
$k = 7$	5.2%	5.2%		0.043 (53.472)	0.991 (9.375)	20	0.049 (56.051)	0.993 (9.932)	22
$k = 8$	5.2%	5.2%		0.044 (59.577)	0.995 (10.194)	23	0.049 (62.523)	0.996 (10.803)	25

και στο 10^ο στάδιο ($Y_1^R = Y_1^E = 0$ και $Y_2^R = Y_2^E = 1$). Η Μαρκοβιανή αλυσίδα μετατοπίζεται στις περιπτώσεις που στο ίδιο χαρακτηριστικό οι δυο θεραπείες έχουν διαφορετικό αποτέλεσμα (στάδια 4, 5, 6, 7, 8, και 9).

Στην περίπτωση αυτή έχουμε

$$\begin{aligned}
 p_{00} &= \pi_{00}^R \cdot \pi_{00}^E + \pi_{01}^R \cdot \pi_{01}^E + \pi_{10}^R \cdot \pi_{10}^E + \pi_{11}^R \cdot \pi_{11}^E, & p_{11} &= \pi_{10}^R \cdot \pi_{01}^E + \pi_{01}^R \cdot \pi_{10}^E, & p_{22} &= 0, \\
 p_{01} &= \pi_{00}^R \cdot \pi_{10}^E + \pi_{00}^R \cdot \pi_{01}^E + \pi_{10}^R \cdot \pi_{11}^E + \pi_{01}^R \cdot \pi_{11}^E, & p_{02} &= \pi_{00}^R \cdot \pi_{11}^E, & p_{12} &= 0, \\
 p_{10} &= \pi_{10}^R \cdot \pi_{00}^E + \pi_{00}^R \cdot \pi_{10}^E + \pi_{11}^R \cdot \pi_{01}^E + \pi_{11}^R \cdot \pi_{10}^E, & p_{20} &= \pi_{11}^R \cdot \pi_{00}^E, & p_{21} &= 0.
 \end{aligned}$$

Για να είναι ο Λ_0 στοχαστικός, θα πρέπει να ισχύει $p_{21} = p_{12} = p_{22} = 0$. Εάν θέσουμε $p_{11} = 0$, $p_{01} = \pi_{00}^R \cdot \pi_{10}^E + \pi_{00}^R \cdot \pi_{01}^E$, $p_{10} = \pi_{10}^R \cdot \pi_{00}^E + \pi_{00}^R \cdot \pi_{10}^E$, και βάλουμε τις υπόλοιπες ποσότητες στην p_{00} , ο σχεδιασμός μετατρέπεται σε καθαρή ανωτερότητα (υπεροχή) της μιας θεραπείας.

Συγκρίνοντας τις 2 επεκτάσεις, από τον Πίνακα 3 παρατηρούμε ότι η προφανής επιλογή για το k είναι 7, ανεξαρτήτως σχεδιασμού, καθώς αυτό είναι το μικρότερο

k που δίνει $\gamma \geq 0.95$. Το c ισούται με 20, για την Επέκταση 1, που αντιστοιχεί σε $\alpha = 0.043$, ενώ ισούται με 22, για την Επέκταση 2, που αντιστοιχεί σε $\alpha = 0.049$. Ο αναμενόμενος αριθμός σταδίων είναι περίπου 10, και για τους δυο σχεδιασμούς, οπότε ο αναμενόμενος αριθμός ασθενών που πρέπει να συμμετάσχουν στην κλινική δοκιμή είναι 20. Γενικά, και οι δυο σχεδιασμοί έχουν σχεδόν την ίδια επίδοση.

3.3 ΠΑΡΟΥΣΙΑ ΣΟΒΑΡΗΣ ΑΝΕΠΙΘΥΜΗΤΗΣ ΠΑΡΕΝΕΡΓΕΙΑΣ

Στην περίπτωση αυτή, η κλινική δοκιμή σταματά όταν εμφανιστεί μια σοβαρή ανεπιθύμητη παρενέργεια (ή ακόμα και θάνατος). Οι Y_1^g , Y_2^g είναι εξαρτημένες με από κοινού πιθανότητες

$$P(Y_1^g = 0, Y_2^g = 0), P(Y_1^g = 1, Y_2^g = 0), P(Y_1^g = 0, Y_2^g = 1), P(Y_1^g = 1, Y_2^g = 1), P(Y_1^g = *, Y_2^g = *),$$

όπου το * σημαίνει την εμφάνιση παρενέργειας. Οι L_S^R και L_S^E παίρνουν τις τιμές 0, 1, 2, και *.

Εκτός από τις πιθανότητες p_{uv} της Εξίσωσης (1) έχουμε και τις

$$\begin{aligned} p_{0*} &= \pi_{00}^R \cdot \pi_{**}^E, & p_{1*} &= \pi_{10}^R \cdot \pi_{**}^E + \pi_{01}^R \cdot \pi_{**}^E, & p_{2*} &= \pi_{11}^R \cdot \pi_{**}^E, \\ p_{*0} &= \pi_{**}^R \cdot \pi_{00}^E, & p_{*1} &= \pi_{**}^R \cdot \pi_{01}^E + \pi_{**}^R \cdot \pi_{10}^E, & p_{*2} &= \pi_{**}^R \cdot \pi_{11}^E, \\ p_{**} &= \pi_{**}^R \cdot \pi_{**}^E, \end{aligned}$$

όπου $\pi_{**}^g = P(Y_1^g = *, Y_2^g = *)$, $g = R, E$.

Δεν χρειάζεται όμως να υπολογίσουμε τις παραπάνω πιθανότητες, καθώς εμφανίζονται μόνο στο διάνυσμα $\mathbf{h}$, που περιλαμβάνει τις πιθανότητες μετάβασης στην απορροφητική κατάσταση (η εμφάνιση μιας παρενέργειας οδηγεί την Μαρκοβιανή αλυσίδα στην απορροφητική κατάσταση), και υπολογίζεται μέσω της $\mathbf{h} = \mathbf{1} - \mathbf{\Lambda}_1 \mathbf{1}$.

Εκτός από τους υπο-κανόνες $(sr)_1 - (sr)_3$, έχουμε επιπλέον τους

- $(sr)_4$: Η μελέτη διακόπτεται υπέρ της «Θεραπείας E » όταν εμφανιστεί μια σοβαρή παρενέργεια εξαιτίας της «Θεραπείας R » ή
- $(sr)_5$: Η μελέτη διακόπτεται υπέρ της «Θεραπείας R » όταν εμφανιστεί μια σοβαρή παρενέργεια εξαιτίας της «Θεραπείας E ».

Η S ισούται με n εάν και μόνο εάν μετά το n -οστό στάδιο, έχουμε

$$\sum_{s=1}^n L_s^R \geq k \quad \text{ή} \quad \sum_{s=1}^n L_s^E \geq k \quad \text{ή} \quad L_s^R = * \quad \text{ή} \quad L_s^E = *$$

και κανένα από τα παραπάνω ενδεχόμενα δεν έχει εμφανιστεί πριν από το n -οστό στάδιο.

Από τον Πίνακα 4 παρατηρούμε ότι η προφανής επιλογή είναι $k = 5$, καθώς είναι η μικρότερη τιμή του k που δίνει $\gamma \geq 0.90$. Το c ισούται με 9 που αντιστοιχεί σε $\alpha = 0.047$. Ο αριθμός των αναμενόμενων σταδίων είναι περίπου 6, οπότε ο αναμενόμενος αριθμός ασθενών είναι 12. Η αριθμητική διερεύνηση έδειξε ότι η επίδοση του κανόνα διακοπής επηρεάζεται σημαντικά από την πιθανότητα εμφάνισης της παρενέργειας.

Πίνακας 4: Ισχύς, αναμενόμενος αριθμός σταδίων και c για $k = 5$ έως 8 για την Επέκταση 3 για συγκεκριμένη H_0 και H_1

	$\pi_{1,1}^E$		4.7%	45.3%	c
	$\pi_{1,1}^R$	$\pi_{1,2}^R$	$\pi_{1,2}^E$	4.7%	45.3%
$k = 5$	4.7%	4.7%	0.047 (37.969)	0.904 (5.855)	10
$k = 6$	4.7%	4.7%	0.049 (42.419)	0.911 (6.316)	10
$k = 7$	4.7%	4.7%	0.049 (52.324)	0.948 (7.623)	13
$k = 8$	4.7%	4.7%	0.047 (57.455)	0.923 (8.146)	13

4. ΣΥΖΗΤΗΣΗ

Στο άρθρο αυτό προτείνουμε έναν νέο κανόνα διακοπής που βασίζεται στην κατανομή του αριθμού των σταδίων μέχρι ένας συγκεκριμένος αριθμός ασθενών να εμφανίσει βελτίωση σε 2 χαρακτηριστικά. Συγκεκριμένα, θεωρούμε μια τυχαιοποιημένη, συγκριτική, πολυσταδιακή κλινική δοκιμή Φάσης II όπου οι ασθενείς κατανέμονται τυχαία σε δυο θεραπείας, μια πειραματική και μια τυπική. Σε κάθε στάδιο της κλινικής δοκιμής εμπλέκονται μόνο 2 ασθενείς. Ο πρώτος παίρνει τυχαία τη μια θεραπεία και ο δεύτερος τη θεραπεία που απομένει. Σε κάθε στάδιο μετράται, ύστερα από ένα λογικό μικρό διάστημα (πριν το επόμενο στάδιο), η απόκριση των ασθενών όσον αφορά 2 χαρακτηριστικά, η οποία αποτελεί 2 διακριτές τυχαίες μεταβλητές με τιμές 1 (επιτυχία) και 0 (αποτυχία).

Η αριθμητική διερεύνηση έδειξε αρκετά καλή απόδοση των επεκτάσεων. Όπως και στον αρχικό κανόνα, οι επεκτάσεις διακόπτουν την κλινική δοκιμή αρκετά γρήγορα (μετά από λίγα στάδια και συνεπώς εμπλέκει λίγους ασθενείς) με μεγάλη ισχύ. Εάν θεωρήσουμε τους υπό-κανόνες $(sr)_1 - (sr)_3$ ανεξάρτητα, η μέθοδός μας μπορεί να χρησιμοποιηθεί και για single-arm μελέτες.

Ως μελλοντικές σκέψεις για διερεύνηση αποτελούν η περίπτωση που η κλινική μελέτη τερματίζεται όταν συνολικά k_1 περιπτώσεις ασθενών εμφανίζουν βελτίωση και στα δυο χαρακτηριστικά ή όταν συνολικά k_2 περιπτώσεις ασθενών δεν εμφανίζουν βελτίωση και στα δυο χαρακτηριστικά, η περίπτωση που η εμφάνιση παρενέργειας τιμωρεί την θεραπεία που την προκαλεί και η περίπτωση που θεωρούμε άνισα βάρη για τα δυο χαρακτηριστικά.

ΑΝΑΦΟΡΕΣ

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** (1), 66-81.

Σαχλάς, Α. Μπερσίμης, Σ. και Παπαϊωάννου, Τ. (2012). Κανόνες Διακοπής για Πολυσταδιακές Κλινικές Δοκιμές Φάσης II, *Πρακτικά 25ου Πανελληνίου Συνεδρίου Στατιστικής*, Βόλος.



ΣΥΜΠΛΗΡΩΜΕΝΗ ΓΟΝΙΜΟΤΗΤΑ ΚΑΙ ΠΙΘΑΝΟΤΗΤΕΣ ΔΙΕΥΡΥΝΣΗΣ ΟΙΚΟΓΕΝΕΙΑΣ ΤΟΥ ΠΛΗΘΥΣΜΟΥ ΤΗΣ ΕΛΛΑΔΑΣ ΚΑΤΑ ΕΠΙΠΕΔΟ ΕΚΠΑΙΔΕΥΣΗΣ, ΑΣΤΙΚΟΤΗΤΑ ΚΑΙ ΝΟΜΟ

Κλέων Τσίμπος

Τμήμα Στατιστικής & Ασφαλιστικής Επιστήμης, Πανεπιστήμιο Πειραιώς
cleon@unipi.gr

ΠΕΡΙΛΗΨΗ

Στην εργασία αυτή χρησιμοποιούνται απογραφικά δεδομένα αριθμού των παιδιών που έχουν φέρει στον κόσμο κατά τη διάρκεια της ζωής τους γυναίκες οι οποίες πρακτικά έχουν ολοκληρώσει την αναπαραγωγική τους περίοδο (45-49 ετών) με σκοπό αφενός την εκτίμηση του επιπέδου συμπληρωμένης γονιμότητας του πληθυσμού της Ελλάδας και αφετέρου την εκτίμηση των πιθανοτήτων διεύρυνσης οικογένειας δηλαδή την πιθανότητα μιας γυναίκας που έχει ήδη i παιδιά να αποκτήσει ένα επιπλέον ($i+1$). Μελετώνται οι γενεές γυναικών 1900-1904 έως 1950-1954 και για τις τελευταίες εξετάζονται διαφοροποιήσεις κατά επίπεδο εκπαίδευσης και αστικότητα σε εθνικό και περιφερειακό επίπεδο.

1. ΕΙΣΑΓΩΓΗ

Κατά την μεταπολεμική περίοδο οι χρονολογικοί δείκτες γονιμότητας στη χώρα μας ακολούθησαν μία καθοδική πορεία ως αποτέλεσμα σημαντικών οικονομικών κοινωνικών και πολιτιστικών αλλαγών και διαφοροποιήσεων στην αναπαραγωγική συμπεριφορά του πληθυσμού. Μέχρι το τέλος της δεκαετίας του 1970 ο μέσος αριθμός παιδιών ανά γυναίκα περιστρέφονταν γύρω από το επίπεδο του 2,3 αλλά από το 1980 και μετά παρατηρήθηκαν έντονες καθοδικές τάσεις με αποτέλεσμα το 2001 η γονιμότητα να φθάσει σε επίπεδο κάτω του 1,3 παιδιών το οποίο στη διεθνή βιβλιογραφία χαρακτηρίζεται ως «the lowest-low level» (Billari 2005). Τα επόμενα χρόνια (περίπου μέχρι το 2008) υπήρξε μία μικρή ανάκαμψη της γονιμότητας, κυρίως λόγω των μεταναστών, αλλά και πάλι οι δείκτες δεν ξεπέρασαν το επίπεδο αναπλήρωσης των γενεών που είναι τα 2,1 παιδιά ανά γυναίκα (Tsimbos 2008).

Οι τάσεις και διαφοροποιήσεις των χρονολογικών δεικτών γονιμότητας του πληθυσμού της Ελλάδας έχουν διεξοδικά εξεταστεί. Ωστόσο λιγότερο συχνά έχουν αναλυθεί στοιχεία που αφορούν τον αριθμό των παιδιών που έχουν αποκτήσει οι γυναίκες σε όλη τη διάρκεια της αναπαραγωγικής τους ζωής. Το δημογραφικό αυτό μέγεθος είναι γνωστό ως *συμπληρωμένη γονιμότητα* (Completed Fertility Rate) οποίο επίσης μπορεί να εκφραστεί πιο αναλυτικά συναρτήσει των αποκαλούμενων *πιθανοτήτων διεύρυνσης της οικογένειας* (Parity Progression Ratios, PPRs) μία τεχνική που πρώτος πρότεινε ο Henry το 1953 (Henry 1976) και που μπορεί να θεωρηθεί ως μία επέκταση του ιστορικού ημερολογίου της γονιμότητας (Wunsch & Termote 1978). Οι PPRs εκφράζουν, βάσει αναδρομικών στοιχείων, την πιθανότητα που έχει μία γυναίκα που έχει ήδη i παιδιά να αποκτήσει και ένα επιπλέον ($i+1$). Υπάρχουν αρκετοί τρόποι υπολογισμού τέτοιων πιθανοτήτων ανάλογα με το είδος των διαθέσιμων στατιστικών στοιχείων, όμως πιο συχνά βασίζεται σε απογραφικά δεδομένα που αφορούν γυναίκες που έχουν ολοκληρώσει την αναπαραγωγική τους ζωή (Newell 1988, Siegel & Swanson 2004).

Στην εργασία αυτή υπολογίζονται πιθανότητες διεύρυνσης της οικογένειας στην Ελλάδα συνδυάζοντας στοιχεία διαδοχικών απογραφών για τις γενέες γυναικών των ετών 1900-1904 έως 1950-1954 και εξετάζονται διαφοροποιήσεις κατά επίπεδο εκπαίδευσης, νομό και επίπεδο αστικότητας.

2. ΣΤΑΤΙΣΤΙΚΟ ΥΛΙΚΟ

Στην εργασία αυτή χρησιμοποιούνται ληξιαρχικά στοιχεία των περιόδων 1980-1982, 1990-1992 και 2000-2002 που αφορούν στον αριθμό των γεννήσεων ζώντων κατά ηλικία της μητέρας (έχουν υιοθετηθεί οι γνωστές και καθιερωμένες πενταετείς ομάδες ηλικιών 15-19, 20-25, ... 45-49). Τα στοιχεία αυτά διατίθενται σε εθνικό και περιφερειακό επίπεδο (κατά νομό). Οι γεννήσεις από μητέρες νεότερες των 15 ετών έχουν συμπεριληφθεί στην πρώτη πενταετή ηλικιακή ομάδα (15-19) και οι γεννήσεις από μητέρες μεγαλύτερες των 50 ετών έχουν συμπεριληφθεί στο τελευταίο πενταετές κλιμάκιο (45-49). Για να αποφευχθούν τυχαίες ετήσιες κυμάνσεις που επηρεάζουν τις τιμές των αντίστοιχων δημογραφικών δεικτών, έχουν ληφθεί υπόψη οι μέσοι όροι των γεννήσεων των τριετιών γύρω από τα κεντρικά έτη των απογραφών. Επιπλέον χρησιμοποιούνται πληθυσμιακά δεδομένα από τις απογραφές των ετών 1971, 1981, 1991 και 2001 και αφορούν στον αριθμό των γυναικών ηλικίας 45-49 ετών (δηλαδή αυτών που στην ουσία έχουν ολοκληρώσει την αναπαραγωγική τους περίοδο) και στον αριθμό των παιδιών που δηλώνουν οι γυναίκες αυτές ότι έχουν αποκτήσει κατά τη διάρκεια όλης της αναπαραγωγικής τους ζωής. Τα απογραφικά δεδομένα των ετών 1971-1991 διατίθενται σε εθνικό επίπεδο ενώ στην περίπτωση της απογραφής του έτους 2001 τα δεδομένα παρέχονται επίσης σε επίπεδο νομού, κατά επίπεδο εκπαίδευσης και αστικότητα. Για τους σκοπούς της ανάλυσης διακρίνονται τρία επίπεδα εκπαίδευσης, και συγκεκριμένα: αγράμματες γυναίκες και γυναίκες που έχουν τελειώσει δημοτικό (χαμηλό επίπεδο), γυναίκες που έχουν απολυτήριο λυκείου (μεσαίο επίπεδο) και γυναίκες που έχουν μετα-λυκειακό δίπλωμα, πτυχίο ΑΕΙ/ΤΕΙ ή

μεταπτυχιακό τίτλο (υψηλό επίπεδο). Σε ό,τι αφορά την αστικότητα, διακρίνονται δύο επίπεδα, δηλαδή αστικές περιοχές (περιοχές με πληθυσμό 10.000 άτομα και άνω) και οι υπόλοιπες περιοχές (χαρακτηρίζονται εδώ ως αγροτικές) οι οποίες περιλαμβάνουν περιοχές με πληθυσμό κάτω των 10.000 ατόμων.

3. ΜΕΘΟΔΟΙ ΑΝΑΛΥΣΗΣ

Στη δημογραφική ανάλυση της γονιμότητας διακρίνονται δύο είδη δεικτών, οι χρονολογικοί και οι γενεαλογικοί δείκτες.

Οι χρονολογικοί δείκτες υπολογίζονται με βάση τις ληξιαρχικές καταγραφές γεννήσεων που πραγματοποιούνται και συγκεντρώνονται κατά τη διάρκεια ενός ημερολογιακού έτους. Οι δείκτες αυτοί εκφράζουν τη γονιμότητα γυναικών διαφορετικών ηλικιών δηλαδή γυναικών που ενώ έφεραν στον κόσμο ένα παιδί την ίδια χρονιά γεννήθηκαν σε διαφορετικά έτη, δηλαδή ανήκουν σε διαφορετικές γενεές (ή κοορτές). Οι γυναίκες διαφορετικών κοορτών προφανώς έχουν διαφορετικές εμπειρίες γονιμότητας, έχουν εκτεθεί σε διαφορετικό κίνδυνο θανάτου και γενικά έχουν διαφορετική συμπεριφορά σχετικά με τη δόμηση της οικογένειας. Αυτό αποτελεί ένα μεθοδολογικό πρόβλημα που αφορά κυρίως την ερμηνεία και προοπτική των αποτελεσμάτων. Έστω ${}_5b_x$ ο αριθμός των γεννήσεων από μητέρες ηλικίας x έως $x+5$ ετών και ${}_5w_x$ ο αριθμός των γυναικών στις ίδιες ηλικίες. Ο ειδικός κατά ηλικία δείκτης γονιμότητας (συμβολίζεται με ${}_5f_x$ και συνήθως εκφράζεται επί 1000 γυναικών) είναι:

$${}_5f_x = \frac{{}_5b_x}{{}_5w_x} \times 1000$$

Ο πλέον γνωστός και καθιερωμένος χρονολογικός - συνθετικός δείκτης γονιμότητας είναι ο Δείκτης Ολικής Γονιμότητας (*TFR*) που στην περίπτωση πενταετών ομάδων ηλικιών ορίζεται ως εξής:

$$TFR = 5 \times \sum_{15}^{50} {}_5f_x$$

Ο δείκτης *TFR* εκφράζει τον αριθμό των παιδιών που αναμένεται να φέρουν στον κόσμο 1000 γυναίκες μέχρι το τέλος της αναπαραγωγικής τους ζωής με την προϋπόθεση ότι θα παραμείνουν αμετάβλητοι οι ειδικοί κατά ηλικία δείκτες γονιμότητας και χωρίς να ληφθεί υπόψη η επίδραση της θνησιμότητας. Σύμφωνα με τη διεθνή βιβλιογραφία, δείκτες *TFR* $\geq 2,1$ ανά γυναίκα χαρακτηρίζονται ως δείκτες που εξασφαλίζουν επίπεδο αναπλήρωσης των γενεών, δείκτες *TFR* με τιμές από 1,3 έως 2,1 χαρακτηρίζονται ως δείκτες χαμηλής γονιμότητας ενώ δείκτες *TFR* $< 1,3$ είναι γνωστοί στη βιβλιογραφία ως «lowest low fertility rates» (Billari 2005).

Μία γενεαλογική διάσταση της γονιμότητας μπορεί να ληφθεί με τα απογραφικά δεδομένα που αναφέραμε πιο πάνω από τα οποία μπορεί να προκύψουν (α) το ποσοστό άτεκνων γυναικών (ως αποτέλεσμα των μακροχρόνιων τάσεων γονιμότητας, εν μέρει γαμηλιότητας και ορισμένων βιολογικών επιδράσεων), (β) οι πιθανότητες διεύρυνσης της οικογένειας (*PPR*) και (γ) ο δείκτης συμπληρωμένης γονιμότητας (*CFR*).

Οι δημογραφικοί αυτοί δείκτες υπολογίζονται με βάση τον αριθμό των παιδιών που δηλώνουν οι γυναίκες ηλικίας 45-49 ετών ότι απέκτησαν κατά τη διάρκεια της αναπαραγωγικής τους ζωής (αριθμός παιδιών: 0, 1, 2, , ..., 9, 10+).

Έστω:

m_{1+} = ο αριθμός γυναικών με τουλάχιστον 1 παιδί

m_{2+} = ο αριθμός γυναικών με τουλάχιστον 2 παιδιά

και γενικά

m_{i+} = ο αριθμός γυναικών με τουλάχιστον i παιδιά

Με τα στοιχεία αυτά υπολογίζονται οι εξής πιθανότητες:

- πιθανότητα μιας γυναίκας με 1 παιδί να αποκτήσει δεύτερο: $a_1 = m_{2+} / m_{1+}$
- πιθανότητα μιας γυναίκας με 2 παιδιά να αποκτήσει τρίτο: $a_2 = m_{3+} / m_{2+}$

ή γενικά:

- πιθανότητα μιας γυναίκας με i παιδιά να αποκτήσει $i+1$ παιδιά: $a_i = m_{i+1} / m_{i+}$

Υπολογίζεται επίσης και η:

- πιθανότητα άτεκνης γυναίκας να αποκτήσει το πρώτο της παιδί:
 $a_0 = m_{1+} / w_{15-49}$

όπου με w_{45-49} συμβολίζεται ο συνολικός αριθμός γυναικών ηλικίας 45-49 ετών.

Με βάση τις πιθανότητες διεύρυνσης της οικογένειας (a_i , $i = 0,1,2,\dots$) υπολογίζεται ο Δείκτης Συμπληρωμένης Γονιμότητας (συμβολίζεται με *CFR*) που εκφράζει τον μέσο αριθμό παιδιών ανά γυναίκα ηλικίας 45-49 ετών:

$$CFR = a_0 + a_0 \cdot a_1 + a_0 \cdot a_1 \cdot a_2 + a_0 \cdot a_1 \cdot a_2 \cdot a_3 + \dots$$

και που ισοδυναμεί με την απλή σχέση:

$$CFR = \frac{k}{w_{45-49}}$$

όπου k ο συνολικός αριθμός παιδιών που δηλώνουν οι γυναίκες ηλικίας 45-49 ετών ότι απέκτησαν κατά τη διάρκεια της αναπαραγωγικής τους ζωής.

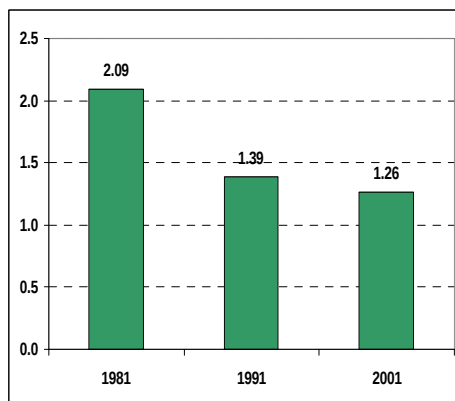
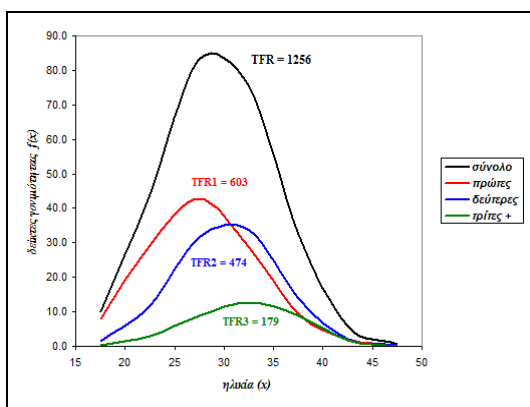
4. ΑΠΟΤΕΛΕΣΜΑΤΑ

Όπως σε όλους του ανθρώπινους πληθυσμούς έτσι και στην Ελλάδα οι ειδικοί κατά ηλικία δείκτες γονιμότητας παρουσιάζουν μία μονοκόρυφη καμπύλη με θετική ασυμμετρία, αν και τα επίπεδα και τα κατά ηλικία πρότυπα γονιμότητας διαφοροποιούνται σημαντικά ανάλογα με τη σειρά γέννησης (Σχήμα 1α). Ο μέσος αριθμός παιδιών ανά γυναίκα μειώθηκε από 2,09 το 1980 (που εξασφαλίζει οριακά την αναπλήρωση των γενεών) στο εξαιρετικά χαμηλό επίπεδο του 1,26 το 2001 (Σχήμα 1β). Σε περιφερειακό επίπεδο η μείωση των δεικτών γονιμότητας πραγματοποιήθηκε επίσης προς την ίδια κατεύθυνση σε όλους τους νομούς της χώρας (Σχήμα 1γ). Είναι χαρακτηριστικό ότι ενώ το 1980 σε 30 από τους 51 νομούς της χώρας ο δείκτης TFR ήταν υψηλότερος του 2,1 το 1991 και το 2001 σε όλους τους νομούς της Ελλάδος η γονιμότητα ήταν κάτω του ορίου αναπλήρωσης. Ειδικότερα το 2001, σε 36 από τους 51 νομούς η γονιμότητα ήταν κάτω του 1,3 (lowest-low). Αν και οι τάσεις αυτές αποδίδονται σε ένα πλέγμα κοινωνικο-οικονομικών και πολιτιστικών παραγόντων, από καθαρά δημογραφικής πλευράς οι εξελίξεις αυτές συνδέονται και με την αύξηση της μέσης ηλικίας της μητέρας, όχι μόνο για το σύνολο των γεννήσεων αλλά και κατά σειρά γέννησης (Σχήμα 1δ).

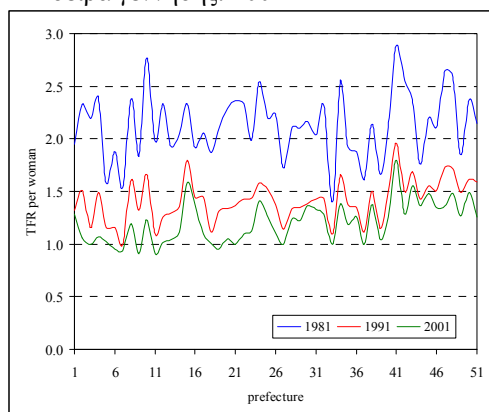
Οι πιθανότητες PPRs κατά σειρά παιδιού για όλες τις εξεταζόμενες γενεές (1900-1904 έως 1950-1954) παρουσιάζονται στο Σχήμα 2α. Για όλες τις εξεταζόμενες γενεές οι πιθανότητες αυτές αντανακλούν μειούμενο επίπεδο γονιμότητας για κάθε σειρά παιδιού αλλά στις πολύ παλαιές γενεές (1900-1904 έως 1920-1924) η μορφή των καμπύλων των PPRs διαφέρει από εκείνη των νεότερων γενεών (1930-1934 έως 1950-1954). Το γεγονός αυτό μπορεί εν μέρει να οφείλεται στις μεταβολές που έχουν επέλθει στην αναπαραγωγική συμπεριφορά των γυναικών αλλά και εν μέρει στην ποιότητα των διαθέσιμων στατιστικών στοιχείων. Είναι χαρακτηριστικό ότι οι καμπύλες, για τις γυναίκες που γεννήθηκαν μετά το 1930, εκφράζουν πιθανότητες διεύρυνσης της οικογένειας που παρουσιάζουν μείωση μέχρι και την πιθανότητα απόκτησης δεύτερου προς τρίτο παιδί αλλά από τη στιγμή απόκτησης ενός τρίτου παιδιού τα ζευγάρια δεν διστάζουν να αποκτήσουν και ένα επιπλέον. Εξετάζοντας συνδυαστικά τα απογραφικά δεδομένα των ετών 1971-2001 προκύπτει ότι ο συγκεφαλαιωτικός δείκτης συμπληρωμένης γονιμότητας για τις γυναίκες που γεννήθηκαν στην αρχή του 20^{ου} αιώνα (1900-1904) ήταν 3,28 παιδιά ανά γυναίκα ηλικίας 45-49 ετών. Ο δείκτης αυτός παρουσίασε σταδιακή αλλά συνεχή μείωση. Για πρώτη φορά ο δείκτης έπεσε κάτω από το επίπεδο των δύο παιδιών (1,95) για τις γενεές 1940-1944 και κατέληξε στο επίπεδο των 1,91 παιδιών για τις γυναίκες που γεννήθηκαν μετά τον πόλεμο (1950-54) Σχήμα 2β. Οι τάσεις αυτές συμβαδίζουν με την πορεία των γενεαλογικών δημογραφικών δεικτών (Billari 2005). Σε ό,τι αφορά τις γυναίκες που γεννήθηκαν την πενταετία 1950-1954 υπάρχει μία σαφής αρνητική σχέση μεταξύ του μέσου αριθμού παιδιών και επιπέδου εκπαίδευσης με τις γυναίκες

χαμηλότερου εκπαιδευτικού επιπέδου να δηλώνουν ότι έχουν αποκτήσει περισσότερα παιδιά (2,13) σε σχέση με τις γυναίκες υψηλότερου εκπαιδευτικού επιπέδου (1,61). Αν και η μορφή των καμπύλων που εκφράζουν τις πιθανότητες PPRs κατά επίπεδο εκπαίδευσης είναι παρόμοια με εκείνη του συνολικού αριθμού γυναικών υπάρχει μία σαφής διαφορά στα επίπεδα γονιμότητας μεταξύ των τριών αυτών εκπαιδευτικών επιπέδων (Σχήμα 3α και 3β). Σύμφωνα με τα στοιχεία της απογραφής του 2001 οι γυναίκες ηλικίας 45-49 που διέμεναν σε αγροτικές περιοχές δήλωναν μεγαλύτερο αριθμό παιδιών από εκείνες των αστικών περιοχών και το εύρημα αυτό ισχύει τόσο σε εθνικό επίπεδο όσο και σε επίπεδο γεωγραφικού διαμερίσματος (Σχήμα 4).

Σχήμα 1 Χρονολογική γονιμότητα στην Ελλάδα 1981–2001

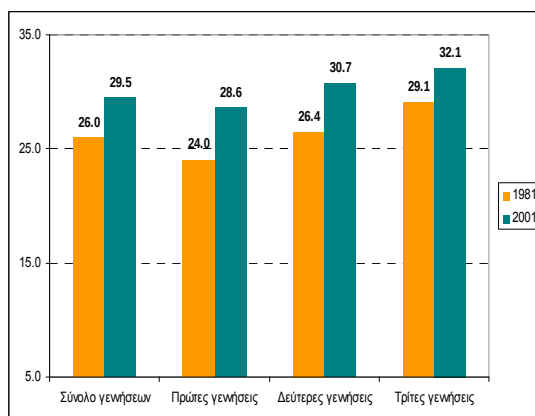


(α) Ειδικοί δείκτες γονιμότητας (%) κατά ηλικία της μητέρας και σειρά γέννησης: 2001



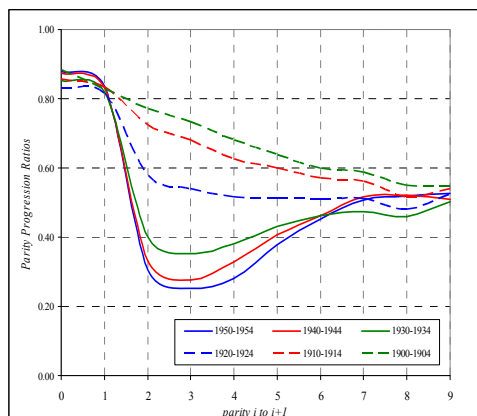
(γ) Δείκτες ολικής γονιμότητας κατά νομό: 1981 – 2001

(β) Δείκτες ολικής γονιμότητας ανά γυναίκα: 1981 - 2001

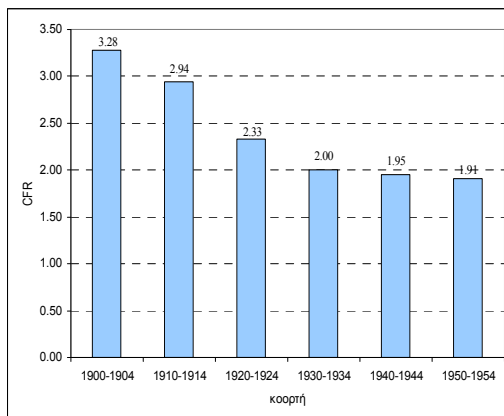


(δ) Μέση ηλικία μητέρας κατά σειρά γέννησης: 1981 και 2001

Σχήμα 2 Πιθανότητες διεύρυνσης της οικογένειας και μέσο επίπεδο συμπληρωμένης οικογένειας στην Ελλάδα: γενεές 1900-1904 έως 1950-1954

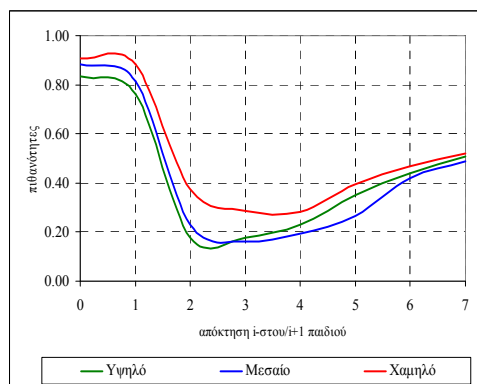


(α) Πιθανότητες διεύρυνσης της οικογένειας κατά σειρά παιδιού και γενεά

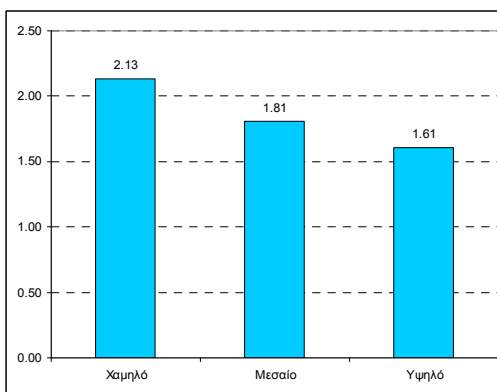


(β) Δείκτες συμπληρωμένης οικογένειας κατά γενεά

Σχήμα 3 Πιθανότητες διεύρυνσης της οικογένειας και μέσο επίπεδο συμπληρωμένης οικογένειας στην Ελλάδα κατά επίπεδο εκπαίδευσης: γενεά 1950-1954

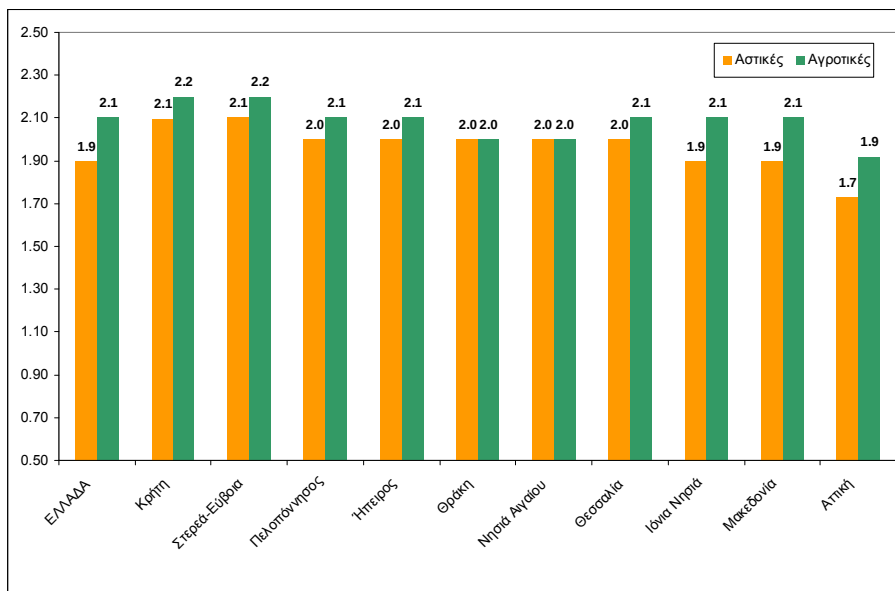


(α) Πιθανότητες διεύρυνσης της οικογένειας κατά σειρά παιδιού και επίπεδο εκπαίδευσης



(β) Δείκτες συμπληρωμένης οικογένειας κατά επίπεδο εκπαίδευσης

Σχήμα 4 Μέσο επίπεδο συμπληρωμένης οικογένειας στην Ελλάδα κατά επίπεδο εκπαίδευσης κατά αστικές και αγροτικές περιοχές: γενεά 1950-1954



5. ΣΥΜΠΕΡΑΣΜΑΤΑ

Στην εργασία αυτή εξετάζονται αναδρομικές παρατηρήσεις που αφορούν τον αριθμό των παιδιών που έχουν αποκτήσει οι γυναίκες ηλικίας 45-49 σε όλη τη διάρκεια της αναπαραγωγικής τους ζωής που προέρχονται από επίσημες στατιστικές (απογραφές του πληθυσμού της Ελλάδας). Από την συνεξέταση των στοιχείων των απογραφών των ετών 2001, 1991, 1981 και 1971 προκύπτει ότι ο μέσος αριθμός παιδιών που απέκτησαν οι γυναίκες που γεννήθηκαν την περίοδο 1900-1904 ήταν 3,28 και μειώθηκε σταδιακά στο επίπεδο των 1,91 παιδιών για τις γυναίκες που γεννήθηκαν την περίοδο 1950-1954. Η πιο λεπτομερής ανάλυση των δεδομένων της απογραφής του έτους 2001 έδειξε ότι ο μέσος αριθμός παιδιών ανά γυναίκα ηλικίας 45-49 ετών ήταν: (α) υψηλότερος στις αγροτικές περιοχές (2,12) σε σύγκριση με τις αστικές (1,87), (β) υψηλότερος μεταξύ των γυναικών χαμηλού εκπαιδευτικού επιπέδου (2,13) σε σχέση με τις γυναίκες μεσαίου (1,81) και υψηλού (1,61) εκπαιδευτικού επιπέδου, (γ) το μεγαλύτερο επίπεδο συμπληρωμένης γονιμότητας παρατηρείται στο νομό Αιτωλοακαρνανίας (2,35) και το χαμηλότερο στο νομό Αττικής (1,74), και (δ) οι πιθανότητες που έχουν οι γυναίκες με δύο παιδιά να αποκτήσουν ένα τρίτο είναι σημείο καμπής για την πληθυσμιακή πολιτικής της χώρας διότι αποτελούν παράγοντα που προσδιορίζει σε μεγάλο βαθμό τις περιφερειακές διαφοροποιήσεις του επιπέδου γονιμότητας του πληθυσμού της χώρας σε περιφερειακό επίπεδο.

ABSTRACT

In this paper we use census data on reported numbers of children ever born alive to women aged 45-49 years who have practically completed their reproductive age. The aim of the study is to estimate the completed fertility rates and the progression parity ratios of these women. We examine birth cohorts 1900-1904 to 1950-1954; for the latter, the above aforementioned fertility measures are estimated by educational level, urbanization level and administrative department (prefecture level).

ΑΝΑΦΟΡΕΣ

- Billari FC (2005). Europe and its fertility, from low to lowest low. *National Institute Economic Review* **194**: 56-73.
- Henry L (1976). *Population, Analysis and Models*. London: Edward Arnold.
- Newell C (1988). *Methods and Models in Demography*, London: Belhaven Press.
- Siegel JS & Swanson DA (2004). *The Methods and Materials of Demography*. New York: Elsevier – Academic Press.
- Tsimbos, C. (2008). Immigrant and Native Fertility in Greece: New Estimates and Population Prospects (2005-2025). *Population Review* **47** (2): 67-84.
- Wunsch GJ & Termote MG (1978). *Introduction to Demographic Analysis, Principles and Methods*. London: Plenum Press.

ΣΥΜΜΕΤΡΙΚΕΣ ΔΙΑΚΡΙΤΕΣ ΚΑΤΑΝΟΜΕΣ ΚΑΙ ΣΥΝΤΕΛΕΣΤΗΣ ΜΕΤΑΒΛΗΤΟΤΗΤΑΣ

Νικόλαος Φαρμάκης

Τμήμα Μαθηματικών Α.Π.Θ.

farmakis@math.auth.gr

ΠΕΡΙΛΗΨΗ

Ο Συντελεστής Μεταβλητότητας έχει αναδειχθεί σε πολύ χρήσιμο εργαλείο στη μελέτη των κατανομών των τυχαίων μεταβλητών, ιδίως των συνεχών (Farmakis, 2003, 2010 και Farmakis, Makris, 2011), όπου από ένα σχετικά μεγάλο δείγμα (με μέγεθος $n > 45$ συνήθως) μπορεί να βγουν ασφαλή συμπεράσματα για την απόδοση της συνάρτησης πυκνότητας πιθανότητας (σ.π.π.) σε μορφή πολυωνυμική. Κρίσιμο σημείο είναι ο εκθέτης k του πολυωνύμου που περιγράφει την σ.π.π. Ο εκθέτης αυτός αποδείχτηκε ότι είναι συνάρτηση απλής μορφής του Συντελεστή Μεταβλητότητας. Αυτά όλα ισχύουν για τις συνεχείς κατανομές και στην παρούσα εργασία γίνεται προσπάθεια να χρησιμοποιήσουμε και τον συντελεστή μεταβλητότητας για να περιγράψουμε μία διακριτή κατανομή. Μάλιστα σε μία προσπάθεια σύνδεσης με τα προηγούμενα των συνεχών κατανομών προσπαθούμε να συνδέσουμε τις διακριτές κατανομές με κάποιες συνεχείς που να αποτελούν, τρόπον τινά, το συνεχές ανάλογο της εκάστοτε διακριτής που μελετούμε. Μία διακριτή τυχαία μεταβλητή π.χ. που παίρνει τιμές στο διάστημα $[a, b]$ μελετάται μέσα από ένα δείγμα αλλά και θεωρητικά και υπολογίζονται στατιστικά και παράμετροι, όπως μέση τιμή, διασπορά, τυπική απόκλιση, συντελεστής μεταβλητότητας και μετά προσδιορίζεται μία σ.π.π. μιας συνεχούς τυχαίας μεταβλητής με τις ίδιες τιμές των παραμέτρων και παρουσιάζεται η πλήρης μορφή της συνεχούς με πολυωνυμική μορφή. Στην παρούσα εργασία περιοριζόμαστε σε συμμετρικές κατανομές.

Λέξεις κλειδιά: Κατανομές, Διακριτή Κατανομή, Συνάρτηση Πυκνότητας Πιθανότητας, Συντελεστής Μεταβλητότητας, Συνάρτηση Κατανομής.

1. ΕΙΣΑΓΩΓΗ

Θεωρούμε γνωστή την έννοια της τυχαίας μεταβλητής (τμ) X και των παραμέτρων της μέση τιμή $E(X) = \mu$, διασπορά $VarX = \sigma^2$, τυπική απόκλιση σ . Ειδικά αν όλες οι τιμές που παίρνει η υπό μελέτη τμ X είναι θετικές (ή τουλάχιστον του ιδίου σημείου όλες)

μπορούμε να έχουμε την σχετικά γνωστή πλέον έννοια του *συντελεστή μεταβλητότητας* C_v που είναι ο λόγος της τυπικής απόκλισης προς τη μέση τιμή: $C_v = \sigma/\mu$. Ο συντελεστής μεταβλητότητας (ΣΜ) είναι μία πολύ χρήσιμη έννοια που μπορεί να δώσει μεγάλη ποσότητα χρήσιμης πληροφορίας από δεδομένα δειγματικά ή και απογραφικά. Είναι όμως λίγο δύσχρηστος ως λόγος δύο παραμέτρων με κατανομή όχι πάντα απλής μορφής και αυτό ήταν η αιτία που δεν τον πολυχρησιμοποιούσαν οι θεωρητικοί Μαθηματικοί ή/ και Στατιστικοί.

Τον χρησιμοποιούσαν από πολύ παλιά οι Βιολόγοι, οι Γιατροί, οι Οικονομολόγοι, οι Κοινωνιολόγοι, οι Περιβαλλοντολόγοι, και άλλοι πολλοί ερευνητές. Τους παρείχε πληροφορία σχετικά με την κατανομή των τιμ που μελετούσαν, ιδίως ως προς το σχήμα της γραφικής παράστασης της συνάρτησης (πυκνότητας) πιθανότητας και κατ' επέκταση της συνάρτησης κατανομής. Τις έννοιες *συνάρτηση κατανομής* (σ.κ.) $F(x)$, *συνάρτηση πιθανότητας* (σ.π. για τις διακριτές τιμ) $f(x)$ και *συνάρτηση πυκνότητας πιθανότητας* (σ.π.π. για τις συνεχείς τιμ) $f(x)$, τις θεωρούμε επίσης γνωστές.

Οι συνεχείς τιμ με συμμετρική μορφή σ.π.π. την

$$f(x) = \begin{cases} h \cdot (x - a)^k, & x \in [a, \frac{a+b}{2}] \\ h \cdot (b - x)^k, & x \in [\frac{a+b}{2}, b], \\ 0, & \text{αλλού} \end{cases} \quad k \neq -1, \quad h = \frac{2^k \cdot (k+1)}{(b-a)^{k+1}} \quad (1.1)$$

παρέχουν την ευκολία να μπορεί να υπολογιστεί ο εκθέτης k με τη βοήθεια μόνο του ΣΜ. Αυτό σημαίνει ότι μπορεί να εκτιμηθεί η τιμή του k και του h μέσα από ένα δείγμα, απ' όπου θα εκτιμηθεί η τιμή του ΣΜ από το $\hat{C}_v = \frac{s}{\bar{x}}$. Στη συνέχεια έχουμε

το $\lambda = C_v^{-2}$ και τέλος το $k = \frac{-5 + \sqrt{1 + 8 \cdot \lambda}}{2} > -1$, Farmakis (2003).

Πρακτικότερα: Αν από σχετικό με το δείγμα ιστόγραμμα διαπιστωθεί το συμμετρικό σχήμα της σ.π.π. της τιμ X , προσεγγίζουμε τη σ.π.π. της X με την (1.1) και από την τιμή του ΣΜ έχουμε εκτίμηση για k και h . Στην εργασία Farmakis (2003) δίνεται και μέθοδος εκτίμησης του κατά πόσον έχουμε «καλή προσαρμογή» των δεδομένων του δείγματος προς τα θεωρητικά δεδομένα που δίνει η ολοκλήρωση της (1.1) στο $[a, b]$. Η καλή προσαρμογή κρίνεται με χρήση της δοκιμασίας χ^2 .

Στην παρούσα εργασία θα ασχοληθούμε με τα διακριτά δεδομένα και με το κατά πόσο μπορεί να συσχετιστούν με τις συνεχείς τιμ. Θα γίνει χρήση του ΣΜ και των δυνατοτήτων ου μην αλλά και των περιορισμών που εισάγει.

2. ΔΙΑΚΡΙΤΟ ΣΥΜΜΕΤΡΙΚΟ ΜΟΝΤΕΛΟ ΓΙΑ ΤΗ σ.π.

Ως διακριτό συμμετρικό μοντέλο σ.π. για μια τμ X νοείται εκείνο που η τμ X παίρνει τιμές συμμετρικές ως προς κέντρο το σημείο $u = \frac{a+b}{2}$. Αυτό σημαίνει ότι για τις τιμές $u-x$, $u+x$ ισχύει $f(u-x_i)=f(u+x_i)$ και για αριθμήσιμο πλήθος τιμών x_i ισχύει $f(u-x_i)>0$.

Σε μια τέτοια τμ X υπολογίζουμε για τον πληθυσμό μεγέθους $n=2\mu+2$: Μέση τιμή $\bar{X}$, διασπορά και τυπική απόκλιση s^2 και s και τον ΣΜ $C_{-y} = s/\bar{X}$. (Τα άκρα a και b του διαστήματος όπου η τμ X παίρνει τις τιμές της περιλαμβάνονται στις $n=2\mu+2$ τιμές του δείγματος $\mu=1,2,3,\dots$).

Στη συνέχεια και με τη γνωστή από την εργασία Farmakis (2003) μέθοδο, υπολογίζουμε τις τιμές του εκθέτη k και του συντελεστή h . Με βάση τα k και h γνωρίζουμε τη σ.π. μιας συνεχούς συμμετρικής κατανομής που θα μπορούσε να εκληφθεί ως το συνεχές ισοδύναμο της σ.π. της διακριτής τμ X , που είναι υπό μελέτη.

Αυτή η σκέψη μας επιτρέπει (κατ' ελάχιστο) να ταξινομήσουμε της διακριτές τμ με συμμετρική σ.π. όπως έγινε και με τις συνεχείς, Farmakis (2003).

Είναι μεγάλο το πλήθος των διακριτών τμ X που έχουν συμμετρική σ.π. αλλά με βάση τα παραπάνω μπορούμε να έχουμε μια ιεράρχησή τους και μάλιστα μπορεί να ισχύσει ο παρακάτω ορισμός:

«Εάν δύο διακριτές τμ X και Y μελετηθούν με βάση τον ΣΜ και οι προκύπτοντες εκθέτες είναι k_1 και k_2 αντίστοιχα και συνδέονται με $k_1 < k_2$, τότε θα λέμε ότι η X προηγείται τη Y και θα συμβολίζουμε $X < Y$ ».

Είναι γνωστό ότι ο εκθέτης $k > -1$. Η τιμή μάλιστα -1 είναι για τον εκθέτη k ένα είδος απόλυτου μηδέν, Farmakis (2003). Η ποσότητα δηλαδή $k+1$ είναι μία κλασματική τμ (ratio variable) και επιτρέπεται να διαιρούνται δύο οποιεσδήποτε τιμές της και να λαμβάνονται υπόψη αναλογίες μεταξύ τους. Αυτό φυσικά διευκολύνει επίσης και τον προγραμματισμό H/Y σε σχέση με τον εκθέτη k . Μετά όλα τα παραπάνω μία μελέτη συμμετρικής διακριτής κατανομής θα τα κάνει ευκρινέστερα.

3. ΔΙΑΚΡΙΤΗ ΟΜΟΙΟΜΟΡΦΗ ΚΑΤΑΝΟΜΗ

Μία κατανομή συχνά πολύ χρήσιμη (και εύχρηστη άλλωστε) είναι η διακριτή ομοιόμορφη κατανομή (δ.ο.κ.). Είναι συμμετρική και αποτελεί πολύ καλό παράδειγμα, για την κατανόηση όσων περιλαμβάνονται στην προηγούμενη παράγραφο.

Διακριτή ομοιόμορφη κατανομή είναι εκείνη που η αντίστοιχη τμ X παίρνει τις τιμές της στο διάστημα $[a, \beta]$ με βάση τον εξής κανόνα:

Οι τιμές a και β περιλαμβάνονται στις τιμές της δ.ο.κ. και ενδιάμεσά του υπάρχουν άλλες 2μ σε πλήθος τιμές της X . Συνολικά έχουμε $2\mu+2$ διαδοχικές τιμές της X που οι

εκάστοτε δύο διαδοχικές απέχουν μεταξύ τους απόσταση $\delta=(\beta-\alpha)/(2\mu+1)$. Οι τιμές αυτές είναι οι

$$\alpha, \alpha+\delta, \alpha+2\delta, \alpha+3\delta, \dots, \alpha+(i-1)\delta, \dots, \alpha+(2\mu+1)\delta=\beta \quad (3.1)$$

και έχουν όλες την ίδια πιθανότητα, ίση με $1/(2\mu+2)$, να είναι η παρατηρούμενη τιμή της τμ X .

Είναι προφανές ότι η μέση τιμή της X στην (3.1) είναι (και λόγω συμμετρίας) ίση με $E(X)=(\alpha+\beta)/2$. (3.2)

Η διασπορά $V(X)$ στην (3.1) με $2\mu+2$ τιμές, $\mu=1,2,3,\dots$ δίνεται από τη σχέση (3.3) και προκύπτει εύκολα μετά από μερικούς υπολογισμούς ως:

$$V(X)=\frac{(n+1)\cdot(\beta-\alpha)^2}{12\cdot(n-1)}, \quad n=2\mu+2 \quad (3.3)$$

και η τυπική απόκλιση ίση με

$$S=\frac{(\beta-\alpha)}{2}\cdot\sqrt{\frac{n+1}{3\cdot(n-1)}}, \quad n=2\mu+2. \quad (3.4)$$

Ο ΣΜ της τμ X είναι ο λόγος των ποσοτήτων (3.4) προς (3.2) και για τις $2\mu+2$ τιμές έχουμε

$$C_{-v}=\sqrt{\frac{n+1}{3\cdot(n-1)}}\cdot\frac{\beta-\alpha}{\beta+\alpha} \quad (3.5)$$

ή

$$C_{-v}=\sqrt{\frac{n+1}{3\cdot(n-1)}}, \quad \text{αν } \alpha=0 \quad (3.6)$$

Σημείωση-Διερεύνηση: Κανονικά είναι $C_{-v}\leq 1$ και το ίσον ισχύει για την τετριμμένη περίπτωση όπου $n=2$. Αν όμως είναι $n=4,6,\dots,2\mu+2$, τότε έχουμε την ισχύ της ανισότητας. Αυτά για $\alpha=0$. Με $\alpha>0$ έχουμε την ανισότητα πάντα. Αυτά όλα σημαίνουν ότι είναι δυνατόν να έχουμε σύνολο παρατηρήσεων ακόμη και το τετριμμένο $\{\alpha, \beta\}$. ■

Παρατήρηση: Η κατανομή της τμ X με τιμές, όπως φαίνεται στην (3.1) τείνει προς την ομοιόμορφη κατανομή, δηλαδή ο εκθέτης k τείνει στο 0, Farmakis (2003), καθώς το πλήθος των παρατηρήσεων αυξάνει.

Αυτό προκύπτει άμεσα από την (3.6) και τη χρήση της ενδιάμεσης παραμέτρου, Farmakis (2003)

$$\lambda = C_{-v}^{-2} = \frac{3 \cdot (n-1)}{(n+1)} \quad (3.7)$$

που στη συνέχεια δίνει τον εκθέτη της αντίστοιχης συνεχούς (συμμετρικής) κατανομής, τον

$$k(n) = \frac{-5 + \sqrt{1+8 \cdot \lambda}}{2} = \frac{-5 + \sqrt{1 + \frac{24(n-1)}{(n+1)}}}{2} = \frac{-5 + \sqrt{\frac{25n-23}{n+1}}}{2}. \quad (3.8)$$

(βλέπε και Παράρτημα για τον εκθέτη k).

Από την (3.8) προκύπτει ότι το $\lim_{n \rightarrow \infty} k = 0$. Ακριβέστερα το $k \xrightarrow{n \rightarrow \infty} 0^-$.

■

Τα αποτελέσματα που προκύπτουν για διάφορες τιμές του k και καθώς το n τείνει στο άπειρο φαίνονται στον παρακάτω Πίνακα 3.1:

Πίνακας 3.1

n	2	4	6	10	100	1000		$\rightarrow \infty$
C_{-v}	1	0.7454	0.6831	0.6383	0.5832	0.5779		$\rightarrow 1/\sqrt{3}$
λ	1	1.8	2.1429	2.4545	2.9406	2.9940		$\rightarrow 3^-$
$k(n)$	-1	-0.5379	-0.3703	-0.2286	-0.0239	-0.0024		$\rightarrow 0^-$

4. ΑΛΛΕΣ ΔΙΑΚΡΙΤΕΣ ΣΥΜΜΕΤΡΙΚΕΣ ΚΑΤΑΝΟΜΕΣ

Η δ.ο.κ. χαρακτηρίζεται από την εμφάνιση των παρατηρήσεων σε ισαπέχουσες μεταξύ τους θέσεις στο διάστημα $[a, \beta]$. Αυτό φαίνεται καθαρά και στην (3.1). Υπάρχει όμως ένας τεράστιος αριθμός διακριτών συμμετρικών κατανομών όπου δεν έχουμε την ίδια «πυκνότητα» παρατηρήσεων στο $[a, \beta]$. Μπορούμε να φανταστούμε τουλάχιστον αριθμήσιμους πλήθους τέτοιες κατανομές που εύκολα μπορούμε να τις μελετήσουμε μέσα από παραδείγματα αρχικά και από γενίκευση στη συνέχεια. Μερικές έχουν αραιά τις τιμές στο κέντρο και πυκνότερα στα άκρα του διαστήματος $[a, \beta]$, $-1 < k < 0$. Άλλες αντίθετα έχουν πυκνότερα τις τιμές στο κέντρο και αραιότερα στα άκρα του $[a, \beta]$, $k > 0$. Η δ.ο.κ. είναι ανάμεσα στις δύο ομάδες και μοιάζει να είναι το σύνορο που τις χωρίζει, επειδή το k τείνει στο 0.

Θα μελετήσουμε με λεπτομέρεια 2 τέτοιες κατανομές και θα δούμε και τον βαθμό k , του ανάλογου συνεχούς των κατανομών αυτών.

4.1 Υποδιπλασιαζόμενο βήμα

Πρόκειται για μία κατανομή τιμών στο $[\alpha, \beta]$ συμμετρικά ως προς το μέσο του $(\alpha+\beta)/2$. Τα άκρα συμπεριλαμβάνονται στις τιμές που παίρνει η τμ X . Στο ημιδιάστημα $\left[\alpha, \frac{\alpha+\beta}{2}\right]$ παρατηρούνται οι ξ τιμές $\alpha,$

$\alpha + \frac{\beta-\alpha}{2^2}, \alpha + \frac{\beta-\alpha}{2^2} + \frac{\beta-\alpha}{2^3}, \dots, \alpha + \sum_{i=2}^j \frac{\beta-\alpha}{2^i}, \dots$ όπου $j=2,3,4,\dots,\xi$, και στο

ημιδιάστημα $\left[\frac{\alpha+\beta}{2}, \beta\right]$ οι συμμετρικές τους. Έχουμε δηλαδή την τμ X με τιμές τις

$$\alpha, \dots, \alpha + \sum_{i=2}^j \frac{\beta-\alpha}{2^i}, \dots, \beta - \sum_{i=2}^j \frac{\beta-\alpha}{2^i}, \dots, \beta, j=2,3,\dots,\xi. \quad (4.1)$$

Προφανές για την τμ X της (4.1) είναι $E(X)=(\alpha+\beta)/2$ και η διασπορά δίνεται από τον τύπο

$$V(X) = \frac{2}{2^\xi - 1} \cdot \sum_{i=1}^{\xi} \left(\frac{\beta - \alpha}{2^i} \right)^2 = \frac{2 \cdot (\beta - \alpha)^2}{2^\xi - 1} \cdot \sum_{i=1}^{\xi} \frac{1}{2^{2i}} = \frac{2 \cdot (\beta - \alpha)^2}{2^\xi - 1} \cdot \frac{1}{2^2} \cdot \frac{1 - 2^{-2\xi}}{1 - 2^{-2}}$$

ήτοι

$$V(X) = \frac{2 \cdot (\beta - \alpha)^2}{2^\xi - 1} \cdot \frac{2^{2\xi} - 1}{3 \cdot 2^{2\xi}} \quad (4.2)$$

και η τυπική απόκλιση και ο ΣΜ αντίστοιχα από

$$S = \sqrt{\frac{2 \cdot (2^{2\xi} - 1)}{3 \cdot (2^\xi - 1)} \cdot \frac{(\beta - \alpha)}{2^\xi}} \quad \text{και} \quad C_{-v} = \sqrt{\frac{2 \cdot (2^{2\xi} - 1)}{3 \cdot (2^\xi - 1)} \cdot \frac{1}{2^{\xi-1}} \cdot \frac{\beta - \alpha}{\beta + \alpha}}. \quad (4.3)$$

Όταν είναι $\alpha=0$, ο ΣΜ δίνεται από

$$C_{-v} = \sqrt{\frac{2 \cdot (2^{2\xi} - 1)}{3 \cdot (2^\xi - 1)} \cdot \frac{1}{2^{\xi-1}}} \quad \text{ή} \quad C_{-v} = \sqrt{\frac{2 \cdot (2^N - 1)}{3 \cdot (N - 1)} \cdot \frac{2}{2^{\frac{N}{2}}}}, \quad \text{με } N=2\xi \quad (4.4)$$

Όλα τα σημεία-τιμές της τμ X είναι σε πλήθος $N=2\xi$ και οι τιμές του C_{-v} για μερικές τιμές του ξ δίνονται στον παρακάτω Πίνακα 4.1, όπου φαίνονται και οι αντίστοιχες τιμές των παραμέτρων λ και $k(\xi)$ και υπονοείται $\alpha=0$:

Πίνακας 4.1

ξ	2	3	4	5	10	20		$\rightarrow \infty$
C_{-v}	0.9129	0.7246	0.6160	0.5441	0.3746	0.2615		$\rightarrow 0$
λ	1.1999	1.9046	2.6354	3.3779	7.1263	14.6237		$\rightarrow \infty$
$k(\xi)$	- 0.8722	- 0.4853	- 0.1504	0.1468	1.3082	2.9312		$\rightarrow \infty$

Από τον Πίνακα 4.1 φαίνεται ότι τα συνεχή ανάλογα των κατανομών που παίρνουμε με τις διάφορες τιμές του $N=2\xi$, $\xi=2,3,4,\dots \rightarrow \infty$ είναι με διάφορες τιμές εκθέτη k , από την τιμή -0.8722 (πολύ κοντά στο απόλυτο όριο -1) μέχρι και το ∞ θεωρητικά. Εύκολα διαχειρίσιμες είναι οι τιμές του k μέχρι 7, Farmakis (2003).

4.2 Βήμα διαιρούμενο δια του πραγματικού αριθμού r

Πρόκειται για μία κατανομή τιμών στο $[\alpha, \beta]$ συμμετρικά ως προς το μέσο του $(\alpha+\beta)/2$. Τα άκρα συμπεριλαμβάνονται στις τιμές που παίρνει η τμ X . Στο ημιδιάστημα $\left[\alpha, \frac{\alpha+\beta}{2}\right]$ παρατηρούνται οι ξ τιμές

$$\alpha, X_2, X_3, \dots, X_\xi \quad \text{με} \quad \frac{\beta+\alpha}{2} - X_j = \frac{\beta-\alpha}{2 \cdot r^{j-1}}, \quad j=2,3,4,\dots,\xi, \quad r>1 \quad (4.5)$$

και στο ημιδιάστημα $\left[\frac{\alpha+\beta}{2}, \beta\right]$ οι συμμετρικές τους.

Έχουμε δηλαδή μέση τιμή την $E(X)=(\alpha+\beta)/2$ (λόγω και της συμμετρίας) και διασπορά την ποσότητα

$$V(X) = \frac{2}{2\xi-1} \cdot \sum_{i=1}^{\xi} \left(\frac{\beta-\alpha}{2 \cdot r^{i-1}} \right)^2 = \frac{(\beta-\alpha)^2}{2 \cdot (2\xi-1)} \cdot \sum_{i=1}^{\xi} \left(\frac{1}{r^2} \right)^{i-1} = \frac{(\beta-\alpha)^2}{2 \cdot (2\xi-1)} \cdot \frac{r^2}{r^{2\xi}} \cdot \frac{r^{2\xi}-1}{r^2-1}$$

οπότε ο ΣΜ είναι η ποσότητα

$$C_{-v} = \frac{r}{r^\xi} \cdot \sqrt{\frac{2 \cdot (r^{2\xi}-1)}{(r^2-1) \cdot (2\xi-1)}} \cdot \frac{\beta-\alpha}{\beta+\alpha} \quad \text{ή} \quad C_{-v} = \frac{r}{r^\xi} \cdot \sqrt{\frac{2 \cdot (r^{2\xi}-1)}{(r^2-1) \cdot (2\xi-1)}}, \quad \text{αν } \alpha=0. \quad (4.6)$$

Αν στην (4.6) τεθεί $r=2$ προκύπτει ο ΣΜ της (4.4). Αν στην (4.6) τεθεί $r=3/2$ τότε είναι

$$C_{-v} = \frac{1}{3^{\xi-1}} \cdot \sqrt{\frac{2 \cdot (3^{2\xi} - 2^{2\xi})}{5 \cdot (2\xi - 1)}} \quad \text{ή} \quad C_{-v} = \frac{1}{3^{\frac{N}{2}-1}} \cdot \sqrt{\frac{2 \cdot (3^N - 2^N)}{5 \cdot (N - 1)}}, \quad N = 2\xi \quad (4.7)$$

Όλα τα σημεία-τιμές της τμ X είναι σε πλήθος $N=2\xi$ και οι τιμές του C_{-v} για μερικές τιμές του ξ δίνονται στον παρακάτω Πίνακα 4.2, όπου φαίνονται και οι αντίστοιχες τιμές των παραμέτρων λ και $k(\xi)$ και υπονοείται $\alpha=0$:

Πίνακας 4.2

ξ	2	3	4	5	10	20		$\rightarrow \infty$
C_{-v}	0.9813	0.8104	0.7030	0.6269	0.4352	0.3038		$\rightarrow 0$
λ	1.0385	1.5227	2.0234	2.5445	5.2799	10.8349		$\rightarrow \infty$
$k(\xi)$	-0.9745	-0.6847	-0.4271	-0.1894	0.7878	2.1819		$\rightarrow \infty$

Από τον Πίνακα 4.2 φαίνεται ότι τα συνεχή ανάλογα των κατανομών που παίρνουμε με τις διάφορες τιμές του $N=2\xi$, $\xi=2,3,4,\dots \rightarrow \infty$ είναι με διάφορες τιμές εκθέτη k , από την τιμή -0.9745 (πολύ κοντά στο απόλυτο όριο -1 και μάλιστα κοντύτερα απ' ότι στον Πίνακα 4.1 με $r=2$) μέχρι και το ∞ θεωρητικά. Εύκολα διαχειρίσιμες είναι οι τιμές του k μέχρι 7, Farmakis (2003).

Σημειώνεται ότι είναι μεν $r>1$ αλλά για τιμές πολύ κοντά στο 1 μερικές Τιμές του ξ δεν δίνουν εφικτή τιμή του $k(\xi)$ διότι βγαίνει $k(\xi)<-1$, πράγμα χωρίς νόημα σύμφωνα με όσα έχουν αποδειχτεί ήδη παλιότερα, Farmakis (2003).

5. ΣΥΖΗΤΗΣΗ ΣΤΙΣ ΔΙΑΚΡΙΤΕΣ ΣΥΜΜΕΤΡΙΚΕΣ ΚΑΤΑΝΟΜΕΣ

Από τα παραπάνω προκύπτει ότι αν οι παρατηρήσεις και γενικότερα τα μέλη μιας κατανομής είναι διακριτές και οι διαδοχικές ισαπέχουν τότε η πολυωνυμική εκτιμήτρια σ.π.π. είναι η δ.ο.κ. και οριακά έχει ανάλογη συνεχή κατανομή την ομοιόμορφη, $k=0$. Κάθε μορφή κατανομής με πεπερασμένο πλήθος σημείων με ίσες αποστάσεις μεταξύ διαδοχικών τέτοιων σημείων έχει ανάλογη συνεχή κατανομή με πολυωνυμική μορφή και αρνητικό εκθέτη. Καθώς πυκνώνουν οι ισαπέχουσες τιμές-παρατηρήσεις ο εκθέτης τείνει στο μηδέν από αριστερά.

Αν τα διαδοχικά στοιχεία μιάς διακριτής κατανομής δεν είναι ισαπέχοντα τότε έχουμε διαφορετική πυκνότητα εμφάνισης των στοιχείων και υπάρχουν πολλές μορφές τέτοιων κατανομών. Αν έχουμε μεγάλη πυκνότητα στο κέντρο τότε ο εκθέτης της ισοδύναμης συνεχούς κατανομής περνάει γρήγορα σε θετικές τιμές μόλις το πλήθος των στοιχείων της κατανομής ξεπεράσει τα 10 ή 20 σημεία.

Αν τα σημεία είναι σαν αυτά της σχέσης (4.1), τότε $r=2>1$ ή σαν αυτά της (4.5) με κάποια τιμή $r>1$, τότε το πλήθος το 2^x μπορεί να τείνει και προς το άπειρο, όπως στον Πίνακα 4.1 και ο ΣΜ τείνει στο 0. Ο εκθέτης $k(x)$ τείνει στο άπειρο επειδή η ενδιάμεση παράμετρος τείνει στο άπειρο επίσης και μάλιστα ταχύτερα του εκθέτη.

Οι συναρτήσεις που έχουν εκθέτη πάνω από το 8 πάντως είναι αρκετά δύσχρηστες. Σε περίπτωση επεξεργασίας δειγματικών δεδομένων τέτοιων κατανομών πρέπει στην ομαδοποίηση των παρατηρήσεων (αν χρειαστεί τέτοια επεξεργασία) να πάρουμε ομάδες με πολύ μικρό εύρος ώστε πολύ στενές περιοχές προς το κέντρο να αναδεικνύονται επειδή έχουν μεγάλη πυκνότητα στοιχείων και ανάλογη σπουδαιότητα. Όσο μεγαλύτερος είναι ο προκύπτων εκθέτης τέτοιων διακριτών κατανομών τόσο περισσότερο αυτές αντιστοιχούν σε συνεχείς κατανομές με πυκνή «παρουσία» σημείων προς το κέντρο της κατανομής και αραιότερη στα δύο άκρα.

ΠΑΡΑΡΤΗΜΑ

Θα παρουσιαστούν τα βασικά συμπεράσματα σχετικά με τις συνεχείς κατανομές τυχαίων μεταβλητών X που εκτείνονται στο διάστημα $[a, \beta]$ και φυσικά είναι της μορφής

$$f(x) = \begin{cases} h \cdot (x-a)^k, & x \in [a, \frac{a+b}{2}] \\ h \cdot (b-x)^k, & x \in [\frac{a+b}{2}, b], \quad k \neq -1, \quad h = \frac{2^k \cdot (k+1)}{(b-a)^{k+1}} \\ 0, & \text{αλλού} \end{cases} \quad (\text{Π1})$$

όπου ο συντελεστής h παίρνει τη μορφή αυτή με βάση τη συνθήκη

$$\int_{-\infty}^{\infty} f(x) dx = 1$$

ενώ το $k > -1$ τελικά.

Με βάση τους τύπους της μέσης τιμής της διασποράς και του συντελεστή μεταβλητότητας και μετά από κάποιες πράξεις καταλήγουμε στο συμπέρασμα:

Θεώρημα: Farmakis (2003). «Ο ΣΜ C_{-v} συνδέεται με τον εκθέτη k με τη σχέση

$$C_{-v} = \sqrt{\frac{2}{(k+3) \cdot (k+2)}} \cdot \frac{\beta - \alpha}{\beta + \alpha} \quad (\text{Π2})$$

ή αν $\alpha=0$

$$C_{-v} = \sqrt{\frac{2}{(k+3) \cdot (k+2)}} \gg \quad (\text{Π3})$$

Αν υιοθετήσουμε για την ευκολία των υπολογισμών την παράμετρο λ ώστε να είναι

$$\sqrt{\frac{2}{(k+3) \cdot (k+2)}} = \sqrt{\frac{1}{\lambda}} \quad \text{ή} \quad \lambda = C_{-v}^{-2}, \text{ με } \lambda > 0 \quad (\Pi 4)$$

τότε έχουμε το συμπέρασμα:

«Ο εκθέτης k συμμετρικής κατανομής (Π1) δίνεται από τη σχέση

$$k = \frac{-5 + \sqrt{1 + 8\lambda}}{2}, \quad \lambda > 0.» \quad (\Pi 5)$$

Αν μείνουμε στις ακέραιες τιμές το λ τότε ουσιαστικά έχουμε μία ακολουθία από τυχαίες μεταβλητές με X_λ , $\lambda=1,2,3,\dots$ και έχουμε μία ταξινόμηση των συμμετρικών συνεχών κατανομών με βάση την παράμετρο λ . Π. χ. με $\lambda=1$ έχουμε την ακραία περίπτωση που μόνο οριακά φανταζόμαστε την ύπαρξή της διότι έχει $k=-1$. Επίσης με $\lambda=3$ έχουμε $k=0$, την ομοιόμορφη κατανομή, με $\lambda=6$ έχουμε $k=1$, την τριγωνική κατανομή, κ. ο. κ.

Όλη η παραπάνω διαδικασία αναδεικνύει τη δειγματοληψία σε παραγωγικότατο εργαλείο προσδιορισμού της σ.π.π. μιας συνεχούς συμμετρικής κατανομής μέσα από ένα δείγμα μετά από κάποιους εύκολους υπολογισμούς και σε ρυθμό αστραπής. Βέβαια η κατανομή δίνεται με προσέγγιση πολυωνυμικής (μόνο) μορφής. Είναι όμως μάλλον η ιδανική για εφαρμογές στα διάφορα φαινόμενα Βιολογίας, Οικονομίας, Ιατρικής, Επιστήμης Περιβάλλοντος, Φυσικής, κλπ.

Αν θέλουμε να επιβεβαιώσουμε τα αποτελέσματά μας μεταβαίνουμε από τη σ.π.π. στην συνάρτηση κατανομής (σ.κ.) παράγοντας ένα ισοπληθές δείγμα με τα θεωρητικά μεγέθη από τη σ.κ. Μία δοκιμασία χ^2 δείχνει αν τα παρατηρούμενα μεγέθη προσαρμόζονται καλά στα θεωρητικά που προέκυψαν από αυτό το δείγμα με την εφαρμογή της παραπάνω μεθόδου. Φυσικά για να προχωρήσουμε σε εφαρμογή της μεθόδου ένα ιστόγραμμα για να εκτιμηθεί αν έχουμε κατανομή που να προσιδιάζει με συμμετρική είναι απαραίτητο στην αρχή.

ABSTRACT

The Coefficient of Variation (symbol C_{-v}) has become a very useful tool in the study of the distributions of random variables especially in the study on the continuous random variables, where from reasonably large sample (size $n > 45$ usually) we can get safe results on the probability density function (pdf) of a random variable X , undertaken that pdf can be approximated by a polynomial function, See also Farmakis (2003), (2010), Farmakis & Makris (2011). A key point is the exponent k (degree) of the polynomial describing the form of pdf. It has been proved that this exponent is connected with C_{-v} via a very simple function when we deal with continuous random variables. In this paper we try to use C_{-v} and k in order to describe a discrete distribution and we also try to connect the discrete distribution with a continuous one, taken as its equivalent continuous distribution. The values of the random variable X

are included in the space $[a, b]$ and symmetric distribution is associated with X . Some illustrative examples are given.

Ευχαριστίες: Ευχαριστούμε τον κριτή για τις εύστοχες παρατηρήσεις και για τις εποικοδομητικές συστάσεις.

ΑΝΑΦΟΡΕΣ

- Breunig R. (2001). “An Almost Unbiased Estimator of the Coefficient of Variation”, *Economic Letters*, Vol. **70**, pp 15-19.
- Farmakis N. (2003). “Estimation of the Coefficient of Variation: Scaling of Symmetric Continuous Distributions”, *Statistics in Transition*, Vol. **6**, No 1, pp 83-96.
- Farmakis N. (2010). “Coefficient of Variation: Connecting Sampling with some Increasing Distribution Models”, *Proceedings of Stochastic Modeling Techniques & Data Analysis International Conference (SMTDA-2010)*, June 8-11, 2010, Chania, Crete, GREECE.
- Farmakis N., Makris G. (2011). “Web-Sampling: Probabilities of Specific Information”, *Proceedings of Applied Stochastic Model & Data Analysis(ASMDA-2011)*, Rome June 2011.

ΠΡΟΣΔΙΟΡΙΣΜΟΣ ΑΡΙΣΤΟΥ ΕΠΙΠΕΔΟΥ ΡΥΠΑΝΣΗΣ ΚΑΙ ΣΥΝΘΗΚΕΣ ΕΞΑΓΩΓΗΣ ΠΕΡΙΟΧΗΣ ΩΦΕΛΕΙΑΣ

Γεώργιος Ε. Χάλκος, Δήμητρα Χ. Κίτσου

Τμήμα Οικονομικών Επιστημών, Πανεπιστήμιο Θεσσαλίας

halkos@uth.gr, dimkitsou@yahoo.gr

ΠΕΡΙΛΗΨΗ

Ο προσδιορισμός του άριστου επιπέδου ρύπανσης είναι μια διαδικασία απαραίτητη στην Οικονομική του Περιβάλλοντος. Αυτή προσδιορίζεται βάσει εκτίμησης δύο σημαντικών συναρτήσεων: του οριακού κόστους ελέγχου της ρύπανσης (Marginal Abatement Cost, MAC) και του οριακού κόστους ζημίας (Marginal Damage cost function, MD). Η αβεβαιότητα στον προσδιορισμό της εξειδίκευσης του μοντέλου, μπορεί να περιοριστεί, διαλέγοντας κατάλληλες συναρτησιακές μορφές των καμπυλών MAC και MD. Στόχος της παρούσας μελέτης είναι ο προσδιορισμός της περιοχής ωφέλειας (Benefit Area, BA) υποθέτοντας διάφορες συναρτησιακές μορφές (προσαρμοστικά μοντέλα). Οι παράμετροι των μοντέλων αυτών εκτιμώνται με τη βοήθεια της μεθόδου ελαχίστων τετραγώνων (OLS) για τη συνάρτηση της καμπύλης MAC και με διακρίβωση (calibration) για τη συνάρτηση MD.

Λέξεις Κλειδιά: Περιοχή ωφέλειας, συνάρτηση ελέγχου, συνάρτηση ζημίας.

1. ΕΙΣΑΓΩΓΗ

Τελευταία η προσοχή της Οικονομικής του Περιβάλλοντος έχει επικεντρωθεί στα διάφορα σημαντικά περιβαλλοντικά προβλήματα όπως μεταξύ άλλων αυτά της όξινης βροχής, της κλιματικής αλλαγής, της τρύπας του όζοντος, της ερημοποίησης, της αποψίλωσης των δασών και τις προκαλούμενες περιβαλλοντικές ζημίες.

Το κύριο πρόβλημα με τα περιβαλλοντικά προβλήματα (και κυρίως αυτά της διασυννοριακής φύσης) είναι η περιορισμένη ικανότητα των χωρών να αντιμετωπίσουν το πρόβλημα μόνες τους. Η αναγκαιότητα συνεργασίας είναι εμφανής. Οι λύσεις συνεργασίας μπορούν να επιτύχουν τους περιβαλλοντικούς στόχους με πιο αποτελεσματικό τρόπο από πλευράς κόστους. Αλλά χρειάζεται ένας μηχανισμός μεταβίβασης πληρωμών για παρακίνηση των χωρών να συνεργαστούν.

Συνεπώς το πρόβλημα είναι να βρεθεί ένα θεσμικό πλαίσιο που θα διευκολύνει στο να γίνονται αφενός οι κατάλληλες συμφωνίες και αφετέρου μια ποσοτική εκτίμηση των επιπτώσεων από τις εκπομπές ρύπων που επιτυγχάνεται με τον προσδιορισμό της περιοχής ωφέλειας (Benefit Area, BA) και την κατάλληλη προσέγγιση των συναρτήσεων κόστους ελέγχου και κόστους ζημίας. Έτσι ένας

τέτοιος ποσοτικός υπολογισμός θα μπορούσε να κατατάξει τις χώρες ανάλογα με την περιοχή ωφέλειας (BA) που δημιουργούν οι εκπεμπόμενοι από αυτούς ρύποι.

Η εκτίμηση των παραμέτρων στα υποτιθέμενα μοντέλα που προσαρμόζουν τις MAC και MD μπορεί να επιτευχθεί με κατάλληλη Στατιστική Ανάλυση όπως η χρήση Ανάλυση Παλινδρόμησης. Μία τέτοια θεσμική συμβολή θα είχε σαν αποτέλεσμα να βελτιωθούν οι περιβαλλοντικές συνθήκες όλων των εμπλεκόμενων χωρών. Διαφορετικά οποιαδήποτε άλλη συμφωνία είναι ανεπιθύμητη ή αναποτελεσματική. Για το λόγο αυτό αναζητάμε ως αποδεκτή λύση την προσέγγιση της αποτελεσματικότητας κατά Pareto.

Το κύριο πρόβλημα στην προώθηση διεθνών συνεργασιών είναι η απουσία μιας πειστικής επιστημονικής έκβασης. Οι οικονομολόγοι του περιβάλλοντος αναγνωρίζουν ότι η αβεβαιότητα η οποία προκύπτει από τη φύση των συναρτήσεων οριακού κόστους ελέγχου (MAC) και του οριακού κόστους ζημίας (MDC) είναι καθοριστικής σημασίας προκειμένου να καταλήξουμε στην κατά Pareto αποτελεσματική πολιτική.

Η αβεβαιότητα σχετικά με τη συνάρτηση MAC μιας πηγής ρύπανσης (π.χ. μιας βιομηχανικής μονάδας) δημιουργείται κυρίως, λόγω της διαθέσιμης μεθόδου ελέγχου της ρύπανσης. Ταυτόχρονα δημιουργείται οικονομική αβεβαιότητα από τη συγκέντρωση του απαιτούμενου κόστους μείωσης της ρύπανσης. Αυτή η θεωρητική σκέψη απεικονίζεται μέσω της καμπύλης αποτελεσματικότητας. Όλες αυτές οι θεωρητικά υπάρχουσες καμπύλες απαιτούν κατάλληλη προσαρμογή των δεδομένων με τη χρήση της μεθόδου ελαχίστων τετραγώνων, όπως θα δούμε στην ενότητα 3.

Τα διάφορα περιβαλλοντικά προβλήματα και οι προκαλούμενες ζημίες δημιουργούν υψηλά κοινωνικά κόστη τα οποία συνήθως δεν λαμβάνονται υπόψη ή πιθανότατα υποεκτιμώνται στις διάφορες περιβαλλοντικές αποτιμήσεις. Αυτή η παράλειψη ή υποεκτίμηση του κοινωνικού κόστους οφείλεται στην αποτυχία της αγοράς λόγω ύπαρξης δημόσιων αγαθών ή των διαφόρων εξωτερικιοτήτων. Η διόρθωση αυτών των αποτυχιών απαιτεί κατάλληλη πολιτική παρέμβαση.

Προτού οι υπεύθυνοι των περιβαλλοντικών αποφάσεων παρέμβουν για να διορθώσουν τις εξωτερικότητες, χρειάζεται μια οργανωμένη περιβαλλοντική πολιτική η οποία να μας οδηγεί στο άριστο επίπεδο ρύπανσης. Αυτό με τη σειρά του απαιτεί τη σύγκριση ανάμεσα στο κόστος ζημίας (η όφελος από τη μείωση της ζημίας) με το κόστος πρόληψης της ζημίας. Οι οικονομικές θεωρίες προτείνουν πως το βέλτιστο επίπεδο ρύπανσης πραγματοποιείται όταν το οριακό κόστος ζημίας ισούται με το οριακό κόστος ελέγχου.

Η οριακή ζημία δείχνει τη ρύπανση ως συνάρτηση της εκπομπής αερίων ενός συγκεκριμένου ρυπαντή. Οι περιβαλλοντικές επιπτώσεις και ζημίες μετρούνται ως οι επιπτώσεις, μεταξύ άλλων, στην ανθρώπινη υγεία, στις λίμνες, στα ποτάμια κ.ά. Γίνονται προσπάθειες για τη μέτρηση των υπαρχουσών τιμών ή άλλων έμμεσων αξιών, χρησιμοποιώντας διάφορες μεθόδους όπως αυτή της υποθετικής αγοράς (Contingent Valuation) ή της έμμεσης ή ωφελμιστικής (hedonic pricing), ή του κόστους ταξιδιού (travel cost) κ.ά. (Halkos and Jones, 2012. Halkos και Matsiori, 2012. Bjornstad και Kahn, 1996. Freeman, 1993). Προφανώς η μέτρηση της ζημίας είναι σημαντική και δύσκολη λόγω μιας σειράς προβλημάτων που παρουσιάζονται

σύμφωνα με τους Halkos (2010), Tol (2002), Farmer et al. (2001), Georgiou et al. (1997) και Barbier (1998).

Με αυτή τη μελέτη θα προσπαθήσουμε να καθορίσουμε το βέλτιστο επίπεδο ρύπανσης κάτω από τις υποθέσεις των γραμμικών, δευτεροβάθμιων και εκθετικών συναρτησιακών μορφών των καμπυλών κόστους ζημίας και κόστους καταπολέμησης των εκπομπών ρύπων.

2. ΠΡΟΣΔΙΟΡΙΣΜΟΣ ΤΟΥ ΑΡΙΣΤΟΥ ΕΠΙΠΕΔΟΥ ΡΥΠΑΝΣΗΣ

Η περιοχή ωφέλειας (Benefit Area) ορίζεται ως το εμβαδόν της τομής των MAC και MD με συντεταγμένες το άριστο επίπεδο που αντιστοιχεί στο βέλτιστο περιορισμό στις ζημίες και το επίπεδο που αντιστοιχεί στο βέλτιστο κόστος (ή όφελος). Οι Halkos και Kitsos (2005) εξέτασαν τρεις περιπτώσεις για τη συνάρτηση ελέγχου $MAC = g(z)$. Συγκεκριμένα:

- Γραμμική [$MAC(z) = \beta_0 + \beta_1 z, \beta_1 \neq 0$];
- Δευτεροβάθμια [$MAC(z) = \beta_0 + \beta_1 z + \beta_2 z^2, \beta_2 > 0$];
- Εκθετική [$MAC(z) = \beta_0 e^{\beta_1 z}, \beta_1 \neq 0$],

με τη συνάρτηση οριακού κόστους ζημίας να είναι μία γραμμική συνάρτηση της μορφής:

$$MD(z) = \varphi(z) = \alpha + \beta z_0$$

Ο προσδιορισμός των παραμέτρων της συνάρτησης του οριακού κόστους ελέγχου επιτυγχάνεται με τη χρήση της μεθόδου ελαχίστων τετραγώνων (OLS)¹ ενώ αυτές της συνάρτησης ζημίας με τη διακρίβωση (calibration).²

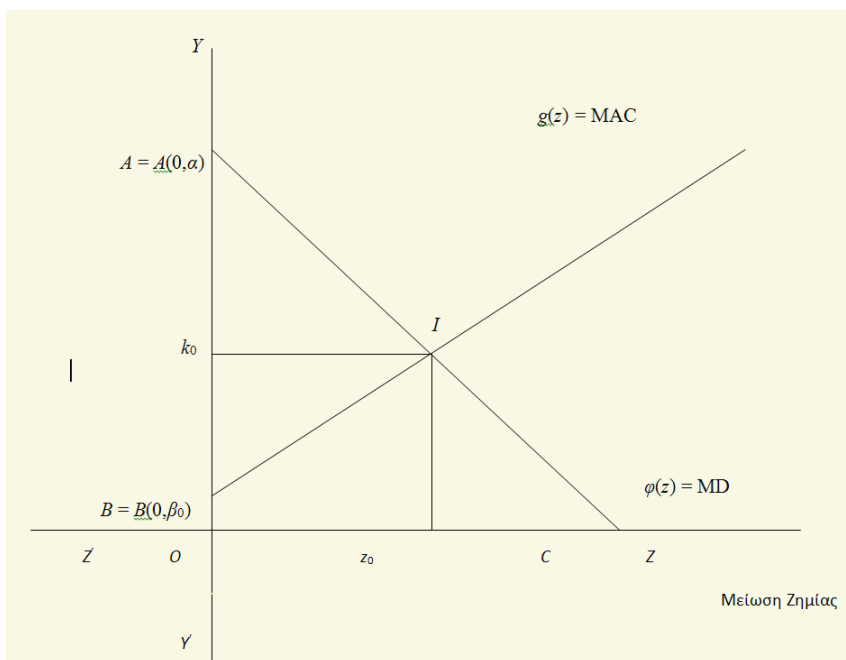
Υπό τις προϋπάρχουσες υποθέσεις υπολογίσθηκε η αντίστοιχη περιοχή ωφέλειας για διαφορετικές πηγές (χώρες, βιομηχανικές μονάδες) υιοθετώντας διαφορετικά μοντέλα και πολιτικές επίλυσης των περιβαλλοντικών προβλημάτων. Στο Διάγραμμα 1 τα σημεία A και B εκφράζουν τα σημεία των MD και MAC που τέμνουν τον άξονα Y περιορίζοντας την ανάλυση σε θετικές τιμές. Για τα σημεία $A=A(0,a)$ και $B=B(0,b)$ οι τιμές a και b είναι οι σταθεροί όροι των καμπυλών MD και MAC. Για την περίπτωση όπου η MAC είναι γραμμική έχουμε $A=A(0,a)$ και $B=B(0,\beta_0)$ (Διάγραμμα 1) και υποθέτουμε ότι $a > \beta_0$.

Το σημείο τομής των καμπυλών MD και MAC είναι το $I = I(z_0, k_0)$. Εξετάστηκε λοιπόν ποιες πρέπει να είναι οι τιμές των σημείων z_0 και k_0 προκειμένου να έχουμε το άριστο επίπεδο στις ζημίες με το αντίστοιχο άριστο κόστος ελέγχου. Στο σημείο I ικανοποιείται η σχέση $MAC(z_0) = g(z_0) = \varphi(z_0) = MD(z_0)$, με z_0 να αντιστοιχεί στο βέλτιστο περιορισμό των ζημιών. Όπως έχουμε τονίσει ήδη, οι συντελεστές της συνάρτησης ελέγχου των εκπομπών ρύπων ($\beta_0, \beta_1, \beta_2$) εκτιμήθηκαν με τη μέθοδο OLS.

¹ Ως εκτιμήσεις οι συντελεστές θα συμβολίζονται με τους αντίστοιχους λατινικούς χαρακτήρες π.χ. b αντί β .

² Για λεπτομέρειες δείτε Hutton and Halkos (1995).

Σχήμα 1 Γραφική παράσταση του άριστου επιπέδου ρύπανσης (γενική μορφή)



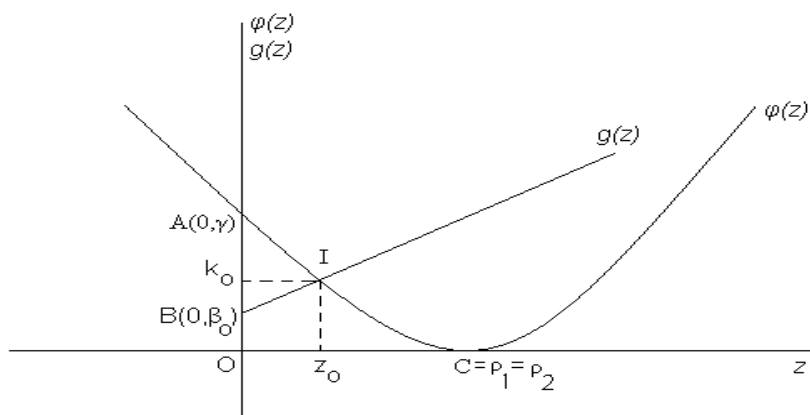
Οι Halkos και Kitsou (2014) εξέτασαν αναλυτικά όλους τους συνδυασμούς συναρτησιακών μορφών των καμπυλών MAC και MD. Για παράδειγμα, στην περίπτωση όπου η MD είναι δευτεροβάθμια συνάρτηση και η MAC γραμμική έχουμε:

$$MAC = g(z) = \beta_0 + \beta_1 z, \quad \beta_1 \neq 0 \quad \text{και} \quad MD = \varphi(z) = \alpha z^2 + \beta z + \gamma, \quad \alpha > 0$$

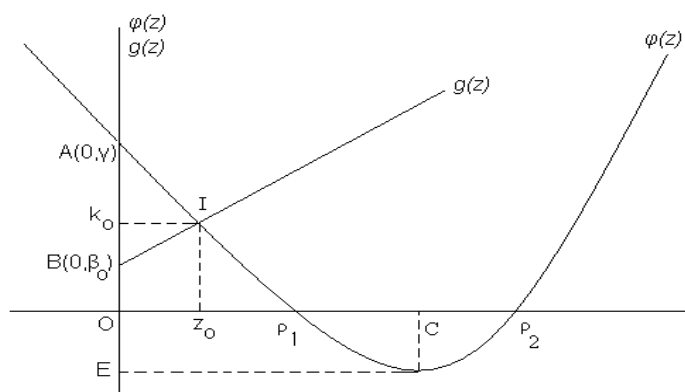
Σε αυτή την περίπτωση οι τιμές των α και β , της τομής των MD και MAC με τον άξονα Y είναι $b = MAC(0) = \beta_0$ και $a = MD(0) = \gamma$, όπως το κάτωθι σχήμα. Προκειμένου να ισχύει η τομή των δύο καμπυλών πρέπει να έχουμε τον περιορισμό $0 < \beta_0 < \gamma$. Θεωρώντας ότι $\alpha > 0$ αναλύθηκαν τρεις περιπτώσεις με βάση την διακρίνουσα της $\varphi(z)$, τη D , η οποία είναι $D = \beta^2 - 4\alpha\gamma$. Συγκεκριμένα αναλύθηκαν οι περιπτώσεις όπου η διακρίνουσα $D = 0$ (Σχήμα 2), και $D > 0$ (Σχήμα 3). Στην περίπτωση κατά την οποία $D < 0$ δεν υπάρχει οικονομικό ενδιαφέρον και άρα δεν υπήρξε περεταίρω ανάλυση. Επιπλέον αναπτύχθηκε και η περίπτωση όπου το $\alpha < 0$ (Σχήμα 4).

Σε όλες τις περιπτώσεις ανάλογα με τη διακρίνουσα και τις ρίζες της ορίστηκαν και υπολογίστηκαν τα αντίστοιχα άριστα σημεία και η περιοχή ωφέλειας σύμφωνα με την κοιλότητα ή την κυρτότητα της καμπύλης οριακής ζημίας.

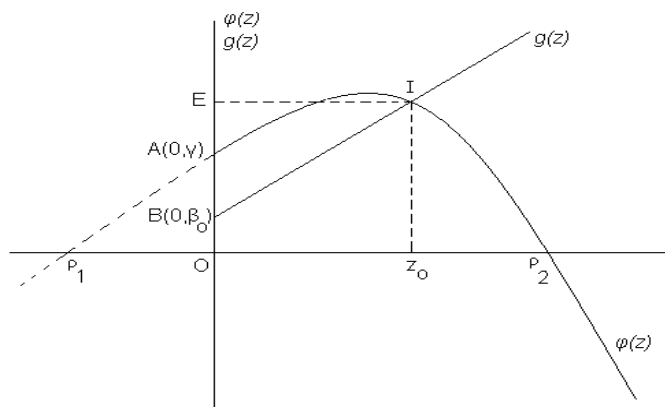
Σχήμα 2. Η $g(z)$ γραμμική και $\varphi(z)$ δευτεροβάθμια με πραγματική διπλή ρίζα.



Σχήμα 3. Η $g(z)$ γραμμική και $\varphi(z)$ δευτεροβάθμια με δύο πραγματικές ρίζες άνισες.



Σχήμα 4. Η $g(z)$ γραμμική και $\varphi(z)$ δευτεροβάθμια με δύο πραγματικές ρίζες άνισες και μέγιστο.



3. ΕΜΠΕΙΡΙΚΗ ΕΦΑΡΜΟΓΗ

Στη συνέχεια θα αναλύσουμε πώς μπορούν να προσεγγιστούν οι καμπύλες ελέγχου και ζημίας. Θα χρησιμοποιήσουμε εκτιμήσεις δεδομένων από εννέα διαφορετικές Ευρωπαϊκές χώρες επιλεγμένες να αντιπροσωπεύουν τα διάφορα γεωγραφικά σημεία της Ευρώπης προκειμένου να χρησιμοποιηθούν στην ανάλυση.³ Εξετάζεται αρχικά ο προσδιορισμός της συνάρτησης του κόστους ελέγχου μέσω της προσέγγισης της μείωσης των τόνων εκπομπών αέριων ρύπων όπως για παράδειγμα εκπομπών θείου (S).

Για τον έλεγχο των εκπομπών του διοξειδίου του θείου στις στατικές πηγές μπορούμε να έχουμε, Halkos (1992, 1993, 1995):

- κατεργασία πριν την καύση (φυσικός καθαρισμός του άνθρακα, αποθείωση πετρελαίου και αντικατάσταση καυσίμου)
- απομάκρυνση του θείου κατά τη διάρκεια της καύσης (μέθοδοι αποθείωσης που εφαρμόζονται σε σταθμούς που χρησιμοποιούν κονιοποιημένο καύσιμο με απευθείας ψεκασμό ασβεστόλιθου και όσες χρησιμοποιούν νέες τεχνολογίες καύσεως όπως η καύση υπό πίεση ρευστοποιημένης στιβάδας ή ακόμη και με αεροποίηση του άνθρακα).
- απομάκρυνση του θείου μετά την καύση του καυσίμου, η οποία μπορεί να γίνει με τη μέθοδο της αποθείωσης αερίου καυσίμου (Flue Gas Desulphurization, FGD).

Το πραγματικό κόστος κάθε διαθέσιμης μεθόδου ελέγχου είναι συνάρτηση εθνικών συνθηκών και οι καμπύλες κόστους εξαρτώνται από το ενεργειακό σενάριο που υιοθετείται. Τα κόστη μείωσης διαφέρουν μεταξύ κρατών ακόμη και για την ίδια τεχνολογία κυρίως λόγω (Halkos 1993, 1995):

- ειδικών συντελεστών, όπως περιεκτικότητα του χρησιμοποιούμενου καυσίμου σε θείο,
- μέγεθος εγκατάστασης,
- και συντελεστές κόστους εργασίας, ηλεκτρισμού και κατασκευής.

Αναλυτικά και σχετικά με τη συνάρτηση του κόστους ελέγχου υποθέτουμε ότι N χώρες παράγουν ηλεκτρισμό (αγαθό, Y_i) με χρήση ορυκτών καυσίμων και εκπομπές θείου (“κακό παραπροϊόν”, E_i) από καύση ορυκτών καυσίμων. Οι εκπομπές είναι συνάρτηση του ηλεκτρισμού που παράγεται, δηλαδή

$$E_i = E_i(Y_i), \quad i = 1, 2, \dots, N.$$

Η διασυννοριακή φύση του προβλήματος αντιπροσωπεύεται (Halkos 1996, 1997) ως:

$$D_i = B_i + d_{ii}(1 - \alpha_i)E_i + \sum_{i \neq j} d_{ij}(1 - \alpha_j)E_j = B_i + \sum_j d_{ij}(1 - \alpha_j)E_j,$$

όπου

- E_i συνολικές εκπομπές θείου (S) στη χώρα i σε τόνους (t)

³ Η εξαγωγή των καμπυλών κόστους ελέγχου των εκπομπών θείου έχει γίνει για όλες τις Ευρωπαϊκές χώρες για το έτος 2000 και μετά από εξέταση κάθε στατικής πηγής (σταθμοί παραγωγής ηλεκτρικής ενέργειας, διυλιστήρια, βιομηχανικές μονάδες κ.ά.). Για περισσότερες πληροφορίες δείτε Halkos (1992, 1993, 1995).

- D_i συνολικές αποθέσεις θείου στη χώρα i σε τόνους (t)
- a_i συντελεστής ελέγχου στη χώρα i
- d_{ij} αναλογία εκπομπών χώρας j που αποθέτονται ("εξάγονται") στη χώρα i (συντελεστής μεταβίβασης, $0 \leq d_{ji} \leq 1$).
- B_i επίπεδο υπαρχουσών αποθέσεων εξαιτίας φυσικών πηγών στην χώρα j , (**background pollution**). Γεγονός ότι ρύπανση μπορεί να παραμένει στην ατμόσφαιρα για μεγάλα χρονικά διαστήματα λαμβάνεται υπόψη με συντελεστή B_i .

Οι αποθέσεις θείου προκαλούν εξωτερικότητες με τη μορφή φυσικών ζημιών, που είναι συνάρτηση του επιπέδου των αποθέσεων. Η συνάρτηση ζημίας είναι αυστηρά κυρτή στις αποθέσεις D_i , δηλαδή

$$Q_i'(D_i) > 0 \text{ και } Q_i''(D_i) > 0.$$

Επεκτείνοντας τα μαθηματικά μοντέλα που περιγράφηκαν πιο πάνω σε στοχαστικά, βρίσκουμε ότι η εξίσωση των ελαχίστων τετραγώνων των κάτωθι τύπων οδηγεί σε ικανοποιητικές προσεγγίσεις για όλες τις χώρες που αναλύονται σε αυτή τη μελέτη.

$$g(z) = AC_i = \beta_{0i} + \beta_{1i}SR_i + u_i \text{ ή } g(z) = AC_i = \beta_{0i} + \beta_{1i}SR + \beta_{2i}SR_i^2 + u_i,$$

SR_i αντιπροσωπεύει το επίπεδο ελέγχου των εκπομπών θείου στη χώρα i (Sulphur Removal, SR_i), AC_i είναι το κόστος ελέγχου της χώρας i και u_i ο διαταρακτικός όρος (σφάλμα).⁴

Επίσης υπολογίζεται η αποτελεσματικότητα (efficiency) ως:

$$Eff = \frac{BA}{\max BA} \times 100.$$

Αυτό είναι ένα μέτρο του ποσοστού το οποίο η υιοθετημένη πολιτική καλύπτει αυτή την πολιτική που θα απέφερε τη μέγιστη περιοχή οφέλους.

Από τον Πίνακα 1 βλέπουμε ότι οι μεγάλες βιομηχανικές προσήνεμες (upwind) χώρες (Γαλλία και Ηνωμένο Βασίλειο) φαίνεται να έχουν μεγάλες περιοχές οφέλους. Κοιτώντας στους συντελεστές μεταβίβασης του European Monitoring and Evaluating Program (EMEP) μπορούμε να διαπιστώσουμε ότι οι χώρες με μεγάλα οφέλη είναι αυτές με μεγάλους αριθμούς στη διαγώνιο δείχνοντας τη σημαντικότητα των εγχώριων πηγών ρύπανσης. Η Πολωνία στην Κεντρική Ευρώπη εμφανίζει τα μεγαλύτερα οφέλη. Οι μεγάλοι εκτός κυρίας διαγώνιου συντελεστές μεταβίβασης δείχνουν τις εξωτερικότητες. Από την άλλη πλευρά οι υπήνεμες (downwind) κοντά στη θάλασσα χώρες ή με ανθεκτικότητα στην οξύτητα εδάφη φαίνεται να έχουν μικρές περιοχές οφέλους.

⁴ Για λεπτομέρειες δείτε Χάλκος (2011).

Πίνακας 1. Εμπειρικά αποτελέσματα εκτίμησης περιοχής ωφέλειας (BA) βάσει εκτιμήσεων OLS για συντελεστές δευτεροβάθμιας συνάρτησης κόστους ελέγχου και εκτιμήσεις διακρίβωσης για γραμμική συνάρτηση ζημίας

Χώρες	a	b	b_0	b_1	b_2	R^2	Eff	z_0	BA
Ισπανία	71.69	0.0072	69.06 (8.41)	0.0039 (1.97)	0.00014 (23.66)	99.2	10.6	149.35	273.74
Γαλλία	33.15	0.2773	21.45 (48.18)	0.1644 (38.64)	0.00134 (118.35)	98.3	43.3	144.62	1115.9
Ελλάδα	3.73	0.0341	2.29 (2.66)	0.0265 (17.27)	0.00099 (64.61)	99.7	1.6	42.12	42.6
Ιταλία	11.01	0.0300	7.78 (17.31)	0.0226 (189.7)	0.00021 (28.06)	99.1	12.9	142.88	332.8
Πολωνία	16.21	0.0231	2.49 (1.38)	0.0178 (42.28)	0.00019 (10.26)	99.5	100.0	277.08	2574.5
Πορτογαλία	1.03	0.0317	0.76 (2.18)	0.0667 (27.33)	0.00689 (15.26)	89.8			
Ρουμανία	9.09	0.0113	5.78 (2.12)	0.00998 (13.60)	0.000153 (10.26)	96.3	13.1	151.46	339.05
Σουηδία	6.4	0.0642	9.61 (7.23)	0.7617 (60.43)	0.01620 (18.52)	97.7			
Ηνωμένο Βασίλειο	19.06	0.0687	9.59 (3.71)	0.0442 (6.99)	0.000212 (11.53)	98.1	80.0	276.80	2060.4

(Λόγοι-t στις παρενθέσεις)

4. ΣΥΜΠΕΡΑΣΜΑΤΑ

Η ανάλυση της αποτελεσματικότητας των περιβαλλοντικών προγραμμάτων απαιτεί σύγκριση του κόστους ζημίας και του κόστους ελέγχου των εκπομπών ρύπων. Η τυπική προσέγγιση του ορισμού του άριστου επιπέδου ρύπανσης είναι η εξίσωση του οριακού κόστους ζημίας με το οριακό κόστος ελέγχου των ρύπων MD=MAC.

Για διαφορετικές υποθέσεις για τη μορφή των καμπυλών MAC και MD και την εξαγωγή της περιοχής οφέλους βρήκαμε ότι το άριστο επίπεδο ρύπανσης εκτιμάται υπό συγκεκριμένες συνθήκες

- i. Σε όλες τις περιπτώσεις $a > \beta_0$. Δηλαδή, ο σταθερός της συνάρτησης ζημίας (μπορούμε να σκεφτούμε τις υπάρχουσες αποθέσεις) είναι μεγαλύτερος από τον αντίστοιχο της συνάρτησης καταπολέμησης στο επίπεδο $z=0$ (μπορούμε να σκεφτούμε τα πάγια κόστη λειτουργίας και συντήρησης μιας μονάδας ελέγχου στο $z = 0$).

- ii. Για περιπτώσεις γραμμικών ή δευτεροβάθμιων και των δύο συναρτήσεων έχουμε $\beta > \beta_1$. Δηλαδή η κλίση της συνάρτησης ζημίας πρέπει να είναι μεγαλύτερη από αυτή της συνάρτησης ελέγχου στο $z = 0$.
- iii. Για τη δευτεροβάθμια περίπτωση απαιτείται ότι $\beta_2 > 0$ ενώ για την εκθετική $\beta_0, \beta_1 > 0$.

Οι δευτεροβάθμιες και εκθετικές συναρτήσεις υπακούουν τον ίδιο τύπο εκτίμησης της περιοχής ωφέλειας. Οι υπολογισμοί αυτοί μπορούν να χρησιμοποιηθούν σαν δείκτες μεταξύ διαφορετικών ανταγωνιστικών πολιτικών, και η πολιτική που προκαλεί τη μέγιστη περιοχή θα είναι η καταλληλότερη (δείτε επίσης Halkos and Kitsou, 2014).

Σημαντική διαπίστωση είναι ότι οι μεγάλες βιομηχανικές προσήνεμες χώρες φαίνεται να έχουν μεγάλα οφέλη ενώ οι υπήνεμες κοντά στη θάλασσα χώρες ή με ανθεκτικότητα στην οξύτητα εδάφη φαίνεται να έχουν μικρές περιοχές οφέλους.

ABSTRACT

The evaluation of the Benefit Area (BA) is essential in Environmental Economics. This paper identifies optimal pollution level and investigates analytically certain restrictions that the existence of this optimal level requires. Uncertainty of the evaluation of the model can be reduced by choosing appropriate approximations for the marginal abatement (MAC) and the marginal damage (MD) cost functions. Parameters estimation requires the use of appropriate methods like Ordinary Least Square (OLS) in the case of abatement cost estimates and calibration of the values in the case of damage cost functions.

ΕΥΧΑΡΙΣΤΙΕΣ

Ευχαριστούμε έναν ανώνυμο κριτή για τα χρήσιμα σχόλια στην αρχική μορφή της εργασίας μας, καθώς και τους συμμετέχοντες στο 26^ο Πανελλήνιο Συνέδριο Στατιστικής για τα ενθαρρυντικά τους σχόλια στην παρουσίαση της μελέτης.

Η παρούσα έρευνα έχει συγχρηματοδοτηθεί από την Ευρωπαϊκή Ένωση (Ευρωπαϊκό Κοινωνικό Ταμείο - ΕΚΤ) και από εθνικούς πόρους μέσω του Επιχειρησιακού Προγράμματος «Εκπαίδευση και Δια Βίου Μάθηση» του Εθνικού Στρατηγικού Πλαισίου Αναφοράς (ΕΣΠΑ) – Ερευνητικό Χρηματοδοτούμενο Έργο: Ηράκλειτος II. Επένδυση στην κοινωνία της γνώσης μέσω του Ευρωπαϊκού Κοινωνικού Ταμείου.



European Union
European Social Fund



MINISTRY OF EDUCATION & RELIGIOUS AFFAIRS
MANAGING AUTHORITY



EUROPEAN SOCIAL FUND

Co- financed by Greece and the European Union

Βιβλιογραφία

- Barbier E. (1998). *The Economics of Environment and Development: Selected Essays*, London: Edward Elgar.
- Bjornstad D. and Kahn J. (1996). *The Contingent Valuation of Environmental Resource: Methodological Issues and Research Needs*, London: Edward Elgar.
- Freeman A. III (1993). *The measurement of Environmental and Resource Values: Theory and Methods*, Washington DC: Resources for the Future.
- Farmer A., Kahn J. R., McDonald J. A. and O'Neill R. (2001). Rethinking the optimal level of environmental quality: Justifications for strict environmental policy. *Ecological Economics*, **36**, 461-473.
- Georgiou S., Pearce D., Whittington D. and Moran D. (1997). *Economic values and the environment in the developing world*, London: Edward Elgar.
- Halkos G. (1992). *Economic Perspectives of the Acid Rain Problem in Europe*, PhD Thesis, Department of Economics and Related Studies, University of York.
- Halkos G. (1993). An evaluation of the direct costs of abatement under the main desulphurisation technologies. MPRA Paper 32588, University Library of Munich, Germany.
- Halkos G. (1995). An evaluation of the direct cost of abatement under the main desulfurization technologies. *Energy Sources*, **17**(4), 391-412.
- Halkos G. (1996). Incomplete information in the acid rain game, *Empirica Journal of Applied Economics and Economic Policy*, **23**(2), 129-148.
- Halkos G. (1997). Modeling optimal nitrogen oxides abatement in Europe, MPRA Paper 33132, University Library of Munich, Germany.
- Halkos G. and Kitsos C. P. (2005). Optimal pollution level: A theoretical identification. *Applied Economics*, **37**, 1475-1483.
- Halkos G. (2010). *Construction of abatement cost curves: The case of F-gases*, MPRA Paper 26532, University Library of Munich, Germany.
- Halkos G. and Jones N. (2012). Modeling the effect of social factors on improving biodiversity protection. *Ecological Economics*, **78**(C), 90-99.
- Halkos G. and Matsiori S. (2012). Determinants of willingness to pay for coastal zone quality improvement. *The Journal of Socio-Economics*, **41**(4), 391-399.
- Halkos G. and Kitsou D. (2014). Uncertainty in optimal pollution levels: Modeling the benefit area. Υπό δημοσίευση στο επιστημονικό περιοδικό *Journal of Environmental Planning and Management* DOI:10. 1080/09640568.2014.881333
- Hutton J. P. and Halkos G. (1995). Optimal acid rain abatement policy in Europe: An analysis for the year 2000. *Energy Economics*, **17**(4), 259-275.
- Kitsou, D. Estimating damage and abatement cost functions to define appropriate environmental policies, PhD Thesis (to appear), Department of Economics, University of Thessaly.
- Tol, R.S.J. (2002). Estimates of the damage costs of climate change: Part I: Benchmark estimates. *Environmental and Resource Economics*, **21**(1), 47-73.
- Χάλκος Γ. (2011). Στατιστική: Θεωρία, εφαρμογές και χρήση στατιστικών προγραμμάτων σε Η/Υ. 3^η Έκδοση, Εκδόσεις Τυπωθήτω, Αθήνα.

εργασίες

στα αγγλικά



ON BAYESIAN TRANSFORMATION SELECTION: PROBLEM FORMULATION AND PRELIMINARY RESULTS

E. Charitidou¹, D. Fouskakis¹, I. Ntzoufras²

¹National Technical University of Athens

echarit@central.ntua.gr, fouskakis@math.ntua.gr

²Athens University of Economics and Business

ntzoufras@aueb.gr

ABSTRACT

The problem of transformation selection is thoroughly treated from a Bayesian perspective. Several families of transformations are considered with a view to achieving normality: the *Box-Cox*, the *Modulus*, the *Yeo & Johnson* and the *Dual* transformation families. Markov Chain Monte Carlo algorithms have been constructed in order to sample from the posterior distribution of the transformation parameter λ_T associated with each competing family T . We investigate different approaches for constructing compatible prior distributions for λ_T over alternative transformation families, using the power-prior and the unit-information prior approaches. In order to distinguish between different transformation families, posterior model probabilities have been estimated. Using simulated datasets, we show the efficiency of our approach.

Keywords: Bayesian transformation selection; MCMC; Power-prior; Prior compatibility.

1. INTRODUCTION

The assumption of normality is fundamental in most parametric statistical procedures with special reference to the - routinely employed - linear model and the associated error terms. Furthermore, normality is essentially related to the notion of conjugacy in Bayesian analysis, which reduces computational complexity. In this paper, several families of transformations are compared with a view of achieving normality of a transformed dataset.

In the literature, the term *transformation selection* so far pertains to the choice of an optimal value for the transformation parameter within a given family. We introduce a two-step approach where a transformation family is selected at an initial level while at a second level the value of the transformation parameter is specified given the selected family. Working within the Bayesian context requires careful choice of prior distributions. In our case, this becomes even more complex since the prior distributions for the

transformation parameter, λ_T , under each family, T , need to be compatible to account for the different interpretation of λ_T given T .

Four uniparametric families of transformations are considered and compared with each other: *Box-Cox* (Box and Cox, 1964), *Modulus* (John and Draper, 1980), *Yeo & Johnson* (Yeo and Jihnsn, 2000) and *Dual* (Yang, 2006). Each family is indexed by T and involves a transformation parameter λ_T . Let us denote by $\mathbf{y} = (y_1, \dots, y_n)^T$ the observed data and by $\mathbf{y}^{(\lambda_T)} = (y_1^{(\lambda_T)}, \dots, y_n^{(\lambda_T)})^T$ the transformed data for a given value of the parameter λ_T within a particular transformation family T . We aim for $\mathbf{y}^{(\lambda_T)}$ to be a sample from a normal distribution $N(\mu_T, \sigma_T^2)$ with unknown parameter vector (μ_T, σ_T^2) under some appropriate value of λ_T .

2. BAYESIAN FORMULATION

Under any transformation T with specific parameter λ_T , the marginal likelihood of the original data is

$$f(\mathbf{y}|\lambda_T, T) = f(\mathbf{y}^{(\lambda_T)}|T) \times \prod_{i=1}^n \left| \frac{\partial y_i^{(\lambda_T)}}{\partial y_i} \right|. \quad (1)$$

$$\propto (S_T^2)^{-\frac{n-1}{2}} \times \prod_{i=1}^n \left| \frac{\partial y_i^{(\lambda_T)}}{\partial y_i} \right| \quad (2)$$

where S_T^2 is the sample variance of the transformed data. Regarding the prior probability of each of the six transformation families, including the Identity (Id) and the Logarithmic (Log) transformation, no special prior weight is assigned to any family, i.e. $\pi(T) = \frac{1}{|T|} = \frac{1}{6}$. As to the prior on the transformation parameters, it has a hierarchical form: $\pi(\theta_T|T) = \pi(\mu_T, \sigma_T^2|\lambda_T, T)\pi(\lambda_T|T)$. The main parameter of interest within a family is λ_T , while (μ_T, σ_T^2) are regarded as nuisance parameters; therefore, we employ an independent Jeffreys prior (reference prior) for those.

On the grounds of the different interpretation of λ_T among families, the concept of the power-prior (Ibrahim and Chen, 2000) is adopted in order to construct compatible prior distributions. In order to fully specify the power-prior, we start with a baseline non-informative prior for λ_T , namely $\pi^N(\lambda_T|T) \propto 1$. The power-prior of λ_T is then proportional to the product of the power-likelihood times the baseline prior. Under the specific choice of $\pi^N(\lambda_T|T)$ this results to:

$$\pi_A(\lambda_T|\mathbf{y}^*, T) = \frac{f(\mathbf{y}^*|\lambda_T, T)^{1/n^*}}{\int f(\mathbf{y}^*|\lambda_T, T)^{1/n^*} d\lambda_T}. \quad (3)$$

In addition, a unit-information normal prior setting is considered, based on the same imaginary data $\mathbf{y}^*$, which theoretically approximates the former power-prior setting. The variance of the latter prior is determined through the observed Fisher information given $\mathbf{y}^*$. As to $\mathbf{y}^*$, it ideally represents available historical data or expert data. In case we have

neither of those, we may simulate some imaginary data from a suitable distribution; for instance we may choose a normal distribution $N(\mu_s, \sigma_s^2)$ so as to reflect the null hypothesis of normality for the observed data. For appropriate selection of the hyperparameters (μ_s, σ_s^2) one could drag information from the actual data $\mathbf{y}$ (minimally empirical values) or use values chosen by the researcher based on substantive considerations.

In the case of the power-prior approach (Prior A), the marginal likelihood is given via the following formula which involves two nested integrals:

$$f(\mathbf{y}|T) = \int \frac{f(\mathbf{y}|\lambda_T, T) f(\mathbf{y}^*|\lambda_T, T)^{1/n^*}}{\int f(\mathbf{y}^*|\lambda_T, T)^{1/n^*} d\lambda_T} d\lambda_T. \quad (4)$$

Approximations of the intractable integrals included in the process are achieved through Chib's estimator (Chib and Jeliazkov, 2001), which incorporates the output of a random-walk Metropolis-Hastings algorithm that simulates from the marginal posterior distribution of λ_T .

3. RESULTS

In order to illustrate our approach, we use simulated data of size 100 and 1000 from a Gamma and a Student distribution. We compare the performance of the various transformation families under each prior via posterior model probabilities. Moreover, we graphically explore the adaptation of the competing models when altering the values of the distribution parameters.

First, we have simulated values from a Gamma(2, 3) distribution to see the behaviour of our approach on highly skewed data (see Table 1). The best adapting class for this dataset is the Box-Cox family for either sample size while the Identity transformation reasonably shows the worst fit. The posterior model probabilities given the Box-Cox family are 60% for the medium size dataset under Prior A and go over 99% for the large size dataset. Quite interestingly, the corresponding value of the posterior median of λ_T is around 0.3 for both sample sizes; only the posterior standard deviation of the transformation parameter is three times less when $n = 1000$ compared to $n = 100$. For the medium size dataset, the Log transformation is the second best after the Box-Cox with posterior model probability 34%; however, for the larger sample size, all models except Box-Cox have posterior model probabilities less than 0.01.

The Student distribution is an example of particular interest since symmetry is accompanied by fat tails. The latter characteristic usually induces failure of transformation to normality under most families. Our example uses a Student distribution t_2 with 2 degrees of freedom and non-centrality parameter equal to -1 . Looking at the figures of Table 2, we observe that the supremacy of the Modulus family is unquestionable for this distribution under both prior setups. For medium and large sample sizes ($n = 100$ & $n = 1000$), the posterior probabilities of the Modulus transformation model rise from 99% to over 99% respectively. The median simulated value of λ_T is about 0.3 in the former case and almost 0 in the latter, meaning that it approximates a form similar, but

Table 1: Posterior model probabilities and log-marginal values of $\mathbf{y}$ for each transformation family T along with Monte Carlo estimates for the posterior median (sd) of λ_T , all estimated using Chib’s approximation method for the Gamma simulated dataset.

G(2,3)		Prior ^a	Box-Cox	Log	Dual	Modulus	YJ	Id
$n = 100$	$P(T \mathbf{y})$	Prior A	0.60	0.34	0.04	0.01	0.02	< 0.01
		Prior B	0.78	0.20	0.01	0.01	< 0.01	< 0.01
	$\log f(\mathbf{y} T)$	Prior A	-104.93	-105.47	-107.60	-108.98	-108.54	-114.26
		Prior B	-104.14	-105.49	-109.24	-108.86	-110.57	-114.26
	λ_T	Prior A	0.34 (0.12)	-	0.01 (0.06)	-0.49 (0.41)-0.45 (0.39)	-	-
		Prior B	0.35 (0.12)	-	0.24 (0.09)	-0.31 (0.37) 0.17 (0.29)	-	-
G(2,3)		Prior	Box-Cox	YJ	Modulus	Log	Dual	Id
$n = 1000$	$P(T \mathbf{y})$	Prior A	> 0.99	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01
		Prior B	> 0.99	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01
	$\log f(\mathbf{y} T)$	Prior A	-2153.59	-2181.84	-2182.01	-2195.63	-2197.82	-2301.69
		Prior B	-2153.49	-2181.73	-2181.96	-2195.63	-2198.04	-2301.69
	λ_T	Prior A	0.33 (0.04)-0.80 (0.12)	-0.81 (0.12)	-	-	0.00 (0.02)	-
		Prior B	0.33 (0.04)-0.79 (0.12)	-0.80 (0.12)	-	-	0.02 (0.02)	-

^a Prior A: Unit-information normal prior; Prior B: Power-prior.

essentially different, to the Logarithmic transformation. It is worth mentioning that according to our findings the Modulus family outstrips the rest of the potential models for simulated values from the Laplace distribution which is another example of a fat tailed density.

To get a more general idea as to the behaviour of the best families regarding the Gamma distribution, we applied the proposed methodology under Prior A for various combinations of the shape and scale distribution parameters (a, b) and for constant sample size $n = 1000$. It is well known that the skewness of the Gamma distribution increases as the shape parameter decreases while an increase of the scale parameter renders the distribution flatter. Figure 1 elegantly pictures the posterior model probabilities $P(T|\mathbf{y})$ of the four best transformations for the Gamma distribution as a function of skewness. The lower part of the graph illustrates the posterior median of λ_T for each parametric family versus skewness. The associated combinations of the distribution parameters (a, b) are also given in the horizontal axis below the skewness values. Note that in the first combination of (a, b) values the shape parameter is taken to be unity, thus degenerating the gamma distribution to an exponential distribution with associated parameter equal to b . For large values of skewness, we observe that the Box-Cox model outperforms the rest. Nonetheless, when the skewness considerably approaches zero, the

Table 2: Posterior model probabilities and log-marginal likelihood values for each trasformation family T along with Monte Carlo estimates for the posterior median (sd) of λ_T , all estimated using Chib's approximation method for the Student simulated dataset.

St ₂		Prior ^a	Modulus	Box-Cox	Dual	Id	YJ	Log
$n = 100$	$P(T y)$	Prior A	0.99	0.01	< 0.01	< 0.01	< 0.01	< 0.01
		Prior B	0.99	0.01	< 0.01	< 0.01	< 0.01	< 0.01
	$\log f(y T)$	Prior A	-250.59	-255.89	-256.39	-258.28	-259.36	-292.14
		Prior B	-251.92	-257.01	-259.77	-258.28	-260.48	-292.14
	λ_T	Prior A	0.36 (0.15)	1.60 (0.22)	1.62 (0.22)	-	1.08 (0.08)	-
		Prior B	0.34 (0.15)	1.62 (0.23)	1.63 (0.21)	-	1.09 (0.08)	-
St ₂		Prior	Modulus	Box-Cox	Dual	YJ	Id	Log
$n = 1000$	$P(T y)$	Prior A	> 0.99	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01
		Prior B	> 0.99	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01
	$\log f(y T)$	Prior A	-3733.00	-3960.67	-3961.52	-4027.56	-4125.02	-4601.98
		Prior B	-3732.75	-3960.92	-3963.83	-4027.63	-4125.02	-4601.98
	λ_T	Prior A	-0.01 (0.04)	2.93 (0.12)	2.93 (0.12)	1.23 (0.01)	-	-
		Prior B	-0.01 (0.04)	2.94 (0.12)	2.94 (0.12)	1.23 (0.01)	-	-

^a Prior A: Unit-information normal prior; Prior B: Power-prior.

posterior model probabilities of the Box-Cox model tend to get close to the rest of the models.

We repeated a similar process as an attempt to relate the degrees of freedom of the central Student distribution to the best transformation model. Figure 2 is also a double graph with a common horizontal axis. The upper part of the graph shows the posterior probabilities of the Modulus and the Id models (since these are the only trully competitive models regarding this particular distribution) versus the degrees of freedom (df) of the distribution for constant sample size $n = 1000$. For fat tailed distributions (i.e. low degrees of freedom) the Modulus family is predominant, whereas as the tails reduce and the distribution all the more approximates the normal we see that the Id model takes over. The lower part of the graph depicts the behaviour of the posterior median of the transformation parameter λ_T versus the degrees of freedom under the Modulus family, clearly showing that the posterior median of λ_T approaches unity as the degrees of freedom rise beyond a certain point. This is fairly reasonable if we combine two facts: i) the Student distribution approaches the normal for increasing degrees of freedom and ii) the parameter λ_T equals 1 for normally distributed data under the Modulus family.

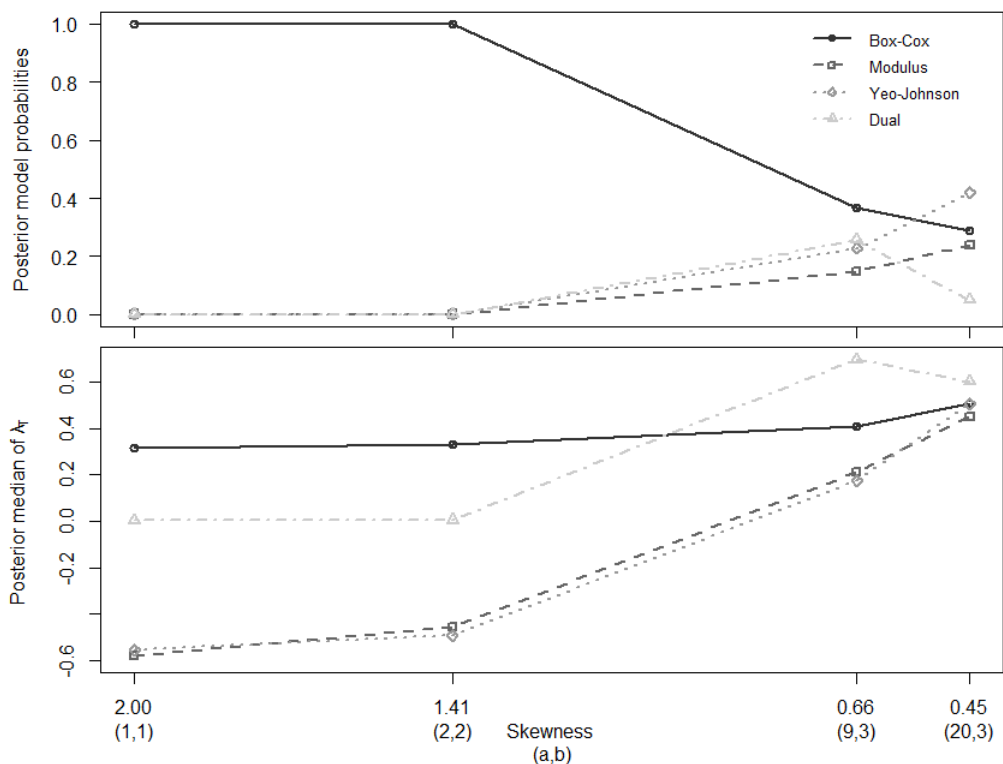


Figure 1: Posterior model probabilities $P(T|y)$ as well as the posterior median of the transformation parameter λ_T under the Box-Cox, the Modulus, the Yeo & Johnson and the Dual families according to different combinations of parameter values (a,b) of the Gamma distribution, thus according to different skewness ($n=1000$).

4. CONCLUSION

The ultimate goal of this article was to reveal the optimal choice of transformation family T and associated transformation parameter λ_T that brings a given dataset closest to normality, strictly working under the Bayesian umbrella. The compatibility issues in transformation selection have been addressed through the power-prior approach. By and large, there is more than adequate convergence of results under the power-prior and the unit-information normal prior.

The high skeweness of the Gamma distribution is best treated by the Box-Cox family whereas considerable drops in the density skewness result in equating the role of the latter family with the others more or less; the posterior median of λ_T is relatively stable around 0.3. The fat tailed Student distribution is incontestably associated with the Modulus family; the posterior median of λ_T is normally less than 0.5 but depends much on the

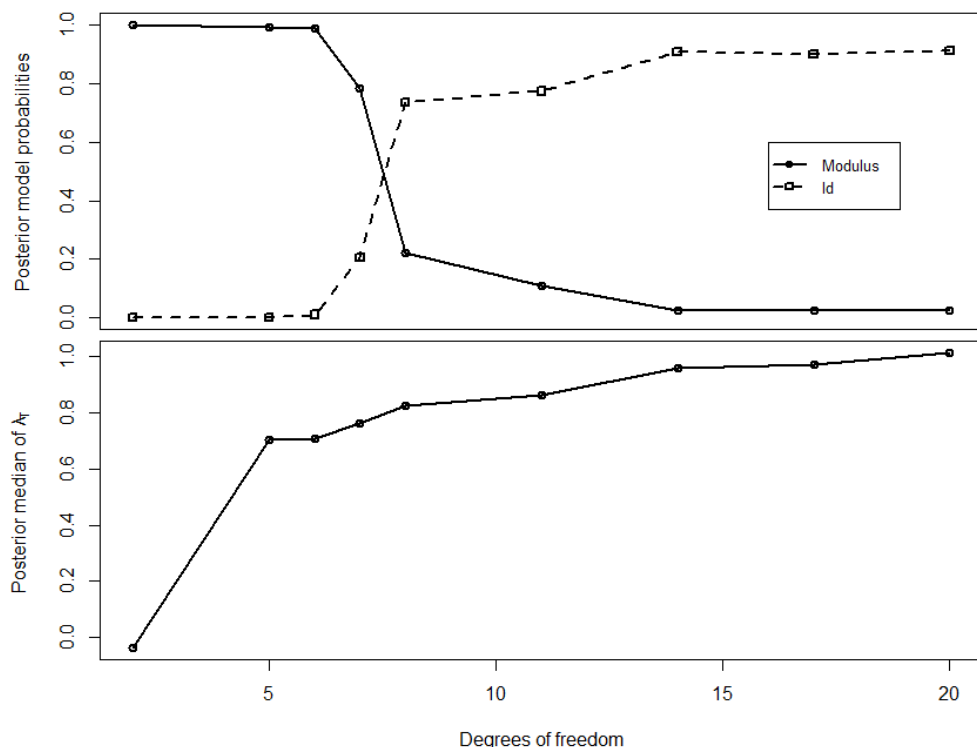


Figure 2: Posterior model probabilities $P(T|y)$ under the Modulus and the Id transformations (above) as well as the posterior median of the transformation parameter λ_T under the Modulus family (below) according to the degrees of freedom (df) of the central Student distribution ($n=1000$).

sample size and the extent of the fat tails of the density. The superiority of Modulus has been verified for other fat tailed distributions such as the Laplace (or double-exponential) distribution. In general, the results entail that the predominance of the Box-Cox transformation in the relevant literature is not always accurate and the selection from a wider transformation set should become common practice.

ΠΕΡΙΛΗΨΗ

Το πρόβλημα της επιλογής οικογένειας μετασχηματισμού εξετάζεται πλήρως από Μπεϋζιανή σκοπιά. Οι κάτωθι οικογένειες μετασχηματισμού λαμβάνονται υπόψη με σκοπό να επιτευχθεί η κανονικότητα ενός συνόλου δεδομένων μέσα από κατάλληλο μετασχηματισμό αυτών: *Box-Cox*, *Modulus*, *Yeo & Johnson* & *Dual*. Αλγόριθμοι Markov Chain Monte Carlo κατασκευάστηκαν ώστε να προσομοιώσουμε δείγμα από

την εκ των υστέρων κατανομή της παραμέτρου μετασχηματισμού λ_T η οποία συνδέεται με κάθε οικογένεια μετασχηματισμού T . Διερευνούμε διαφορετικές προσεγγίσεις με στόχο την κατασκευή συμβατών πρότερων κατανομών για την παράμετρο λ_T μεταξύ των οικογενειών, κάνοντας χρήση της power-prior μεθόδου αλλά και μιας κανονικής προσέγγισης αυτής. Ο διαχωρισμός και η ανάδειξη της βέλτιστης οικογένειας βασίζεται στον υπολογισμό των εκ των υστέρων πιθανοτήτων των οικογενειών. Χρησιμοποιώντας προσομοιωμένα δεδομένα, αναδεικνύεται η αποτελεσματικότητα της μεθοδολογίας που περιγράφεται στην παρούσα εργασία.

Acknowledgements: This work has received funding by the Research Committee of the National Technical University of Athens (Π.Ε.Β.Ε. 2010 Scheme).

REFERENCES

- Box, G.E.P. and Cox, D.R. (1964). An Analysis of Transformations (with discussion). *J. R. Stat. Soc. Ser. B*, **26**, 211-252.
- Chib, S. and Jeliazkov, I. (2001). Marginal likelihood from the Metropolis-Hastings output. *J. Amer. Statist. Assoc.*, **96**, 270-281.
- Ibrahim, J.G. and Chen, M.H. (2000). Power-Prior Distributions for Regression Models. *Statist. Sci.*, **15**, 46-60.
- John, J.A. and Draper, N.R. (1980). An alternative family of transformations. *J. R. Stat. Ser. C*, **29**, 190-197.
- Yang, Z. (2006). A modified family of power transformations. *Econ. Lett.*, **92**, 14-19.
- Yeo, I.K. and Johnson, R.A. (2000). A new family of power transformations to improve normality or symmetry. *Biometrika*, **87**, 954-959.



SMALL SAMPLE PROPERTIES OF LEAST SQUARES AND RIDGE PREDICTORS UNDER NORMAL AND NONNORMAL AUTOCORRELATED DISTURBANCES

George S. Donatos¹, George C. Michailidis²

¹University of Athens, Dept. of Economics

gdonat@econ.uoa.gr

² University of Michigan, Dept. of Statistics

gmichail@umich.edu

ABSTRACT

In this paper we study through a Monte Carlo simulation the small sample properties of the generalized and ordinary versions of some ridge predictors and compare them with least squares under autocorrelated normal and nonnormal disturbances, with small and large variances, for varied degree of multicollinearity. Our findings indicate that among the examined predictors for the criteria of bias and dispersion the least squares exhibit the best performance followed by the LW predictor, while the HKB predictor exhibits the worst performance. The McDG and RIDG predictors perform satisfactorily in most settings. All the predictors perform well under normal and uniform disturbances, while they show poor results for lognormal disturbances. Overall, their small sample properties have rather strong differences than those of the corresponding estimators.

1. Introduction

The forecasting process is closely associated with time series data which frequently are subject to autocorrelation. Moreover, multivariate real data usually suffer from the problem of almost linear dependencies among the predictor variables (multicollinearity).

In most empirical applications estimation and prediction is carried through models and in case the data analyst is faced with the second problem he might resort to the technique of ridge regression that offers some potentially useful solutions. Moreover, the presence of small samples and the fact that in many situations the distribution of the disturbance term need not be normal (see Fielitz and Roselle

(1983), and Tucker and Pond (1988)) requires a thorough examination of the ridge predictors in such a setting.

It should be noted that theory suggests that the least squares (LS) predictors will not necessarily be imprecise due to the presence of multicollinearity at least as long as the values of the regressors for which a prediction is desired follow similar patterns to the ones already present in the design matrix. However, although the problem of prediction has rarely been assessed in the presence of strong multicollinearity, the little evidence at hand suggests that ridge predictors could improve upon LS although less than when the criterion is efficiency of estimation (see Goldstein and Brown (1978)). Since the literature has not dealt with the problem of prediction under the simultaneous presence of autocorrelation and multicollinearity with normal and nonnormal disturbances: the present study, attempts to shed some light to this interesting for many practitioners topic.

The setting of our study is the linear model

$$y = X\beta + \varepsilon \quad (1)$$

where y is an $n \times 1$ vector of observations of the variable to be predicted, X a $n \times p$ full rank design matrix, β a $p \times 1$ vector of unknown coefficients and ε a $n \times 1$ vector of disturbances generated by a first order autoregressive scheme given by

$$\varepsilon_i = \rho \varepsilon_{i-1} + u_i, \quad |\rho| < 1, \quad i = 1, \dots, n \quad (2)$$

with $E(u) = 0$ and $E(uu') = \sigma^2 I_n$. The available estimators for model (1) are the least squares $\hat{\beta}$ and the generalized least squares (GLS) $\hat{\beta}_G$ estimators (for their definition and properties see Amemiya (1985)). The respective predictors are defined by $\hat{y} = X\hat{\beta}$ and $\hat{y}_G = X\hat{\beta}_G$.

To overcome the problem of multicollinearity Hoerl and Kennard (1970) proposed the ridge estimators $\hat{\beta}(k) = (X'X + kI_p)^{-1} X'y$, $k > 0$, which defines a class of biased estimators with the property that for suitable choice of the biasing parameter k it both stabilizes the estimated coefficients and improves upon the mean square error (MSE) of the LS estimator. In the presence of serially correlated disturbances we introduce the generalized ridge estimator $\hat{\beta}_G(k) = (X'\Omega^{-1}X + kI_p)^{-1} X'\Omega^{-1}y$, $k > 0$, where Ω is given by

$$\Omega = (1 - \rho^2)^{-1} \begin{pmatrix} 1 & \rho & \dots & \rho^{n-1} \\ \rho & 1 & \dots & \rho^{n-2} \\ \dots & \dots & \dots & \dots \\ \rho^{n-1} & \rho^{n-2} & \dots & 1 \end{pmatrix}$$

The respective ridge predictors are defined by $\hat{y}(k) = X\hat{\beta}(k)$ and $\hat{y}_G(k) = X\hat{\beta}_G(k)$.

One of the major topics in the ridge regression literature is the determination of the biasing parameter k . In this study, we consider the LS predictor along with five ridge predictors, namely the HKB, HK, LW, RIDG and McDG. For a detailed description of HKB, HK, LW, RIDG and McDG estimators see Donatos and Michailidis (1990).

The paper is organized as follows: in section 2 the design of the Monte Carlo experiment is presented. Section 3 contains the comparison of the predictors, while some concluding remarks are given in Section 4.

2. Design of the Simulation Experiment

The design matrix X is generated by the following procedure

$$x_{tj} = (1 - \alpha)^{1/2} z_{tj} + \alpha z_{tj}, \quad t = 1, \dots, 20; j = 1, \dots, 4 \quad (3)$$

where z_{tj} are independent uniform pseudorandom numbers in the interval (0, 1) and α takes the following values: $\alpha = .4, .6, .8, .9$.

For each set of predictors two choices for the “true” coefficient vector are considered. The first one, denoted by $\beta = bmax$, corresponds to the normalized eigenvector of the largest eigenvalue of the matrix $X'X$, while the second one, denoted by $\beta = bmin$ to the eigenvector of the smallest eigenvalue. Newhouse and Oman (1971) note that the mean squared error function subject to the constraint $\|\beta\| = 1$, is minimized by $bmax$ and maximized by $bmin$. In the absence of another systematic way of selecting the “true” underlying parameter vector, these extreme choices result in very economical and informative simulation designs. The values of the coefficients corresponding to the four choices of α are given in Table 1.

Table 1: Values of α and Coefficient Vectors
The Upper Entry is $bmax$ and the Lower $bmin$

α	β_1	β_2	β_3	β_4
.40	.1636623	.678855	.6937014	.1764905
	-.0995342	-.697941	.7091283	-.0103813
.60	.4410097	.533364	.438193	.5736027
	.0362193	-.599579	-.2846343	.7471121
.80	.4970149	.497003	.5083127	.4975767

	.4179148	.139671	-.8441387	.3054
.90	.5014548	.498993	.5005008	.4991051
	-.7153518	-.115042	.6699431	.1619057

Regarding the disturbances, a set of 5000 samples of size 20 of independently distributed, uniform random numbers in the unit interval are generated through a standard “random number generator” computer program (Michael (1972)). Then, this set of independently uniform random numbers is transformed to sets of independently distributed normal, uniform, lognormal and Laplace variates with zero mean and given variance $\sigma^2 = 1$ and 2 (see Knuth (1969)). By setting $\varepsilon_0 = 0$ in (2) and using the following values for $\rho = \pm.9, \pm.7, \pm.5, \pm.3, \pm.1$, we finally obtain the ε_i 's used in (1) and then the y_i 's. In order to make the generalized LS estimator and the generalized ridge estimators operational, we need to estimate ρ . This is achieved by the method of Prais-Winsten (1954) that suggested estimating ρ by :

$$\hat{\rho} = \frac{\sum_{i=1}^n \hat{\varepsilon}_{i-1} \hat{\varepsilon}_i}{\sum_{i=1}^n \hat{\varepsilon}_i^2}$$

where $\hat{\varepsilon}_i$'s are the LS residuals of (1), while values of $\rho \geq .5$

Finally, given that the “true” parameter vectors are known, since they are defined at the first stage of the experiment, the bias, mean squared error (MSE) and variance (Var) along with the multiple correlation coefficient (R^2) are computed for the estimates of each predictor (see Donatos (1989)).

3. Comparison of Predictors

Given the small sample summary statistics of the bias, variance, MSE and R^2 for the generalized and ordinary versions of LS and the five ridge predictors, we examine their small sample rankings. The Tables give the final ranking of the predictors. In order to make the presentation more concise we lumped together the design matrices corresponding to $\alpha = .40$ and $.60$ under the title “low multicollinearity” and those corresponding to $\alpha = .80$ and $.90$ under the title “high multicollinearity”. Similarly, values of the autocorrelation coefficient $|\rho| \leq .5$ are put together under the category “low autocorrelation”, while values of $|\rho| \geq .5$ under “high autocorrelation”. The upper part of each setting of every Table reflects the ranking of the predictors based on the following scoring: the best predictor in the respective setting receives the highest score, the second best the second highest score etc. For example a certain setting might be high multicollinearity, high autocorrelation, $\beta = bmax$ and variance of the disturbance term $\sigma^2 = 1$, which combines the results corresponding to values of $\alpha = .8$ and autocorrelation coefficients ($|\rho| \geq .5$); the predictor that is the best for each case in this setting will receive 6 points, the second best 5 points etc. Therefore, the numbers

shown in that part of the Tables represent cumulative scores of the performance of the predictors. On the other hand, the lower part of the Tables show how frequently a particular predictor in that setting exhibits a certain property in terms of MSE. Unfortunately, the limited numbers of pages doesn't allow us to present the Table of Generalized predictor for Normal errors and the Tables of Ordinary and Generalized Predictors for (i) Uniform errors, (ii) Laplace errors, (iii) Lognormal Errors. Especially the Table 2 reports the results, regarding the statistics of bias, Var, MSE of Ordinary Predictors for Normal errors.

We can summarize our main findings with respect to the various distributions of the disturbance term as follows:

(a) Normal disturbances

Among the ordinary predictors the LS one is superior to the others under all criteria, while the HKB predictor is the worst one especially for the case of low multicollinearity. The HK and LW predictors exhibit a relative poor performance for both coefficient vectors ($\beta = bmax$ and $\beta = bmin$). On the other hand the McDG and RIDG predictors give satisfactory results for low autocorrelation.

Regarding the generalized versions, the LS predictor gives the best results for all criteria but only for $\beta = bmin$. It also exhibits a strong performance for the criteria of bias and R^2 and fares well for the dispersion criteria for $\beta = bmax$. The LW predictor follows LS closely regarding the dispersion criteria but only for $\beta = bmax$, while it performs poorly for the other coefficient vector. The McDG predictor performs well for low autocorrelation for both coefficient vectors. Finally, it should be noted the poor performance of the HKB predictor and the erratic behavior of the HK predictor.

(b) Uniform disturbances

The ordinary LS predictor outperforms the other ones under all criteria. The worst performance among the ordinary predictors comes from the HKB predictor for low autocorrelation and from the RIDG (case $\beta' = bmax$) and the LW (case $\beta = bmin$) predictors for high autocorrelation respectively. The McDG predictor fares well for low autocorrelation and in particular when combined with high multicollinearity levels.

The generalized versions of the LS and LW predictors perform satisfactorily in most settings under all criteria. It should also be noted the good performance of the HK predictor when $\beta = bmin$ and especially for low values of the autocorrelation coefficient. Also for low autocorrelation and high variance of the disturbance term the McDG and RIDG predictors give satisfactory results. In settings characterized by high autocorrelation the HKB and HK predictors perform poorly in general.

(c) Laplace disturbances

Within the class of ordinary predictors, the HK predictor is the least biased, and performs better than the rest of the predictors for the other criteria under low levels of multicollinearity. On the other hand for high multicollinearity levels the LS predictor

performs satisfactorily. In almost all the settings the HKB predictor exhibits the poorest performance followed by the LW predictor for $\beta = bmin$. For low autocorrelation the good performance of the RIDG predictor should also be mentioned. The remaining predictors exhibit an inconsistent behavior.

Regarding the generalized versions, the LS predictor uniformly dominates the other predictors for the criterion of R^2 . Also, for the case $\beta = bmin$ and for the remaining criteria the LS predictor outperforms the other ones, followed by the McDG and RIDG predictors. For the case $\beta = bmax$ and the criteria of bias and dispersion the LW predictor performs better than the other ones. For low levels of autocorrelation the RIDG and HK predictors exhibit rather good results. Finally, it should be mentioned the poor overall performance of the HKB predictor and of the LW predictor for $\beta = bmin$.

(d) Lognormal disturbances

The ordinary LS predictor yields the best results, while the HKB predictor the worst results especially for low levels of multicollinearity. It is noteworthy the overall good performance of the RIDG predictor and the good results yielded by the HK and McDG predictors for low autocorrelation.

The generalized LS predictor outperforms the other ones for all criteria, although for $\beta = bmax$ and low variance of the disturbance term the LW predictor follows closely, while for high variance of the disturbance term the HK and RIDG predictors come after the LS predictor. For $\beta = bmin$ and low variance of the disturbance term the RIDG and McDG predictors fare well, while for high variance of the disturbance term the HK and RIDG predictors give satisfactory results. Finally, it should be noted the poor overall performance of the HKB predictor.

Regarding the performance of the various predictors considered in the study for the different forms of the error distribution the following remarks can be made. For all the specifications of the study the predictors exhibit their poorest performance under lognormal disturbances. With respect to the measure of bias the predictors show the best results for normal and uniform disturbances. More specifically the following predictors – HKB, HK, LW – present marginal differences with respect to these two distributions, while for the remaining predictors it is observed that their performance is best for high levels of autocorrelation under normal disturbances and for low levels under uniform ones. Regarding the dispersion measures the best results are obtained in the normal and uniform cases and in some settings under Laplace disturbances as well. The following pattern emerges in general: the predictors' performance is very satisfactory for high autocorrelation under normal disturbances and for low autocorrelation under uniform disturbances. It should also be mentioned the very good performance of the LW and HK predictors under Laplace disturbances.

One important aspect of this study is that it allows comparisons with the small sample performance of the respective generalized and ordinary ridge estimators (see Donatos and Michailidis, 1993). With respect to bias and under normal disturbances

the LS estimator outperforms the other estimators and a similar result holds for the LS predictor. On the other hand both the HKB estimator and predictor yield poor results. However, under lognormal and Laplace disturbances the HKB estimator is the least biased, while the LS estimator the most biased. The reverse relationship holds for the HKB and LS predictors. With respect to the dispersion criteria the HKB estimator exhibits a strong performance and the LS estimator a poor one; the reverse is true for the predictors. It should also be noted that the finding that for small values of the autocorrelation coefficient ρ the ordinary ridge estimators outperform both their generalized counterparts and generalized LS, especially for high levels of multicollinearity can not be extended to the ridge predictors.

Table 2, Ordinary Predictors

Normal Errors	Low Multicollinearity						High Multicollinearity					
	LW	McDG	HKB	HK	RIDG	LS	LW	McDG	HKB	HK	RIDG	LS
	S ² =1 . bmax											
	High Autocorrelation											
Bias	36	23	8	23	31	47	38	28	15	14	26	47
Var	31	22	12	25	33	45	41	27	15	16	26	43
MSE	34	21	11	23	31	48	40	25	14	16	28	45
R ²	32	23	11	22	33	47	36	26	16	19	24	47
min MSE	0	0	0	0	0	8	3	0	0	0	0	5
max MSE	1	2	5	0	0	0	0	1	4	3	0	0
	Low Autocorrelation											
Bias	60	37	21	15	47	72	58	39	36	12	37	70
Var	60	42	27	14	41	68	60	40	37	13	32	70
MSE	61	49	24	12	37	69	59	44	32	12	33	72
R ²	60	42	27	14	41	68	59	41	37	12	32	71
min MSE	3	0	0	0	0	9	0	0	0	0	0	12
max MSE	0	0	4	8	0	0	0	0	0	12	0	0
	S ² =2 . bmax											
	High Autocorrelation											
Bias	32	33	12	16	28	47	34	38	20	9	21	46
Var	21	22	16	26	48	35	25	20	14	26	48	35
MSE	25	29	13	20	39	42	35	27	15	19	32	40
R ²	21	24	15	23	48	37	24	24	13	22	38	37
min MSE	0	0	0	0	2	6	2	0	0	0	1	5
max MSE	1	0	7	0	0	0	0	0	6	2	0	0

Normal Errors	Low Multicollinearity						High Multicollinearity					
	LW	McDG	HKB	HK	RIDG	LS	LW	McDG	HKB	HK	RIDG	LS
	Low Autocorrelation											
Bias	56	27	12	36	53	68	54	41	22	21	48	66
Var	43	37	21	23	72	56	45	38	24	15	72	58
MSE	51	34	16	29	60	62	47	43	22	21	56	63
R ²	45	33	18	27	72	57	46	35	20	21	72	58
min MSE	0	0	0	0	5	7	0	0	0	0	4	8
max MSE	0	1	9	2	0	0	0	1	5	6	0	0
	S ² =1 . bmin											
	High Autocorrelation											
Bias	24	27	9	35	26	47	9	29	16	33	34	47
Var	15	26	15	37	28	47	8	29	18	34	32	47
MSE	19	26	12	35	28	48	10	29	17	33	32	47
R ²	15	27	15	36	27	48	8	26	18	35	33	48
min MSE	0	0	0	0	0	8	0	1	0	0	0	7
max MSE	4	1	3	0	0	0	6	0	2	0	0	0
	Low Autocorrelation											
Bias	42	34	14	57	33	72	17	48	21	57	39	70
Var	31	47	21	51	33	69	12	49	28	55	38	70
MSE	37	37	19	53	34	72	16	49	23	55	38	71
R ²	32	47	20	50	33	70	12	48	27	55	38	72
min MSE	0	0	0	0	0	12	0	1	0	0	0	11
max MSE	1	0	9	0	2	0	9	0	3	0	0	0
	S ² =2 . bmin											
	High Autocorrelation											
Bias	27	36	11	17	30	47	12	40	25	16	28	47
Var	15	24	18	27	48	36	8	25	21	31	48	35
MSE	21	27	17	17	42	44	12	29	22	21	40	44
R ²	13	26	15	30	45	39	8	27	19	38	34	42
min MSE	0	0	0	0	2	6	0	0	0	0	2	6
max MSE	2	1	5	0	0	0	6	0	1	1	0	0
	Low Autocorrelation											
Bias	32	33	16	52	51	68	14	46	27	43	56	66
Var	17	44	26	38	72	55	12	46	31	34	72	57

Normal Errors	Low Multicollinearity						High Multicollinearity					
	LW	McDG	HKB	HK	RIDG	LS	LW	McDG	HKB	HK	RIDG	LS
MSE	20	39	23	46	59	65	12	46	28	39	65	62
R ²	19	38	22	48	71	54	12	44	28	58	40	70
min MSE	0	0	0	0	4	8	0	0	0	0	6	6
max MSE	3	1	8	0	0	0	12	0	0	0	0	0

Finally, for both the estimators and the predictors the best results are yielded under normal and uniform disturbances and the worst under lognormal disturbances. Unfortunately, no other simulation studies have examined ridge predictors and therefore no comparisons can be made.

4. Concluding Remarks

The purpose of this simulation study was to gain some insight regarding the performance of some ridge predictors along with LS under autocorrelated normal and nonnormal disturbances.

In most settings, and for all the examined criteria LS perform better than the ridge predictors, followed closely by the LW predictor. The RIDG and McDG predictors perform well in some settings only, while the HKB performs poorly overall. Moreover, the generalized versions of the predictors outperform their ordinary counterparts. All the examined predictors perform better under normal and uniform disturbances, while the least satisfactory results come under lognormal disturbances.

The findings of the present simulation study confirm the fact that even under autocorrelated normal and nonnormal disturbances and in the presence of high multicollinearity levels the gains of ridge predictors in predictive accuracy over LS can be characterized at most as marginal. Moreover, in all other settings the LS predictor is superior to the ridge predictors. However, it remains an open question the performance of the ridge predictors for data points different than the design points. Finally, another important finding is that the ridge algorithms that perform well in estimation perform poorly in prediction and the RIDG and McDG algorithms emerge as an interesting compromise.

ΠΕΡΙΛΗΨΗ

Στην εργασία αυτή διερευνώνται, με χρήση πειραμάτων Monte Carlo, οι ιδιότητες γενικευμένων και απλών προβλεπτών ridge για μικρά δείγματα, σε σχέση με τις ιδιότητες γενικευμένων και απλών προβλεπτών ελαχίστων τετραγώνων για αυτοσυσχετιζόμενα κανονικά και μη κανονικά σφάλματα, με την παρουσία χαμηλής

και υψηλής πολυσυγγραμμικότητας. Τα ευρήματα έδειξαν ότι μεταξύ των προβλεπτών, που εξετάστηκαν ως προς τα κριτήρια της αμεροληψίας και αποτελεσματικότητας, ο προβλέπτης των ελαχίστων τετραγώνων δείχνει να πλεονεκτεί, ακολουθούμενος από τον προβλέπτη LW, ενώ ο προβλέπτης HKB εμφανίζεται να μειονεκτεί, έναντι των άλλων. Οι προβλέπτες McDG και RIDG εμφανίζουν ιδιαίτερα ικανοποιητική συμπεριφορά στις περισσότερες περιπτώσεις. Όλοι οι προβλέπτες παρουσιάζουν ικανοποιητικά αποτελέσματα για σφάλματα που ακολουθούν την κανονική και ομοιόμορφη κατανομή, ενώ δεν έχουν ικανοποιητικές επιδόσεις για σφάλματα που ακολουθούν την λογαριθμοκανονική κατανομή.

References

- Amemiya, T., (1986). *Advanced Econometrics*, Harvard University Press.
- Donatos, G.S., (1989). "A Monte Carlo Study of k-class estimators for small samples with normal and nonnormal disturbances", *The Statistician*, **38**, 11-20.
- Donatos, G.S. and Michailidis G.C., (1990). "Small sample properties of ridge estimators with normal and nonnormal disturbances", *Communications in Statistics, Simulation and Computation*, **19**, 935-950.
- Donatos, G.S. and Michailidis G.C., (1993). "A simulation study of least squares and ridge estimators for normal and nonnormal autocorrelated disturbances", *Journal of Statistical Simulation and Computation*, **47**, 49-66.
- Fieltz, B. and Roselle J., (1983). "Stable distributions and the mixture of distributions hypotheses for common stock returns", *Journal of the American Statistical Association*, **78**, 28-36.
- Goldstein, M. and Brown, P.J., (1978). "Prediction with shrinkage estimators", *Mathematische Operationforschung und Statistik, Series Statistics*, **9**, 3-7.
- Hoerl, A.E. and Kennard, R.W., (1970). "Ridge regression: Biased estimation for nonorthogonal problems", *Technometrics*, **12**, 56-67.
- Knuth, D.E., (1968). *The Art of Computer Programming, 2 Seminumerical Algorithms*. Addison-Wesley, Reading, Mass.
- Mikhail, W.M., (1972). "Simulating small sample properties of econometric estimators", *Journal of American Statistical Association*, **67**, 620-624.
- Newhouse, J.P. and Oman, S.D., (1971). "An evaluation of ridge estimators", Rand Report. No R-716-PR, 1-28.
- Prais, S.J. and Winsten, C.B., (1954). "Trend estimators and serial correlation", Unpublished Cowles Commission Study Paper. Stat. No **383**, Chicago.
- Tucker, A. and Pond L., (1988). "The probability distribution of foreign exchange price changes", *Review of Economics and Statistics*, **70**, 638-647.



A MARKOV CHAIN APPROACH FOR MODELING USA BANKS' CAPITALIZATION

Konstantinos Drakos, Vasileios Koutras

Department of Accounting and Finance

Athens University of Economics and Business

kdrakos@aueb.gr, vkoutras@aueb.gr

ABSTRACT

In the present article we exploit a Markov Chain setup for US banks' capitalization and conduct a comparison of the two past decades (1992-2000 versus 2001-2009) in terms of mobility and persistence. The states of the Markov Chain are based on the FDIC discretized version of Capital Ratio. We document a substantial increase in the 00's on system mobility, and what is more important, mobility takes the form of increased probability of capitalization deterioration.

Keywords: Capitalization, FDIC, Markov Chain, Transition Matrix, USA Banks.

1. Introduction

The Basel agreement has put in place a capitalization threshold, relative to risk-weighted assets, above which banks are advised to operate. In particular, banks are expected to maintain a Capital Ratio (CAR hereafter), defined as $\text{Tier1} + \text{Tier2} / \text{Risk-Weighted Assets}$, of at least 8%. The Federal Deposit Insurance Corporation (FDIC hereafter), the US banking sector regulator, has adopted a finer classification of capitalization based on a bucketing approach. In particular, according to the FDIC a bank is classified in one of the following buckets: (1) *Critically Undercapitalized* if $\text{CAR} < 2\%$, (2) *Significantly Undercapitalized* if $2\% \leq \text{CAR} < 6\%$, (3) *Undercapitalized* if $6\% \leq \text{CAR} < 8\%$, (4) *Adequately Capitalized* if $8\% \leq \text{CAR} < 10\%$ and (5) *Well Capitalized* if $\text{CAR} \geq 10\%$. This bucketing, at least for regulatory purposes, it is preferred to the simple dichotomous Basel classification that characterizes a bank either as undercapitalized or not, depending on whether its Capital Ratio falls below or above 8%. The intuition behind this bucketing is that not all undercapitalized banks should be treated as identical.

In the present study we employ a panel dataset whose cross section consists of all US commercial banks and its time series dimension of the period 1992-2009

(annual intervals). Using a Markov Chain setup whose states correspond to the FDIC buckets, we estimate the transition probabilities between capitalization buckets; including the default state. The benefit of this approach is that it allows the quantification of migration probabilities of banks across the capitalization spectrum, among which those related to the default state are of special interest. Our analysis includes estimating direction-specific migration probabilities, i.e. improvements vs. deteriorations of capitalization. This information is particularly useful both for market participants, but especially for the regulator, since it helps tracing any similarities and/or differences effectively over the whole period in which the Basel requirements are in place.

The remainder of the paper is organized as follows. Section 2 describes the data used for the analysis. Section 3 presents the empirical methodology followed, while Section 4 discusses the empirical results.

2. Data issues

The dataset consists of all FDIC-insured commercial banks in USA for the period 1992-2009 (183,739 bank-year observations). Let the indices k, t denote the cross-sectional unit (bank) and time period (year) respectively. Then, exploiting the FDIC buckets, we may discretize the CAR variable by defining a new variable $CAP_{k,t}$ as follows:

$$CAP_{k,t} = \begin{cases} 1, & \text{if bank } k \text{ is well capitalized in year } t \\ 2, & \text{if bank } k \text{ is adequately capitalized in year } t \\ 3, & \text{if bank } k \text{ is undercapitalized in year } t \\ 4, & \text{if bank } k \text{ is significantly undercapitalized in year } t \\ 5, & \text{if bank } k \text{ is critically capitalized in year } t \end{cases}$$

Table 1 shows the distribution of banks across CAP categories by year. The annual overall unconditional distributions are quite similar across years, with the vast majority of banks belonging to the Well Capitalized category, followed by those belonging to the Adequately Capitalized category.

Table 1. Distribution of banks across capitalization bucket & year

	Critically Undercapitalized	Significantly Undercapitalized	Under-capitalized	Adequately Capitalized	Well Capitalized	# of Defaults	# of banks
1992	0.78	0.66	0.67	3.41	93.20	180	13,822
1993	0.53	0.23	0.18	1.73	96.96	50	13,196
1994	0.06	0.11	0.22	1.54	97.96	15	12,529
1995	0.02	0.04	0.18	1.13	98.57	8	11,902
1996	0.01	0.08	0.11	1.48	98.26	6	11,394
1997	0.03	0.04	0.11	1.78	98.03	1	10,871
1998	0.01	0.09	0.21	1.91	97.75	3	10,420
1999	0.02	0.04	0.09	2.15	97.63	8	10,184
2000	0.03	0.05	0.05	2.26	97.54	7	9,867
2001	0.04	0.04	0.16	2.22	97.50	4	9,581
2002	0.02	0.02	0.06	1.39	98.50	11	9,325
2003	0.02	0.03	0.03	0.99	98.92	3	9,154
2004	0.00	0.02	0.02	0.94	99.02	4	8,951
2005	0.00	0.02	0.06	0.76	99.16	0	8,809
2006	0.01	0.00	0.07	0.64	99.28	0	8,657
2007	0.00	0.06	0.11	1.03	98.80	3	8,513
2008	0.10	0.37	0.74	2.28	96.51	25	8,288
2009	0.25	0.94	1.15	2.15	95.51	143	7,994
All years	0.11	0.16	0.23	1.66	97.73	471	183,739

Notes: (a) Numbers denote percentages, (b) The number of banks decreases each year due to M&A's and bank defaults, (c) Sums may differ from 100 due to rounding errors.

3. Empirical methodology

3.1 A Markov Model for Capital Ratio

Let us first present the basic terminology and techniques we will use from the theory of Markov models (see e.g. Ross (1996), Noris (1998)). Consider a sequence of random vectors $\{\mathbf{d}_t\}_{t \geq 0}$ as denoting the distribution of banks across capitalizations buckets (states) each year, and the state-space $\mathfrak{S} = \{1, 2, \dots, n\}$. The parameters of the Markov model $[\{\mathbf{d}_t\}_{t \geq 0}, \mathfrak{S}]$ are its transition probability matrix $\mathbf{P}$ and its initial distribution π_{0i} , $i \in \mathfrak{S}$. The evolution of the distribution $\mathbf{d}_t$ over time can be assessed by exploiting the recurrence $\mathbf{d}_t = \mathbf{P} \cdot \mathbf{d}_{t-1}$. The typical element p_{ij} of the transition probability matrix $\mathbf{P}$ indicates the transition probability that a bank is in state (capitalization bucket) j at time t , given it was in state i in time $t-1$. By definition $p_{ij} \geq 0$, $\forall i, j$ and since the set of states is exhaustive, the row elements of

the matrix $\mathbf{P}$ add up to unity: $\sum_{j=1}^n p_{ij} = 1, \forall i$. It is not unusual for a Markov chain to possess an absorbing state, in which once the chain enters, it is impossible to escape. In the case of banks this absorbing state is of crucial importance since it corresponds to the default of a financial institution. The absorbing state is typically treated as the last state in the state space $\mathfrak{S} = \{1, 2, \dots, n\}$, in which case the transition probability matrix takes the form:

$$\mathbf{P} = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1n-1} & p_{1n} \\ p_{21} & p_{22} & \cdots & p_{2n-1} & p_{2n} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 0 & 1 \end{bmatrix}.$$

Let now $N_{ij} \equiv \sum_t \mathbf{1}(d_{t-1} = i, d_t = j)$ be the count of all transitions from state i to state j in the reference period, where $\mathbf{1}(\alpha)$ is an indicator function that takes the value of unity when the argument α is true, and zero otherwise. Let us also define $N_{i\bullet} \equiv \sum_{j \in \mathfrak{S}} N_{ij}$, $N_{ij} = \sum_{j=1}^n N_{ij}$ as the number of transitions starting from state i . Then the likelihood function, takes on the form $\mathcal{L} = p_0 \prod_{i \in \mathfrak{S}} \prod_{j \in \mathfrak{S}} p_{ij}^{N_{ij}}$ and the log-

likelihood reads $\ln \mathcal{L} = \sum_{i \in \mathfrak{S}} \sum_{j \in \mathfrak{S}} N_{ij} \ln p_{ij}$.

The likelihood maximization yields:

$$\hat{p}_{ij} = \frac{N_{ij}}{\sum_{j \in \mathfrak{S}} N_{ij}} = \frac{N_{ij}}{\sum_{j=1}^n N_{ij}} = \frac{N_{ij}}{N_{i\bullet}}$$

as the asymptotically unbiased and normally distributed ML estimator of p_{ij} (see Anderson and Goodman (1957)).

3.2 Capital Ratio Migrations

In a transition matrix, important information is encapsulated in its main diagonal, for it denotes the probability that the chain remains in the same state between two successive periods. This property is usually referred as persistence. In the context of bank capitalization, persistence is perceived as a situation where a bank tends to remain in its current capitalization bucket. On the other hand, the off-

diagonal elements of the transition matrix describe the ability of the Markov chain to shift in deferent states between successive periods.

In the extreme case where the off-diagonal elements are zero, the probabilities of remaining in the same bucket between any two consecutive periods would equal 1; therefore we have the absolute (or complete) persistence case, which will be described by a transition probability matrix coinciding with the identity matrix **I**.

The degree of mobility exhibited in an empirical transition matrix can be assessed by the so-called indices, Jafry and Schuermann (2004), Shorrocks (1978), Sommers and Conlisk, (1979), Theil (1972), Trück and Rachev (2006, 2009). There are two classes of mobility indices that have been suggested in the relevant literature: the **Summary Mobility Indices** and the **Eigenvalue Based Indices**.

The Summary Mobility Indices include the **Immobiity Ratio (IR)**, **Moving Up (MU)**, and **Moving Down (MD)**. Their main function is to assist us in conducting a preliminary analysis of the empirical transition matrix of interest, say $\mathbf{P} = (p_{ij})$, and they are expressed as follows:

$$IR = \left[\frac{\sum_{i=1}^n p_{ii}}{\sum_{i=1}^n \sum_{j=1}^n p_{ij}} \right] \times 100 = \left[\frac{\sum_{i=1}^n p_{ii}}{n} \right] \times 100, \quad MU = \left[\frac{\sum_{i < j} p_{ij}}{\sum_{i=1}^n \sum_{j=1}^n p_{ij}} \right] \times 100 = \left[\frac{\sum_{i < j} p_{ij}}{n} \right] \times 100,$$

$$MD = \left[\frac{\sum_{i > j} p_{ij}}{\sum_{i=1}^n \sum_{j=1}^n p_{ij}} \right] \times 100 = \left[\frac{\sum_{i > j} p_{ij}}{n} \right] \times 100.$$

The immobiity ratio index gives the percentage of the banks, whose bucket capitalization classification remained intact for the past year. The other two indices reflect the percentage of the banks, which changed capitalization bucket, moving to a higher/lower one within a year, and therefore pertain to the system's mobility.

It should be stressed that, according to the coding adopted in the present analysis, a moving of $CAP_{k,t}$ to a higher (lower) state designates a deterioration (improvement) of the bank capitalization. Having this in mind, the Moving Up index reflects the percentage of the banks, whose capitalization bucket changed, moving to a lower one during the last period while the Moving Down index reflects the percentage of the banks, whose capitalization bucket changed, moving to a higher one during the last period.

The Eigenvalue Based Indices are expressed in terms of the eigenvalues of the empirical transition matrix. The most popular indices in this class are the **Prais-Shorrocks Index**, **Sommers-Conlisk Index**, **Shorrocks Index** and the **Half Life Index**, which we shall denote by M_{PS} , M_{SC} , M_S and M_h respectively. If λ_i , $i = 1, 2, \dots, n$ are the transition matrix eigenvalues in a descending order (in absolute value), then the above mentioned indices are defined as follows:

$$M_{PS} = \frac{n - \text{tr}(P)}{n - 1} = \frac{n - \sum_{i=1}^n \lambda_i}{n - 1}, \quad M_{SC} = 1 - |\lambda_2|, \quad M_S = 1 - |\det(P)| = 1 - \left| \prod_{i=1}^n \lambda_i \right|,$$

$$M_h = e^{-h}, \text{ where } h = \frac{-\log 2}{\log |\lambda_2|}.$$

All indices, with the exception of the M_h , assume values in the $[0, 1]$ interval, with 1 denoting the highest degree of mobility and 0 the lowest. The M_h indicates the speed of convergence towards the equilibrium distribution $\Pi = \lim_{r \rightarrow \infty} P^r$.

Another index which is indirectly related to the eigenvalues of a matrix is the **Singular Value Decomposition Index (SVD)** which is given by the formula $M_{SVD} = \sum_{i=1}^n \sqrt{\lambda_i^*} / n$ where λ_i^* denote the (positive) eigenvalues of the matrix $(P - I)'(P - I)$. The matrix $P - I$ is usually referred to as mobility matrix. The square roots of the (non-negative) eigenvalues of the matrix $(P - I)'(P - I)$ are called singular values and therefore M_{SVD} is simply the average of the singular values of the transition probability matrix P .

4. Empirical results

4.1 Estimation of sample transition matrices

As mentioned earlier, our main objective is the comparison of the capitalization bucket migrations for the periods 1992-2000 and 2001-2009. To achieve that we shall first compute the empirical transition probability matrices referring to those two periods. Using the data of the period 1992-2000 we estimated the transition probabilities between all 6 states (see Section 3.1), thereof obtaining the empirical transition probability matrix P_{pre} shown in Table 2. As expected, banks belonging to any of the undercapitalized categories run the highest default risk within a year (5.49%, 29.93% and 20.45%). Surprisingly, that the Significantly Undercapitalized category has the highest probability of default (about 30%), while one would expect the Critically Undercapitalized group to exhibit the highest default probability. The Adequately Capitalized and Well Capitalized categories have very low probabilities of default, estimated at 0.66% and 0.02% respectively.

Persistence is substantially stronger at both ends of the capitalization spectrum. In particular, Well Capitalized banks tend to remain in the same category almost with certainty (with probability of 98.82%), while Critically Undercapitalized banks retain their capitalization bucket with probability of about 64%. The other category with

notable, but much lower, persistence is that of Adequately Capitalized banks which tend to retain their status with a probability just above 34%.

Table 2. Transition Probability Matrix for the period 1992-2000 (matrix $\mathbf{P}_{pre}$)^{(a), (b)}

From / To	Well Capitalized	Adequately Capitalized	Undercapitalized	Significantly Undercapitalized	Critically Undercapitalized	Default
Well Capitalized	0.9882	0.0106	0.0007	0.0002	0.0001	0.0002
Adequately Capitalized	0.6049	0.3411	0.0282	0.0144	0.0048	0.0066
Undercapitalized	0.5055	0.2582	0.0659	0.0989	0.0166	0.0549
Significantly Undercapitalized	0.3139	0.1241	0.0949	0.1241	0.0437	0.2993
Critically Undercapitalized	0.1023	0.0114	0.0341	0.0113	0.6364	0.2045
Default	0	0	0	0	0	1

Notes: (a) Numbers denote percentages (b) Sums do not add up to 100 due to rounding errors

Working in a similar fashion with the 2001 to 2009 part of the dataset we computed the empirical transition probability matrix $\mathbf{P}_{post}$, shown in Table 3.

Table 3. Transition Probability Matrix for the period 2001-2009 (matrix $\mathbf{P}_{post}$)^{(a), (b)}

From / To	Well Capitalized	Adequately Capitalized	Undercapitalized	Significantly Undercapitalized	Critically Undercapitalized	Default
Well Capitalized	0.9851	0.0102	0.0021	0.0011	0.0004	0.0011
Adequately Capitalized	0.6795	0.1993	0.0396	0.0291	0.0035	0.0490
Undercapitalized	0.2581	0.1075	0.0538	0.1720	0.0430	0.3656
Significantly Undercapitalized	0.3333	0.0000	0.0000	0.0000	0.0000	0.6667
Critically Undercapitalized	0.3846	0.0000	0.0769	0.0000	0.0000	0.5385
Default	0	0	0	0	0	1

Notes: (a) Numbers denote percentages (b) Sums do not add up to 100 due to rounding errors

This matrix gives a similar picture for the movements between *CAP* categories in the period 2001-2009, albeit the transitions to lower capitalization buckets are now much more intense than the ones observed in $\mathbf{P}_{pre}$. Banks belonging to any of the undercapitalized categories exhibit again the highest default risk within a year (36.56%, 66.67% and 53.85%) which are much higher than the respective risks

estimated for the period 1992-2000. The rather counter-intuitive fact of not observing the highest probability of default in the Critically Undercapitalized group is still present in this period (Significantly Undercapitalized category has instead, the highest probability of default). Adequately Capitalized and Well Capitalized categories have probabilities of default estimated at 4.9% and 0.11% respectively.

Persistence is now observed only on the upper end of the capitalization spectrum, with Well Capitalized banks tending to remain in this category almost with certainty (98.51%), while Critically and Significantly Undercapitalized banks have zero probability to retain their capitalization bucket. The other category with noteworthy persistence is that of Adequately Capitalized banks which tend to retain their status with a probability of almost 20%.

A substantial difference between matrices $\mathbf{P}_{pre}$ and $\mathbf{P}_{post}$ is that in the latter there exist some zero entries (besides the ones appearing in the absorbing state row) which means that some states are not accessible by other states in one step, i.e. between successive years.

4.2 Calculation of Mobility indices

Using the empirical transition matrices $\mathbf{P}_{pre}$ and $\mathbf{P}_{post}$ we calculated the values of the three summary mobility indices, the eigenvalue based indices as well as the SVD mobility index, which are reported in Table 4.

Table 4. Mobility indices for transition probability matrices		
Index	1992-2000	2001-2009
Panel I: Summary Mobility Indices		
Immobility Ratio	0.52595	0.373033
Moving Up (deterioration)	0.130617	0.320317
Moving Down (improvement)	0.343433	0.30665
Panel II: Eigenvalue Based Mobility Indices		
Prais-Shorrocks(M_{PS})	0.56886	0.75236
Sommers-Conlisk(M_{SC})	0.0007007766	0.0043276
Shorrocks(M_S)	0.999568	1
Half Life (h)	988.781	159.822
Panel III: Singular Value Based Indices		
Mobility Index(M_{SVD})	0.567265	0.752142

Let us start the discussion of the results with the summary mobility indices. It is clear that the MU ratio is greater for the period 2001-2009 as compared to the period 1992-2000. This implies that after 2001 there were more transitions towards lower capitalization buckets as compared to the pre 2001 period. On the other hand, the MD ratio is rather similar between the two transition matrices. Therefore, based on the

summary mobility indices, we may conclude that mobility has substantially increased in the post 2001 era, with the majority of transitions involving movements towards deterioration of capitalization.

The same conclusions are drawn, should one resort to the eigenvalue-based mobility indices. Recall that all eigenvalue-based mobility indices M_{PS} , M_{SC} and M_S take on values in $[0, 1]$ interval; yet, some of them may take either very large values (M_S) or very small values (M_{SC}), therefore our conclusions should be drawn by comparing their values for the two periods we are interested in rather than considering their closeness to 1 as a mobility evidence. The Prais-Shorrocks index (M_{PS}), reveals a medium level of mobility (56.9%) for the 1992-2000 transition matrix. For the period 2001-2009 the same index shows a much greater mobility (75.2%). The singular value based index M_{SVD} whose values are very close to the ones attained by the Prais-Shorrocks index leads to exactly the same conclusions. The M_{SC} index paints the same picture as the M_{PS} mobility index, if we take into account the fact that its values for the period 2001-2009 is much higher than the one exhibited for the period 1992-2000 (note that the respective ratio is about $0.0043/0.0007=6.143$).

Needless to say, the eigenvalue and singular value based indices, cannot differentiate between alternative configurations in the off-diagonal elements of the transition probability matrices and consequently do not provide any information with respect to the distribution of probabilities across neighboring and distant states.

The estimated half-life measure M_h shows that it has decreased substantially from 988 for the period 1992-2000 to 159 for the period 2001-2009. This finding reveals that for the post 2000 era there is considerably higher mobility than in the pre 2000. To confirm this, we mention that, looking at the transition probability matrices of order $2h$ (which is almost 320 for the period 2001-2009 and 1980 for the post 2001 period), one can easily verify that in the matrix $\mathbf{P}_{post}^{320}$ the probabilities of default (no matter which the initial state was) exceed 75%, and the same holds true when we consider the matrix $\mathbf{P}_{pre}^{1980}$.

To recap, looking at the mobility of the US banking sector, defined as the tendency to change capitalization bucket, we document (exploiting a battery of mobility indices) a ***substantial increase in the post 2000 era. What is more important is that this mobility takes the form of increased probability of capitalization deterioration. Moreover, during the post 2001 period these deteriorations were more likely to transcend one capitalization bucket.*** That is during the 00's, in comparison to the 90's, the US banking sector has exhibited greater mobility, with capitalization deteriorations (including defaults) been more likely and more abrupt.

Our analysis has several policy relevant findings. Firstly, it offers (annual) capitalization bucket-specific transition probabilities towards all possible states. This is useful for the regulator who can anticipate likely movements across the

capitalization spectrum, with special reference to the default state. Secondly, it provides a comparison of capitalization movement experiences between the past two decades, therefore highlighting possible developing trends.

REFERENCES

- Anderson, T. and L. Goodman, (1957). Statistical inference about Markov Chains. *Annals of Mathematical Statistics*, **28**, 12-40.
- Jafry, Y., and T. Schuermann, (2004). Measurement, estimation and comparison of credit migration matrices. *Journal of Banking and Finance*, **28**, 2603–2639.
- Noris, J., (1998). *Markov Chains*, Cambridge University Press.
- Ross, S., (1996). *Stochastic Processes* (2nd Edition), John Wiley & Sons Inc., New York.
- Shorrocks, A., (1978). The measurement of mobility. *Econometrica*, **46**, 1013-1024.
- Sommers, P., and J. Conlisk, (1979). Eigenvalue immobility measures for Markov Chains. *Journal of Mathematical Sociology*, **6**, 253-276.
- Theil, H., (1972). *Statistical Decomposition Analysis*, North Holland, Amsterdam.
- Trück, S., and T. Rachev, (2006). Changes in migration matrices and Credit VaR - a new class of difference indices. Working Paper, Queensland University of Technology, Australia.
- Trück, S., and T. Rachev, (2009), *Rating Based Modeling of Credit Risk Theory and Applications of Migration Matrices*, Elsevier Inc.



STATISTICAL INVARIANCE DUE TO AFFINE TRANSFORMATIONS

Christos P. Kitsos

Department of Informatics, Technological Educational Institute of Athens, Greece

xkitsos@teiath.gr

ABSTRACT

The target of this paper is to use the Affine Geometry to prove transformation invariance properties in Statistics. A number of Mathematical properties are introduced and the application to Statistics is discussed.

Keywords: Affine geometry, invariance, logit model.

1. INTRODUCTION

In principle, there are statistical approaches, either on design theory or in estimation, with a strong mathematical insight. There also cases where the developed procedure, despite the fact that “works well”, although there is a lack of theoretical insight. This paper is based on these cases: tries to offer a strong mathematical background for the proposed method.

Through the usual Euclidean distance $d(x, y) = \sum_{i=1}^n (x_i - y_i)^2$, $x, y \in \mathbb{R}^n$, there are defined other distance measures, such that

$$d_1(x, y) = e^{d(x, y)} - 1, \quad x, y \in \mathbb{R}^n, \text{ or}$$

$$d_2(x, y) = \min \{ |x_i - y_i|, \quad i = 1, 2, \dots, n \}, \quad x = (x_i)_{i=1}^n, y = (y_i)_{i=1}^n \in \mathbb{R}^n.$$

Example 1. In statistical theory the concept of distance is fundamental: In principle, statistical estimation is based on the minimum distance of $d(F_\theta, F_{\theta_2})$, with F_θ , $\theta \in \Theta$ the parameter space, from the class of distribution functions from the class $P = \{P_\theta, \theta \in \Theta\}$ of all probability measures. In such a case let us define the empirical distribution

$$\text{Emp}(X) = \frac{1}{n} \sum_{j=1}^n I_{A_j}(\omega),$$

with $A_j = \{x_j(\omega) \leq x\}$, $x_1, x_2, \dots, x_n$ are coming from F , and I_A the characteristic function of A . Then, two typical examples of estimation adopting the notion of distance, besides the Maximum Likelihood Estimation (MLE) and the Least Squares Estimation (LSE) are the :

$$d(F_\theta, \text{Emp}) = \begin{cases} \sup\{|F_\theta - \text{Emp}|\}, & (\text{Kolmogorov-Smirnov}) \\ n \sum_{j=1}^n \frac{(P_\theta - \text{Emp})^2}{\text{Emp}}, & (\chi^2\text{-Statistic}) \end{cases}$$

The values of F_θ, P_θ , and Emp are at $x \in \mathbb{R}$. See for details, Blyth (1970).

Example 2. Recall the Latin Square of order $k \in \mathbb{N}$: By definition is a $k \times k$ array in which each of the k symbols (usually Latin letters) occurs once in each row and once in each column. It can be proved that for $m \geq 2$, the $m \times m$ array defined by

$$L(i, j) = i + j, \quad i, j \in \mathbb{Z}_m,$$

with $\mathbb{Z}_m$ being the set of integers modulo m , is a Latin Square.

For a $k \times k$ Latin Square, recall the main part of an ANOVA Table, when the Analysis of Variance (or Regression Analysis) is adopted, see Graybill (1976) among others:

Table 1. ANOVA (main part).

Source of Variation	df	SS
Rows	$k - 1$	$\ \bar{y}_{\cdot j} - \bar{y}_{\dots}\ ^2$
Columns	$k - 1$	$\ \bar{y}_{i \cdot} - \bar{y}_{\dots}\ ^2$
Treatment	$k - 1$	$\ \bar{y}_{i \cdot} - \bar{y}_{\dots}\ ^2$
Error	$\ \bar{y}_{i \cdot} - \bar{y}_{\dots}\ ^2$	by subtraction
Total	$k^2 - 1$	$\ y - \bar{y}_{\dots}\ ^2$

df: degrees of freedom, SS: sum of squares

Topology is a more “primitive” concept than the one of distance. The induced, by $d(x, y)$, topology on $\mathbb{R}^n$ is the same with that of $d_1(x, y)$, $d_2(x, y)$, i.e. the topology roughly speaking, is not very depended of the form of the defined distance on $\mathbb{R}^n$.

Now, the idea is to think of a set of points, say M , so that locally to resemble with $\mathbb{R}^n$. This is behind of what we call the set M is a manifold:

- each point, say P of M can have an open neighborhood,

- ii. which has a continuous 1-1 and onto map, say f , of an open set of $\mathbb{R}^n$ (for some n).

We emphasize that the map f needs to be 1-1; it is not asked the lengths or the angles to be preserved, but only states that $f(U) \subseteq \mathbb{R}^n$ for the manifold M . The pair (U, f) is known as a chart. So, we only express a practical need: to present the set M “like” $\mathbb{R}^n$, in terms of a co-ordinate chart mapping. For any P of M we define (U, f) and $f(P)$ is considered with coordinates $(x_1(P), x_2(P), \dots, x_n(P))$.

Given two charts (U, f) , (V, g) and (V, g) then: $g \circ f^{-1} : \mathbb{R}^n \rightarrow \mathbb{R}^n$ with

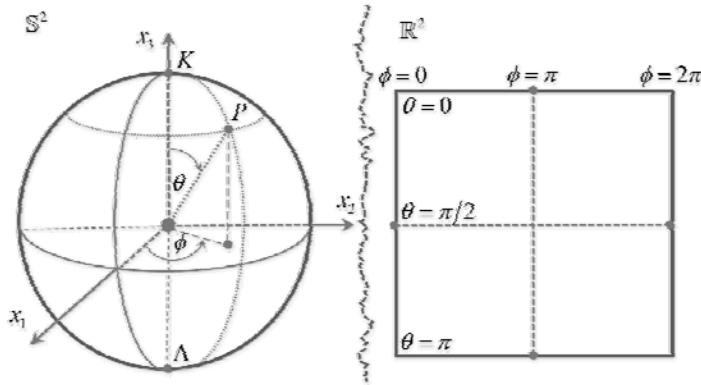
$$(x_1, x_2, \dots, x_n) \mapsto (y_1, y_2, \dots, y_n), \quad y_i = y_i(x_1, x_2, \dots, x_n), \quad i = 1, 2, \dots, n$$

Example 3. The sphere S^2 is a manifold. Indeed: Consider $S^2 : x_1^2 + x_2^2 + x_3^2 = \text{const.}$ and

$$\forall P \in S^2 \quad \exists N_p(r) : N_p(r) \xrightarrow[\text{onto}]{1-1} D \subseteq \mathbb{R}^2.$$

To avoid singular points we restrict angles θ, ϕ to $(\theta, \phi) \in (p, \pi) \times (0, 2\pi)$, Figure 1.

Figure 1. Coordinate mapping of the sphere, $S^2 \rightarrow \mathbb{R}^2$ ($\phi = 0$: equator).



Example 4. An affine space L in $\mathbb{R}^2$ is defined by the mean value of two elements x_1, x_2 . Any two points such that $x_1 + x_2 = 2\bar{x}$ or $x_2 = 2\bar{x} - x_1$ define the line L_1 , Figure 2. If we add a constant h to both values we are moved to line L_2 as

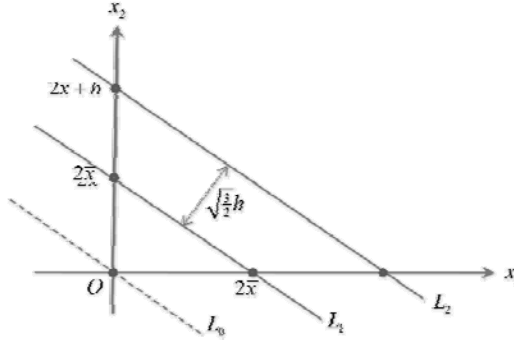
$$\frac{1}{2}[(x_1 + h) + (x_2 + h)] = \bar{x} + h$$

Notice that the parallel line to orbit L_1 passing through the origin O is a vector space, while $L_1, L_2, \dots$ are not. For the orbit L_2 holds $L_2 = L_1 + \text{const} * L_1$, see also Example 5 below.

Now, let us consider two curves passing from the point $P \in M$ of the form

$$x_i = x_i(q), \quad x_i = x_i(r), \quad i = 1, 2, \dots, n.$$

Figure 2. L_0 : vector space, $L_1: \frac{1}{2}(x_1 + x_2) = \bar{x}$, $L_2: L_1 + \sqrt{3/2}h$.



Then the Lie bracket is defined to be

$$\left[\frac{d}{dq}, \frac{d}{dr} \right] = \frac{d}{dq} \frac{d}{dr} - \frac{d}{dr} \frac{d}{dq}.$$

Moreover the set $B = \{\partial/\partial x_i\}$, $i = 1, 2, \dots, n$ forms a base to the vector space created from the directional derivatives along a curve. Consider the set of derivatives along a curve at point P , say $\mathcal{D}_p$, and the vector space of all tangent vectors at P , say $\mathcal{T}_p$. Then there exist a map of $\mathcal{T}_p$ to $\mathcal{D}_p$.

The bracket notation applied to Statistics by Fraser (1968) as in the following.

Example 5. Consider the affine transformation $[\delta, k]$, such that

$$[\delta, k]x = \delta \mathbf{1} + kx, \quad x, \mathbf{1} \in \mathbb{R}^n, \quad k \neq 0.$$

Then a group G of transformations can be formed with elements: identity: $[0, \mathbf{1}]$; inverse: $[\delta, k]^{-1} = [-k^{-1}\delta, k^{-1}]$; composition: $[\Delta, K][\delta, k] = \Delta + k\delta + Kkx$.

So, each realization x , with error e from $f(e)de$, measuring the unknown value θ , is

$$x = [\theta, \mathbf{1}]e \quad \text{With error distribution} \quad \prod_{i=1}^n f(e_i) \prod_{i=1}^n de_i.$$

An orbit is then defined from the group of transformations G , can be for $x, \mathbf{1} \in \mathbb{R}^n$ of the form

$$Gx = [\theta, \mathbf{1}]x = \theta \mathbf{1} + x = \{\theta \mathbf{1}\} + x = G\theta + x.$$

So the affine character has been described. Based on these affine characteristics, Fraser (1968) introduces the location variable $r(x)$, which leads to the “reference” point $d(x)$ on the defined orbit, and came across some nice results of the form:

$$[\bar{x}, s_{\bar{x}}]^{-1} x = \left(\frac{x_1 - \bar{x}}{s_x}, \dots, \frac{x_n - \bar{x}}{s_x} \right) := d(x).$$

For the measurement model with normal error (from $\mathcal{N}(\mu, \sigma^2)$) the reduced model is

$$[\bar{x}, s_{\bar{x}}]^{-1} = [\mu, \sigma][\bar{e}, s_e] \Leftrightarrow \begin{cases} \bar{x} = \mu + \sigma \bar{e}, \\ s_{\bar{x}} = \sigma s_e, \end{cases}$$

with $\bar{e} = Z/\sqrt{n}$, $s_e = x_{n-1}/\sqrt{n-1}$, $Z \sim \mathcal{N}(0,1)$, $x_{n-1}^2 \sim \chi_{n-1}^2$.

This affine geometry point of view is essential to define invariance, as the orbits are defined through a group of transformations G , such that $Gx = G\mathbf{0} + x$. Some of these ideas are applied to the next sections.

2. INVARIANT STATISTICAL APPLICATIONS

Among the statistical applications with no mathematical justification there is a very nice and practical one. Hanes et al. (2007) consider the logistic dose response model

$$\text{logit}(p) = \beta_0 + \beta_1 d_1 + \beta_2 d_2,$$

with $d_1, d_2 \geq 0$. This application is one of the few considering an extension of the simple logit model. To overpass the hidden difficulties for the chosen model, Hanes et al. (2007) simply replace $\beta_i d_i$, $i=1,2$ with z_i , $i=1,2$ and they came across to known results, evaluating Fisher's information,

$$I(\theta) = \mathbf{v}\mathbf{v}^T, \text{ with } \mathbf{v} = \left[\frac{\exp(\theta \mathbf{z}^T)}{1 + \exp(\theta \mathbf{z}^T)} \right]^{1/2} \mathbf{z} \quad (2.1)$$

and $\theta = (\beta_0, 1, 1)$, $\mathbf{z} = (1, z_1, z_2)$. Relation (2.1) is in accordance with the theoretical result: for the Logit and Probit model (and not only), there is always a vector w such that $I(\theta) = \mathbf{w}\mathbf{w}^T$, see Kitsos (1986). In Kitsos (2011) the author worked on the investigation of the background of the problem. The transformation of the vector β to vector θ and $\mathbf{d} = (1, d_1, d_2)$ to $\mathbf{z} = (1, z_1, z_2)$ was the idea to work with. Here are the main steps. For the Logit transformation of the form

$$y = \log \left\{ \frac{p(x)}{1 - p(x)} \right\} = \beta_1 + \beta_1 x_2 + \dots + \beta_p d_p + \sigma e, \quad e \sim \mathcal{N}(0,1), \quad (2.2)$$

with the group (under matrix multiplication) of the affine transformations G formed as

$$G = \left\{ g = \begin{pmatrix} 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & 0 \\ \beta_1 & \beta_2 & \dots & \beta_p & \sigma \end{pmatrix} = \begin{pmatrix} \mathbf{I}_p & \mathbf{O}^T \\ \boldsymbol{\beta} & \sigma \end{pmatrix}, g \in \mathbb{R}^{p+1} \right\}, \quad (2.3)$$

Then we denote

$$\mathbf{I}_p = \text{diag}(1, \dots, 1) \in \mathbb{R}^{p \times p}, \quad \mathbf{O} = (0, \dots, 0) \in \mathbb{R}^{1 \times p}, \quad \sigma \in \mathbb{R}^+.$$

It can be proved that $\exists g^{-1} : gg^{-1} = g^{-1}g = \text{id.}$ with

$$g^{-1} = \begin{pmatrix} \mathbf{I}_p & \mathbf{O}^T \\ \sigma^{-1}\boldsymbol{\beta} & \sigma \end{pmatrix}. \quad (2.4)$$

We need a transformation of model (2.2) to move on a set where affine transformation is possible, see Figure 3. Therefore (2.2) can be written technically, and this is the key idea, as

$$\begin{pmatrix} x'_1 \\ x'_2 \\ \vdots \\ x'_p \\ y' \end{pmatrix} = \begin{pmatrix} 1 & 0 & \dots & 0 & 0 \\ 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & 0 \\ \beta_1 & \beta_2 & \dots & \beta_p & \sigma \end{pmatrix} \begin{pmatrix} x'_1 \\ x'_2 \\ \vdots \\ x'_p \\ e' \end{pmatrix}, \text{ or} \quad (2.5)$$

$$\Psi = gE, \quad (2.6)$$

with $g \in G$ and the definition of Ψ and E obvious. Therefore, the orbit $G\Psi$ is

$$G\Psi = \left\{ \sum_{i=1}^p \beta_i x_i + \sigma e, \quad \sigma \in \mathbb{R}^+ \right\}.$$

From the affine geometry point of view it is known that for “another model” ω (practically another hyper plane or another orbit), then

$$G\Psi \cap G\omega = \emptyset, \text{ or } G\Psi = G\omega. \quad (2.7)$$

That hyperplanes $G\Psi$ and $G\omega$ are either identical or disjoint (we are referring to a completely another situation). Notice that, as neither $\beta_i \geq 0$, $i = 1, 2, \dots, p$, nor $\sum_{i=1}^p \beta_i = 1$, the set $G\Psi$ cannot be considered as an affine simplex, but expands a p -dimensional hyperplane.

Under this theoretical framework, the following can be proved, see Kitsos (2011).

Theorem 2.1. Consider any positive definite matrix $\mathbf{M}$ and a vector $\mathbf{c}$ with appropriate dimensions such that $\mathbf{N} := \mathbf{c}^+ \mathbf{M}^{-1} \mathbf{c}$ is valid. Then $\mathbf{N}$ remains invariant if $\mathbf{c}$ and $\mathbf{M}$ transformed under g .

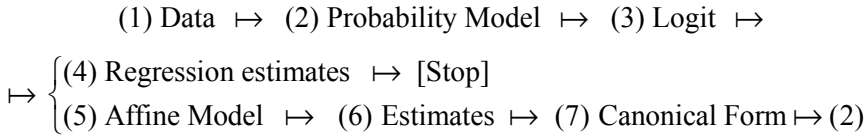
Corollary 2.1. If $\mathbf{M} = \mathbf{M}(\theta, \xi)$ is the average per observation information matrix, with ξ being a design measure, then $\mathbf{N} = \mathbf{N}(\theta, \xi)$ is the transformed form of $\mathbf{M}(\theta, \xi)$. Moreover, if $p = 2$, the co-optimal design remains invariant under g .

The above Theorem 2.1 provides evidence that a transformation g exists as far as the information matrix is concerned. That is, the D-optimal design criteria remain invariant under g transformation. Still we need to prove that $g \in G$ can be transformed, so that the parameter $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_p)$ can be $\boldsymbol{\theta} = (\beta_1, 1, 1, \dots, 1)$.

Theorem 2.2. There is a transformation $g \in G$ as in (2.3), with element $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_p)$ with parameter vector equal to $\boldsymbol{\theta} = (\beta_1, 1, 1, \dots, 1)$.

Proof. See Kitsos (2011).

This result is the basic one where Hanes et al. (2007) were really correct when arbitrary used their transformation with no further explanation. Moreover, this transformation to $\boldsymbol{\theta} = (\beta_1, 1, 1, \dots, 1) \in \mathbb{R}^p$ provides evidence that always exist the so-called “canonical form”, Kitsos (1986, 2006, 2011), Ford et al. (1992), of the Logit model. Figure 3 explains the involved transformations to binary response data set, so that to obtain the canonical form. The flow chart of the procedure is the following:



3. INVARIANCE ON A CYCLE

Recall that the manifold of all rotations $\text{SO}(3)$ is different from the manifold whose coordinate system is rotated $\mathbb{E}^3$. For the orthogonal group in n -dimensions $\text{O}(n)$ for the matrix $\mathbf{A} \in \text{O}(n) \Rightarrow \det \mathbf{A} = \pm 1$. Moreover, there exists $\mathbf{B} \in \text{SO}(n)$ such that the canonical form $\mathbf{B}^{-1} \mathbf{A} \mathbf{B}$ can be evaluated. The matrix $\mathbf{B}$ consist of a block with elements $+1$, -1 , or submatrix $\mathbf{Q} = (Q_{ij})$, $Q_{11} = Q_{22} = \cos \theta$, $Q_{21} = -Q_{12} = -\sin \theta$. We work on a cycle as in Figure 4.

Let the measurement $\omega = (\omega_1, \omega_2)$ of an angle with error $e = (e_1, e_2)$, with $|e_i| \leq 1$, $i = 1, 2$. Then the group of transformation G is $\text{O}(2)$, i.e.

$$\text{O}(2) = G = \{g = \mathbf{Q}, \theta \in [0, 2\pi]\} \Leftrightarrow e = g^{-1} \omega. \quad (3.1)$$

For multiple observations $\boldsymbol{\Omega} = (\omega_{ij})$, $i = 1, 2, j = 1, 2, \dots, n$, and errors $\mathbf{E} = (e_{ij})$, the general affine measurement model (recall (2.6)) is

$$\Omega = gE, \quad \prod_{i=1}^n f(e_i) \prod_{i=1}^n de_i. \quad (3.2)$$

For the circle, a Normal error distribution is adopted, i.e.

$$f(e; \lambda) = \frac{1}{2\pi B_0'(\lambda)} \exp\{\lambda \cos e\} de,$$

with B_0' being the imaginary Bessel function of zero order. When $\lambda = 0$, the distribution is uniform on the circle. This measurement model still needs more investigation and extension to the sphere.

Figure 3. The transformation to Canonical Form of the Logit model.

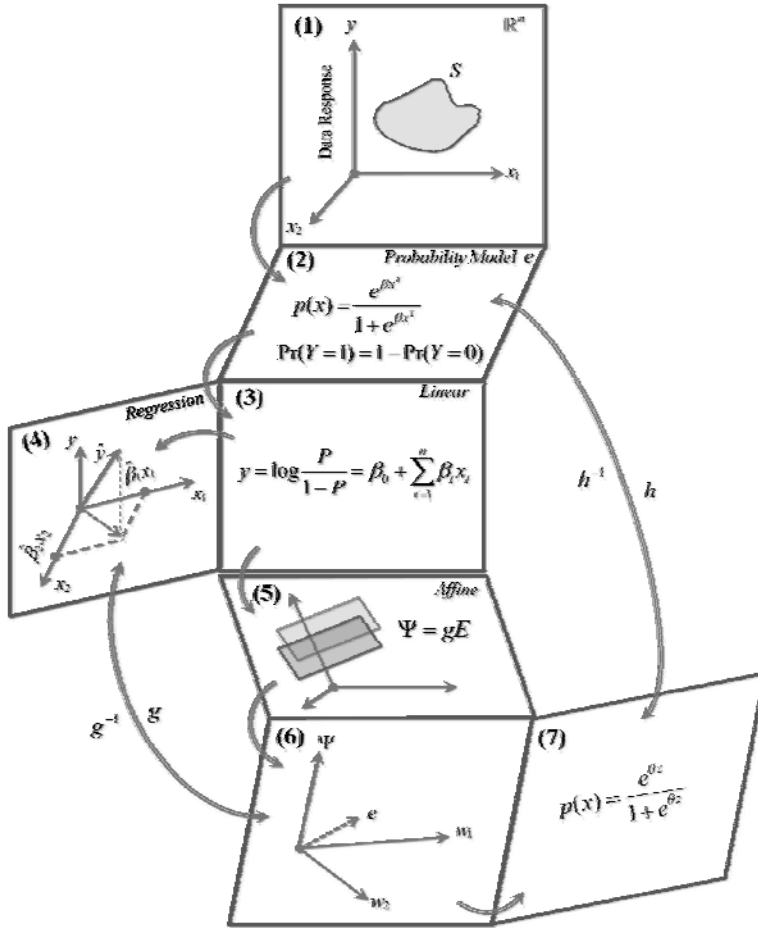
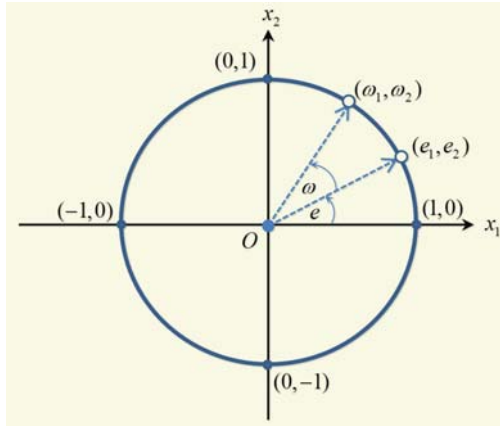


Figure 4. Measurement on a cycle.



This approach works well to a number of cases and still needs more investigation both on the cycle and on the sphere. It works well on tolerance regions (hyper-ellipse), Muller and Kitsos (2004) and the design criteria defined through tolerance regions. Thus there is some work towards this direction, but certainly needs more investigation, part of our future work.

ΠΕΡΙΛΗΨΗ

Ο στόχος αυτής της εργασίας είναι η χρησιμοποίηση της Ομοπαράλληλικής Γεωμετρίας (Affine Geometry) για να αποδείξει ιδιότητες αναλλοίωτων μετασχηματισμών στη Στατιστική. Ορισμένες Μαθηματικές έννοιες εισάγονται και η εφαρμογή τους στην Στατιστική συζητείται.

REFERENCES

- Bechger T.M., Verhelst N.D. and Verstrales H.H.F.M. (2001). Identifiability of nonlinear logistic test models. *Psyciometrika*, 68(3), 357-372.
- Blyth C.R. (1970). On the inference and decision models of Statistics. *Ann. Math. Stat.*, **41**, 1034-1058.
- Ford I., Torsney B. and Wu C.F.J. (1992). The use of canonical form in the construction of locally optimal designs for non-linear problems, *JRSS*, **54**(2), 569-583.
- Fraser D.A.S. (1968). *The Structure of Inference*, Wiley.
- Graybill F.A.(1976). *Theory and Application of the Linear Model*, Duxbury Press.
- Haines M.L., Kabera G., Ndlovu P. and O'Brien E.T. (2007). D-optimal designs for logistic regression in two variables. In *Advances in Model-Oriented Design and*

- Analysis* (MODA8), Diaz J.M. and Torsney B. (Eds.), Heidelberg, Physica-Verlag, 91-98.
- Kitsos C.P. (2011). Invariant canonical form for the multiple logistic regression. *Mathematics in Engineering, Science and Aerospace* (MESA), **38**(1), 267-275.
- Kitsos C.P. (2006). On the logit methods for Ca problems. In: *Statistical Methods for Biomedical and Technical Systems*, Limassol, Cyprus, 335-340.
- Kitsos C.P. (2005). Optimal designs for bioassays in carcinogenesis. In: *Quantitative Methods for Cancer and Human Health Risk Assessment*, Edler L. and Kitsos C.P. (Eds), John Wiley, England, 267-279.
- Kitsos, C.P. (2001). Design aspects for the Michaelis–Menten model. *Biometrical Letters*, **38**, 53-66.
- Kitsos, C.P. (1986). *Design and Inference in Nonlinear Problems*, PhD Thesis, University of Glasgow, UK.
- Muller C.H. and Kitsos C.P. (2004). Optimal design criteria based on tolerance regions. In: *Advances in Model-Oriented Design and Analysis* (MODA7), Di Bucchianno A., Lauter H. and Wynn H. (Eds), Physica-Verlag, 107-115.



STATISTICAL INFERENCE FOR MULTI-STATE SYSTEMS

A. Makrides¹, A. Karagrigoriou²

¹Department of Mathematics and Statistics, University of Cyprus
andreasmaths@hotmail.com

²Division of Statistics & Actuarial - Financial Math., Univ. of the Aegean, Greece
alex@ucy.ac.cy

ABSTRACT

Markov processes are widely used for reliability analysis because the number of failures in arbitrary time intervals in many practical cases can be described as a Poisson process and the time up to the failure and repair time are often exponentially distributed. In this work we focus on the estimation of both the intensity rates and transition probabilities via output performance observations in non-homogeneous systems, using the Weibull class of distributions.

Keywords: Multi-state system, Reliability theory, Weibull, transition intensities, transition probabilities.

1. INTRODUCTION

All systems are designed to perform their intended tasks in a given environment. Most researchers deal with traditional binary reliability models allowing for only two possible states for a system and its components: perfect functionality and complete failure. However, many real-world systems are composed of multi-state components, which have different performance levels and several failure modes with various effects on the entire system performance. A system with finite number of performance levels is called a multi-state system (MSS). The basic concepts of MSS reliability were introduced in the 70's and the 80's (Murchland, 1975; Natvig, 1982 and 1985). Readers interested in the history of MSS reliability theory are referred to the work of Lisnianski and Levitin (2003) and Natvig (2011).

The entire set of possible system states can be divided into two disjoint subsets corresponding to acceptable and unacceptable system functioning. The system entrance into the subset of unacceptable states constitutes a failure. MSS reliability can be defined as the system's ability to remain in acceptable states during the operation period.

The MSSs are examined via Markov processes. Markov processes are widely used for reliability analysis because the number of failures in arbitrary time intervals in many practical cases can be described as a Poisson process and the time up to the failure and repair time are often exponentially distributed. One of the ways to investigate processes with arbitrarily distributed sojourn times is to use a semi-Markov process model (Levy, 1954). Indeed, in such a situation, other distributions may be considered, which in some cases, describe more appropriately the time between failures. In such a setting the appropriate transition matrix of the *embedded jump chain* and the distribution function for the sojourn times should be investigated since *together characterize uniquely*, the semi-Markov process. The intensity rate (or transition intensity), which is described as the instantaneous transition probability (known as failure rate or in survival analysis, as hazard rate or function) is of primary importance in identifying the ideal distribution of the failure time. The intensity rate is directly related to the parameters of the underlying distribution since it depends on the probability density function. The set of intensity rates is used to evaluate the corresponding transition probabilities between states. Our intention in this work is the estimation of both the intensity rates and transition probabilities via output performance observations in non-homogeneous systems using various distributions. Such distributions include the Weibull, Pareto, and Rayleigh which are frequently encountered in reliability and technical problems. Here, we explore the Weibull class of distributions.

This paper is organized as follows. Section 2 is devoted to statistical concepts of intensity rates and transition probabilities associated with various distributions of sojourn times. More specifically, in Section 3, the one-step transition probabilities $Q_{ik}(t)$ between states i and k and the resulting kernel matrix are obtained and the parameter estimators of the distribution of the sojourn times are furnished for the case of the Weibull class of distributions. An illustrative example is provided in Section 4. The paper ends with a discussion and concluding remarks and comments.

2. SOJOURN TIMES, INTENSITIES, TRANSITION PROBABILITIES

Consider a discrete-state continuous-time not necessarily homogeneous process with N possible states $\{1, 2, \dots, N\}$ where usually the last state " N " is associated with nominal performance of the system and the first state " 1 " is associated with total failure. Let also T be the total observation time. The quantity of interest in such settings is the transition probability of moving from state i to state j . We define the following quantities:

T_{ij} : time spent at state i before moving (directly) to state j .

k_i : accumulated number of system exits from state i to any other state during observation time T .

k_{ij} : accumulated number of system transitions from state i to state j during observation time T .

$T_i^{(m)}$: system's sojourn time of the m^{th} residing in state i during observation time T .

Let $F_{ij}(t)$ be the cumulative distribution function (cdf) of sojourn time T_{ij} in state i when the unit transits from state i to state j (Limnios and Oprisan, 2001). Let also $f_{ij}(t)$ be the corresponding probability density function of T_{ij} and θ be the m -dimensional parameter involved in the underlying distribution. The quantity of interest in the present setting is the intensity rate (or failure rate or hazard rate) which is defined by

$$\lambda_{ij}(t) = \frac{f_{ij}(t)}{1 - F_{ij}(t)}$$

where $R_{ij}(t) = 1 - F_{ij}(t)$ represents the reliability function. Naturally, the intensity rate depends on the unknown parameter θ . As a result, in order to construct estimators of intensity rates, appropriate estimates of the parameter θ are required.

If the MSS is initially in state i at time instant $t = 0$, the probability $Q_{ik}(t)$ that the unit will transit (first transit) from state i to a fixed (specific) state k before time t is called a one-step transition probability

$$Q_{ik}(t) = \Pr(X_t = k \mid X_0 = i) \quad (1)$$

In other words, $Q_{ik}(t)$ represents the probability that the *first* transit from state i will be in state k before time t conditioning on the fact that the unit is in state i at initial time instant $t = 0$. One can obtain the one-step probability $Q_{ik}(t)$, with k fixed, as the probability that under condition $T_{ik} \leq t$ the random variable T_{ik} is the minimum of the random variables T_{ij} , $j = 1, \dots, N$. In other words, if $T_{i(m)}$ represents the m^{th} order statistic, $m = 1, \dots, N$ of the N random variables $T_{i1}, \dots, T_{iN}$, then $T_{i(1)} = T_{ik}$. Hence, for each $i = 1, \dots, N$,

$$\begin{aligned} Q_{ik}(t) &= P\{T_{ik} \leq t, T_{ik} = \min\} = P\{T_{ik} \leq t, T_{ik} \leq T_{il}, \forall l \neq k\} \\ &= \int_0^t [1 - F_{i1}(u)] \dots [1 - F_{i,k-1}(u)] \times [1 - F_{i,k+1}(u)] \dots [1 - F_{iN}(u)] dF_{ik}. \quad (2) \end{aligned}$$

Then, we consider the distribution $F_i(t)$ of the time till the first exit from state i irrespectively of the state to which the process arrives. For this purpose we take the summation of Q_{ik} for all k :

$$F_i(t) = \sum_{k=1}^N Q_{ik}(t). \quad (3)$$

Assume now that the process moves from i to j with $j \neq i$, namely $Q_{ii}(1) = 0$. Then, T_{ik} is the minimum of the $N-1$ random variables $T_{i1}, \dots, T_{i,i-1}, T_{i,i+1}, \dots, T_{iN}$, namely $T_{ik} = \min\{T_{i1}, \dots, T_{i,i-1}, T_{i,i+1}, \dots, T_{iN}\}$. Notice that in such a case, both $Q_{ik}(t)$ and $F_i(t)$ given in (2) and (3) involve $N-1$ instead of N terms.

In our attempt to determine estimators for the parameter θ and consequently of the intensity rates $\lambda_{ij}(t)$, we will be using the steady-state probabilities π_{ik} which represent the fraction of time the system, in the long run (long-run behavior), spends in state k , $\pi_{ik} = \lim_{t \rightarrow \infty} Q_{ik}(t)$. The steady state probabilities π_{ik} can be naturally estimated as the ratio of the corresponding number of transitions,

$$\hat{\pi}_{ik} = \frac{k_{ik}}{k_i} \quad (4)$$

where k_{ik} is the accumulated number of system transitions from state i to state $k \neq i$ during observation time T , and k_i is the accumulated number of system exits from state i to any other state during observation time T . The determination of the required transition probabilities is based on the intensity matrix $Q = (\lambda_{ij})_{i,j=1,\dots,N}$ or even better, on the estimator $\hat{Q} = (\hat{\lambda}_{ij})_{i,j=1,\dots,N}$ of the intensity matrix Q .

More specifically, the eigenvalues γ_i of the estimator of the matrix Q , are obtained and then, the corresponding eigenvectors are evaluated. If L is the matrix of eigenvectors and $E(t)$ a $N \times N$ diagonal matrix with (i, i) element equal to $\exp(-\gamma_i t)$ then the transition probabilities are given by the matrix $L \cdot E(t) \cdot L^{-1}$ where L^{-1} denotes the inverse of L .

Note that the case of constant intensity rates (i.e. exponential distribution), is associated with time-homogeneity. Hence, the origin $t = 0$ of the time coordinate can be located at any instant, and the equations remain valid for any time t so that $\lambda_{ij}(t) = \lambda$, $\forall t$. On the other hand, if a transition rate is a non-constant function of time, the equations will be valid only for one particular choice of the origin $t = 0$, which must be the instant when all the components of a system are new, so that the time parameter t can be identified with the age of the components. It is possible to contrive homogeneous Markov models that simulate systems with variable transition rates. Such an issue will be the object of a future project.

3. THE WEIBULL CLASS OF DISTRIBUTIONS

If the sojourn time T_{ij} follows the Weibull distribution then the cdf is given by

$$F_{ij}(t) = 1 - e^{-(t/a_{ij})^c} \quad (5)$$

where the scale parameter is denoted by a_{ij} and the shape parameter is denoted by c and it is assumed to be constant, irrespectively of the states i and j . The case of a shape parameter that depends on the states i and j is avoided here, not only because it increases considerably the computational burden but also due to limited number of transitions between each pair of states that may lead to unreliable and unstable estimators. Note that the exponential and Rayleigh distributions are special cases of (5) if the constant shape parameter is taken to be equal to 1 and 2 respectively. The intensity rate of the sojourn time for the case of the Weibull distribution associated with (5) is given by

$$\lambda_{ij}(t) = \frac{c}{a_{ij}} \left(\frac{t}{a_{ij}} \right)^{c-1}. \quad (6)$$

Note that in the case of the exponential distribution, the intensity rate is constant and equal to $1/a_{ij}$.

Following the methodology of the previous section for the case of the Weibull distribution, we have that

$$Q_{ik}(t) = \int_0^t \frac{c}{ca_{ik}^c} \exp \left\{ -x^c \sum_{j=1}^N \frac{1}{a_{ij}^c} \right\} dx^c = \frac{1}{a_{ik}^c \sum_{j=1}^N \frac{1}{a_{ij}^c}} \left[1 - e^{-t^c \sum_{j=1}^N \frac{1}{a_{ij}^c}} \right]. \quad (7)$$

As a result, using (3) we have

$$F_i(t) = 1 - e^{-t^c / \left(\sum_{j=1}^N \frac{1}{a_{ij}^c} \right)^{-1/c}} \quad (8)$$

which represents a Weibull distribution function with parameters c and $\left(\sum_{j=1}^N \frac{1}{a_{ij}^c} \right)^{-1/c}$

and with mean given by

$$\mu_i = \left(\sum_{j=1}^N \frac{1}{a_{ij}^c} \right)^{-1/c} \cdot \Gamma \left(1 + \frac{1}{c} \right)$$

where $\Gamma(\cdot)$ is the standard Gamma function.

In other words, the above quantity $F_i(t)$ represents the distribution of the time till the first exit from state i irrespectively of the state to which the process arrives, for the case of the Weibull distribution.

For the estimation of the mean $\mu_i, \forall i$ we take the standard moment estimator given by $\hat{\mu}_i = \sum_{j=1}^{k_i} T_i^{(j)} / k_i$. Thus,

$$\sum_{j=1}^N \frac{1}{a_{ij}^c} = \left[\frac{\Gamma\left(1 + \frac{1}{c}\right) k_i}{\sum_{j=1}^{k_i} T_i^{(j)}} \right]^c. \quad (9)$$

Using expression (9) one can estimate the summation of the scale parameters for all transitions that exit from state i . Although the resulting estimator is of some interest, the estimation of the individual scale parameters a_{ij} for each $\{i, j\}$ is of greater importance. In order to estimate individual scale parameters, we make use of the long-run (steady-state) probabilities, in the following way,

$$\pi_{ik} = \lim_{t \rightarrow \infty} Q_{ik}(t) = \lim_{t \rightarrow \infty} \frac{1}{a_{ik}^c \sum_{j=1}^N \frac{1}{a_{ij}^c}} \left[1 - e^{-t^c \sum_{j=1}^N \frac{1}{a_{ij}^c}} \right] = \frac{1}{a_{ik}^c \sum_{j=1}^N \frac{1}{a_{ij}^c}} \quad (10)$$

which, in turn, gives

$$a_{ik}^c = \frac{1}{\pi_{ik} \sum_{j=1}^N \frac{1}{a_{ij}^c}}. \quad (11)$$

Substituting (4) and (9) into (11) the following estimate is obtained for the shape parameter a_{ik} with $i, k = 1, \dots, N$ with $i \neq k$:

$$\hat{a}_{ik} = \frac{1}{\left(\frac{k_{ik}}{k_i} \right)^{1/c} \left[\frac{\Gamma\left(1 + \frac{1}{c}\right) k_i}{\sum_{j=1}^{k_i} T_i^{(j)}} \right]}, \quad (12)$$

and using (6), the estimate for the intensity rate between states i and k is

$$\hat{\lambda}_{ik}(t) = \frac{c}{\hat{a}_{ik}} \left(\frac{t}{\hat{a}_{ik}} \right)^{c-1}. \quad (13)$$

Note that since $\sum_{k=1}^N \lambda_{ik}(t) = 0, \forall i$, then $\lambda_{ii}(t) = -\sum_{k \neq i} \lambda_{ik}(t)$. Similarly,

$$\hat{\lambda}_{ii}(t) = -\sum_{k \neq i} \hat{\lambda}_{ik}(t).$$

If c is given, this value is used in (13). Otherwise, an appropriate least squares estimator is used instead. One such estimator of c is given by

$$\hat{c} = \sum \log(T_{ij}) \log \left(\log \left(\frac{1}{1 - \hat{F}(T_{ij})} \right) \right) - \frac{n\bar{X} \cdot \bar{Y}}{\sum T_{ij}^2 - n\bar{X}^2} \quad (14)$$

where n is the sample size, T_{ij} is the sojourn time,

$$\bar{X} = \frac{1}{n} \sum \log(T_{ij}) \quad \text{and} \quad \bar{Y} = \sum \log \left(\frac{1}{1 - \hat{F}(T_{ij})} \right),$$

with the usual empirical estimator used to estimate the distribution function.

Special cases of the Weibull distribution are the exponential and Rayleigh distributions. For the exponential distribution it is easy to see that the estimator of the (i,j) transition rate is

$$\hat{\lambda}_{ik}^{exp} = \frac{1}{\hat{a}_{ik}} = \left(\frac{\sum_{j=1}^{k_i} T_i^{(j)}}{k_{ik} \Gamma(2)} \right)^{-1} = \frac{k_{ik}}{\sum_{j=1}^{k_i} T_i^{(j)}} \equiv \hat{\mu}_i^{-1} \quad (15)$$

which coincides with the reciprocal of the mean moment estimator.

Consider now the Rayleigh distribution, the cumulative distribution of which is given by $F_{ij}(t) = 1 - e^{-t^2/2a_{ij}^2}$ with intensity rate given by

$$\lambda_{ij}(t) = \frac{2}{\sqrt{2}a_{ij}} \left(\frac{t}{\sqrt{2}a_{ij}} \right). \quad (17)$$

In this case, with $c = 2$ and $\sqrt{2}a_{ik}$ in place of a_{ik} in (7), we have for $i, k = 1, \dots, N$ with $i \neq k$, that the estimator of the intensity rate (17) is obtained analogously as in (13), namely

$$\hat{\lambda}_{ik}(t) = \frac{t}{\hat{a}_{ik}^2}.$$

4. ILLUSTRATIVE EXAMPLE

In this section for illustrative purposes we consider a setting with four (4) states. More specifically, a sample of size $n = 295$ has been randomly selected such that an exponential distribution has been used for all transitions except the transition from

state "4" to state "3" for which a Weibull distribution was used with shape parameter $c = 0.6$. The intensity rates associated with the true model are given by the following (true) intensity rate matrix:

$$Q = \begin{pmatrix} -0.065 & 0 & 0 & 0.065 \\ 0.024 & -0.11 & 0 & 0.086 \\ 0.022 & 0 & -0.11 & 0.098 \\ 0.003 & 0.006 & 0.125 & -0.134 \end{pmatrix}.$$

Using the formulas presented in Section 3 we obtain the required matrix $\hat{Q}$ and obtain the transition probabilities according to (13) and (15). The results for $t = 10$ and $t = 360$ in Tables 4.1 and 4.2 refer to the case where all transition probabilities (even the one from state "4" to state "3") are computed with the Exponential formula (eq. (15)). In Tables 4.3 and 4.4 all transition probabilities are based on the Weibull distribution (eq. (13)) with c estimated by (14). Finally, in Tables 4.5 and 4.6 we provide the results associated with the correct model, namely where all transition probabilities are computed with the exponential formula except the case from state "4" to state "3" which is computed with the Weibull formula.

Note that as time progresses, i.e as t grows, the transition probabilities start to stabilize. As it is clear from Tables 4.1 and 4.3, even for $t = 10$ the transition from state i to state 4 is more likely for any state i . When $t = 360$ (1 year) we observe (Tables 4.2 and 4.4) that the probabilities are already stable and indeed, state 4 is the most likely state with state 2 to be the most unlikely for the system to arrive. On the other hand, by Tables 4.5 and 4.6 (which correspond to the true model) we observe that the conclusions based on Tables 4.1 - 4.4 (which correspond to approximations of the true model) are only partially correct. Indeed, the state 2 is obviously the less likely to occur but states 3 and 4 are almost equally possible to occur with probabilities around 35%.

State	1	2	3	4
1	0.5609	0.0050	0.0839	0.3502
2	0.0779	0.4038	0.1024	0.4159
3	0.1255	0.0072	0.4003	0.4675
4	0.0518	0.0166	0.2633	0.6685

Table 4.1: Exponential Distribution, $t = 10$.

State	1	2	3	4
1	0.1483	0.0203	0.2727	0.5663
2	0.1483	0.0203	0.2728	0.5664
3	0.1485	0.0203	0.2729	0.5667
4	0.1484	0.0203	0.2728	0.5665

Table 4.2: Exponential Distribution, $t = 360$.

State	1	2	3	4
1	0.5305	0.0080	0.0838	0.3779
2	0.3573	0.1950	0.0807	0.3672
3	0.3900	0.0096	0.1731	0.4275
4	0.2006	0.0196	0.1857	0.5946

Table 4.3: Weibull Distribution, $t = 10$.

State	1	2	3	4
1	0.3488	0.0174	0.1486	0.4972
2	0.3488	0.0174	0.1486	0.4972
3	0.3488	0.0174	0.1486	0.4972
4	0.3489	0.0174	0.1487	0.4974

Table 4.4: Weibull Distribution, $t = 360$.

State	1	2	3	4
1	0.5927	0.0092	0.1428	0.2552
2	0.1369	0.3399	0.1949	0.3288
3	0.1675	0.0122	0.5011	0.3192
4	0.0983	0.0287	0.4407	0.4323

Table 4.5: Exp & Weibull Distributions, $t = 10$.

State	1	2	3	4
1	0.2488	0.0255	0.3835	0.3422
2	0.2490	0.0255	0.3835	0.3424
3	0.2488	0.0255	0.3835	0.3422
4	0.2488	0.0255	0.3835	0.3422

Table 4.6: Exp & Weibull Distributions, $t = 360$.

Finally, we must indicate that the matrix $\hat{Q}$ associated with the true model (Tables 4.5 and 4.6) is the one which is close to the true matrix Q . In what follows the estimates of Q are provided for each of the three cases examined:

$$\hat{Q}_{12} = \begin{pmatrix} -0.0599 & 0 & 0 & 0.0599 \\ 0.0135 & -0.0022 & 0 & 0.0787 \\ 0.0270 & 0 & -0.1268 & 0.0998 \\ 0.0022 & 0.0033 & 0.0611 & -0.0666 \end{pmatrix}$$

$$\hat{Q}_{34} = \begin{pmatrix} -0.0748 & 0 & 0 & 0.0748 \\ 0.0981 & -0.1675 & 0 & 0.0694 \\ 0.1483 & 0 & -0.2611 & 0.1128 \\ 0.0048 & 0.0058 & 0.0780 & -0.0886 \end{pmatrix}$$

$$\hat{Q}_{56} = \begin{pmatrix} -0.0599 & 0 & 0 & 0.0599 \\ 0.0245 & -0.1117 & 0 & 0.0872 \\ 0.0321 & 0 & -0.1163 & 0.0839 \\ 0.0025 & 0.0083 & 0.1303 & -0.1411 \end{pmatrix}$$

5. CONCLUSIONS

In this paper we focus on multi-state systems (MSS) with N states. The entire set of possible system states can be divided into two disjoint subsets corresponding to acceptable and unacceptable system functioning. In such cases, the underlying distribution for the sojourn time is of vital importance. This work is devoted to the evaluation of the intensity rates and transition probabilities via output performance observations using the Weibull class of distributions which is frequently encountered in reliability.

ΠΕΡΙΛΗΨΗ

Η εργασία ασχολείται με τη στατιστική συμπερασματολογία σε τεχνικά συστήματα πολλαπλών καταστάσεων (επιπέδων λειτουργίας). Συνήθως ο χρόνος μετάβασης από μια κατάσταση σε άλλη ακολουθεί την εκθετική κατανομή. Στην παρούσα εργασία γίνεται μελέτη της οικογένειας των κατανομών Weibull.

Acknowledgements: The authors wish to express their appreciation to the Editor and an anonymous referee for their valuable and constructive suggestions and comments.

REFERENCES

- Barlow, RE and Proschan, F. (1965). *Mathematical Theory in Reliability*. Wiley, New York.
- Levy, P. (1954). Process semi-markoviens. In: *Proc. of the Intern. Congress on Math.*, Amsterdam, 416-426.
- Limnios, N. and Oprisan, G. (2001). *Semi-Markov Processes and Reliability*, Birkhauser, Boston.
- Lisnianski, A., Frenkel, I. and Ding, Y. (2010). *Multi-state System Reliability Analysis and Optimization for Engineers and Industrial Managers*. Springer, London.
- Lisnianski, A. and Levitin, G. (2003). *Multi-state System Reliability: Assessment, Optimization and Applications*. Singapore, World Scientific.
- Murchland, J. (1975). Fundamental concepts and relations for reliability analysis of Multistate systems. In: Barlow RE, Fussell JB and Singpurwalla N (eds) *Reliability and fault tree analysis: theoretical and applied aspects of system reliability*. SIAM, Philadelphia: 581-618.

- Natvig, B. (1982). Two suggestions of how to define a multistate coherent system. *Adv Appl Probab* **14**: 434–455.
- Natvig, B. (1985). Multi-state coherent systems. In: Johnson N, Kotz S (eds) *Encyclopedia of Statistical Sciences*, vol 5. Wiley, New York: 732–735.
- Natvig, B. (2011). *Multistate Systems Reliability. Theory with Applications*. Wiley, New York.



STATISTICAL MODELING OF SOCIAL NETWORKS: AN APPLICATION TO CO-AUTHORSHIP DATA

G. Oikonomopoulou¹ and D. Karlis²

Department of Statistics,
Athens University of Economics and Business,
Athens, Greece
emails: ¹georgia.oikonomopoulou@gmail.com,
²karlis@aueb.gr

ABSTRACT

This paper discusses the use of statistical models to visualize social network data and to summarize its topological structure. The Latent Position Cluster Model (LPCM) of Handcock et al (2007) is used, in order to position a network in the latent “social space” to represent the available data but also to determine the existence of clusters. The model also assigns the actors in the respective clusters. We apply this model to a real network dataset representing the co-authorship network between faculty members of the Department of Statistics in Athens University of Economics and Business (AUEB) aiming at describing the academic synergies between those members.

Keywords: latent space model; latent position cluster model; co-authorship structure;

1. INTRODUCTION

Social networks represent the interaction among a collection of social entities or “actors”. Usually, social network data consist of a set of n actors and are characterized by a set of binary link variables $y_{i,j}$ measured on each ordered pair of actors $i, j = 1, \dots, n$, indicating the presence or absence or the value of a relationship from actor i to actor j . Such data are modeled by an “interaction graph” in which nodes represent actors and edges represent some kind of interaction between pairs of those actors. Actors are typically unique individuals or groups of people, but sometimes can also be non-human like animals. Applications of network data could be, for example, networks representing the relations of people in clubs or companies, meetings between members of terrorist cells, friendship among school children, sexual contacts within a community, corporate alliances among businesses, email exchanges between individuals, co-authorship on scientific articles, trade agreements among nations, just to name few. Other examples,

concerning cross-citations of 42 economics journals over the years 1995-1997, are given by Ambroise and Matias (2004) as well as Latouche et al (2012) who tried to cluster the French political blogs.

Apart from the main analysis concerning model fitting, the topological structure of the network is studied as a descriptive task. Graphical tools and simple metrics are applied in order to visualize the data and to summarize network's topological structure. Many researchers have proposed different stochastic models in order to analyze social network data, representing the same time, the networks' structural features. In particular, the Latent Position Cluster Model (LPCM) proposed by Handcock et al (2007), is discussed. This model accomplishes to capture network features such as transitivity, homophily on attributes and clustering at the same time and does not require the number of clusters to be known. It extends the Latent Space Model (LSM) of Hoff et al (2002), using the idea of model-based clustering.

The structure of the paper is organised as follows: In section 2 we discuss the LPCM while Section 3 presents an application of the model to the co-authorship network data concerning the faculty members of the Department of Statistics in AUEB. Some remarks can be seen in Section 4.

2. THE LATENT POSITION CLUSTER MODEL

2.1 The Latent Space Model

The LSM dates back to Hoff et al (2002). They suggested an approach to modeling networks based on assuming the existence of a latent (Euclidean usually) space of characteristics of the actors. Actors' relationships are modeled as a function of distances between these characteristics as well as functions of observed covariates.

In particular, the model supposes each actor to have an unobserved position in a 2-dimensional Euclidean latent "social space". In addition, it is assumed that the presence or absence of a tie between two actors is independent of all other ties in the network, given the positions z_i and z_j in the "social space" of two individuals i and j . The model involves locating each actor in a latent "social space" such that actors who are close have a higher probability to form links than those distant.

Data are considered on the relations between n actors. In particular, the observations $y_{i,j}$ are dichotomous, where 1 implies a direct link between i and j actors. The model can be applied for non-binary data or non-symmetric relationships. A covariate information $x_{i,j,k}$, which is the k -th covariate, may be used to describe the relationship between i, j actors. The $n \times n$ sociomatrix is denoted as Y with elements $y_{i,j}$. In this case, the $y_{i,j}$ are binary and the probability that there is a link between i and j , i.e. the probability $P(y_{i,j} = 1)$, is modeled.

The class of the models considered assumes that the conditional probability

of a tie between two actors, given the covariates, depends only on the distance between them in an unobserved “social space”. The model also assumes that ties between actors occur independently, given their distance apart and the covariates.

At the beginning, the discussed model is presented in a general form. It is assumed that

$$P(Y_{i,j} = y_{i,j} | \beta, |z_i - z_j|) = f(y_{i,j} | E(Y_{i,j} | \beta, |z_i - z_j|))$$

i.e. the probability of a link is modeled via some probability function $f(\cdot)$ and the expectation depends on some parameters β as well as the distance between the latent positions z_i and z_j . In order to use the more familiar GLM frame, we assume a typical link function:

$$E(Y_{i,j} | \beta, |z_i - z_j|) = g^{-1}(\eta_{i,j}(\beta, |z_i - z_j|))$$

where

$$\eta_{i,j}(\beta, |z_i - z_j|) = \sum_{k=1}^p \beta_k x_{i,j,k} - \beta_0 |z_i - z_j|$$

is the typical link function in the GLM setting. Since we are working with probabilities a logit link function is assumed. Hence, the probabilities of a tie are modeled using the logistic regression model formula

$$\log odds P(y_{i,j} | Z_i, Z_j, \beta) = \sum_{k=1}^p \beta_k x_{i,j,k} - \beta_0 |z_i - z_j|$$

where $\log odds(p) = \log\left(\frac{p}{1-p}\right)$, $x_{i,j,k}$ is possible covariate information. We also assume for identifiability reasons that the positions have unit root mean square, i.e.

$$\sqrt{\frac{1}{n} \sum_i |z_i|^2} = 1$$

This particular approach assumes an Euclidean distance, however other distances are also applicable. In particular, the usage of the Euclidean space considerably helps on the interpretation of the model. It is noted that under this model the probability of a tie between actors, in an unobserved Euclidean “social space”, depends on the distance between them.

2.2 The Latent Position Cluster Model

An important feature in many networks is the existence of clustering between actors, hence interest often focuses on finding clusters of actors or ties, while the number of groups in the data is typically unknown. As a result, the LPCM that Handcock et al (2007) proposed, is used.

This model is an extended version of the LSM, by incorporating a Gaussian mixture model structure for the latent positions of actors in “social space” to accommodate the clustering of nodes in the network. Clusters are included explicitly in the model rather than found by post-hoc analysis of the network model.

To represent clustering, it is also assumed that the actors’ positions in the d -dimensional space come from a finite mixture of G multivariate normal distributions, each representing a different cluster of actors (Banfield and Raftery, 1993). That is

$$z_i \sim \sum_{g=1}^G \lambda_g \text{MVN}_d(\mu_g, \sigma_g^2 I_d),$$

where I_d is the $d \times d$ identity matrix, $\lambda_g > 0$ is the probability that an actor belongs to the g th group with $g = 1, \dots, G$ and $\sum_{g=1}^G \lambda_g = 1$. $\text{MVN}_d(\mu, \Sigma)$ denoted the jpdf of a multivariate Gaussian distribution with mean vector μ and covariance matrix Σ .

Model based clustering is becoming increasingly popular for many types of data (see Melynikov and Maitra, 2010). In this frame, the main idea is that the positions of the actors depend on their assigned cluster and thus it is more probable that members of the same cluster are connected, than those of different. The number of clusters is considered unknown a priori and it is estimated from the data. In similar finite mixture problems, the estimation is applied sequentially for different number of clusters, while that number is selected by applying information or other criteria.

2.3 Estimation

As Handcock et al (2007) proposed, there are different estimation methods for the LPCM. These methods concern latent position estimation. The most popular methods among them, are the “two-stage maximum likelihood” and the “Bayesian estimation”.

The first method is applied in two stages. The first one, is to bring off the maximum likelihood estimation of the latent positions for the non-clustering social space model of Hoff et al (2002). The second stage is to find a maximum likelihood estimator of the mixture model conditionally on the latent positions estimated in the first stage. This is based mainly on a standard EM type algorithm for finite mixtures.

Despite being fast and simple, the above method has a disadvantage. It does not estimate the latent positions and the cluster model simultaneously which results in losing information from the cluster structure that may be useful in estimating the latent positions and vice versa. Details about the steps of the algorithm can be found at Handcock et al (2007).

The second method is a fully Bayesian method based on MCMC sampling. The approach uses an hierarchical Bayesian model. As Handcock et al (2007) proposed for the LPCM, prior distribution is imposed on the latent space positions and the covariate coefficients, while the hyperpriors are imposed on the latent space prior parameters. Details on the prior specification can be found in Handcock et al (2007). The Bayesian approach can be time consuming for large networks.

3. APPLICATION: CO-AUTHORSHIP ON SCIENTIFIC ARTICLES NETWORK

In this section, the model is applied in a real co-authorship network dataset among faculty members in Department of Statistics in AUEB. The LPCM is fitted, estimating both the actors' latent positions in "social space" and the clusters they are assigned to.

Co-authorship issues have been addressed at the past from several different views. A simple way these issues to be handled is by creating a socio-matrix where the co-authorships are indicated. Then standard statistical techniques, like distance based clustering or any other statistical method of similar nature (i.e correspondence analysis or scaling methods) are applied. Alternatively, network methods may be applied, for example like in Newman (2004). In this section, the LPCM is applied to co-authorship data.

Sociometric information was collected about the relation of "co-authorship" on scientific articles among faculty members. The form of data produced is a 25×25 sociomatrix Y where its entries $y_{i,j}$ are either zero or one, indicating the presence or absence of a 'link' from i to j . In particular, it is considered that an actor has a social relation to another if they have collaborated at least once during their academic life in order to write an article. Data derive from searching the Scopus data base and last search was made in December 2012. In this application the number of common articles is ignored, as well as other related issues like the journal where the article is published in.

The data are considered coming from a mixture of probability distributions where each one represents a different cluster. The primary objective is to choose the number of clusters that best fit to our data. The primary concern is the way faculty members can be classified into groups in relation to their co-authorship interactions. Recasting this problem as one of model selection, the number of clusters that best fit to the data is estimated. In other words, a different statistical model is fitted every time a particular number of clusters is considered. Following this, we compare the arising models using BIC.

Figure 1 presents the results of fitting the LSM to the data. Without assuming any clustering (i.e. $G = 1$), the latent positions of the faculty members are visualized in the "social space". Lines imply that the two authors have a joint paper. Members without a link are those that never co-authored a paper. As expected, due to the unit mean root assumption of the model, the actors are positioned in

the periphery of the plot thus creating an outer circle. It is underlined that this model does not imply any clustering and just represents the data in an optimal network.

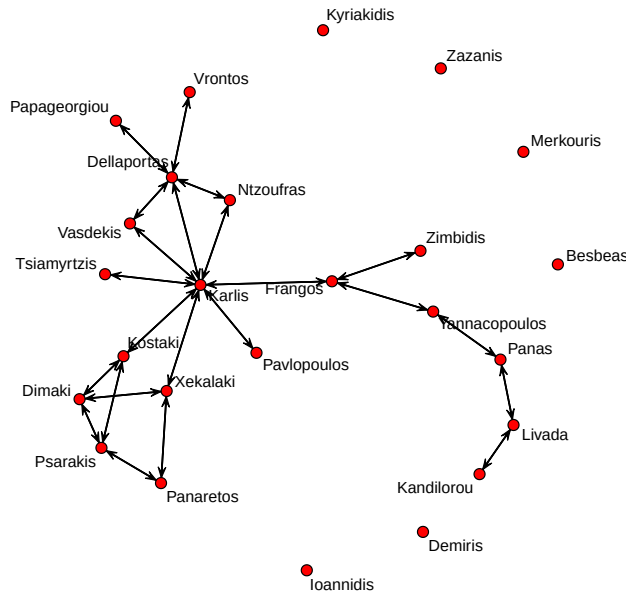


Figure 1: The latent positions of the faculty members in the “social space”. Lines imply that the two authors have a joint paper. Members without a link are those that never co-authored a paper.

Since the number of clusters is unknown, four different models are fitted sequentially to the co-authorship network dataset. In particular models for different number of clusters, $G = 2, 3, 4, 5$ are fitted. Then, using the methods described in section 2.3 the positions of the actors are estimated and the results for both methods are plotted (see Fig.2).

The choice of the number of clusters is based on BIC which indicates a clear choice of four clusters. The library `latentnet` is used in R to fit the models. Details about the library can be found in Krivitsky and Handcock (2008) and Handcock et al (2004).

Based on BIC and the maximum posterior probabilities produced by MCMC sampling, it is estimated that the faculty members create four compact clusters. In this way, it is considered that there is a similarity among the individuals of each particular cluster. The first group has to do with Bayesian (mostly) topics, the second one has to do with various topics while the third with finance and actuarial. The final group indicates those members that have not published with

other members.

Between the four clusters, there is one whose members are co-authors with actors from other clusters. Due to this fact, this specific cluster is regarded as the most central group and the main link inside the social network.

Figure 2 (right) compares the four models fitted to the network data using Bayesian criteria. The same conclusions are obtained based on the highest posterior probability of each candidate model. Moreover, Table 1 presents the uncertainty in the cluster assignments of each actor for the final four-cluster model. Following this, the results are summarized, deciding which model best fits to the data. At the same time, a clear picture of the members' classification into clusters is acquired.

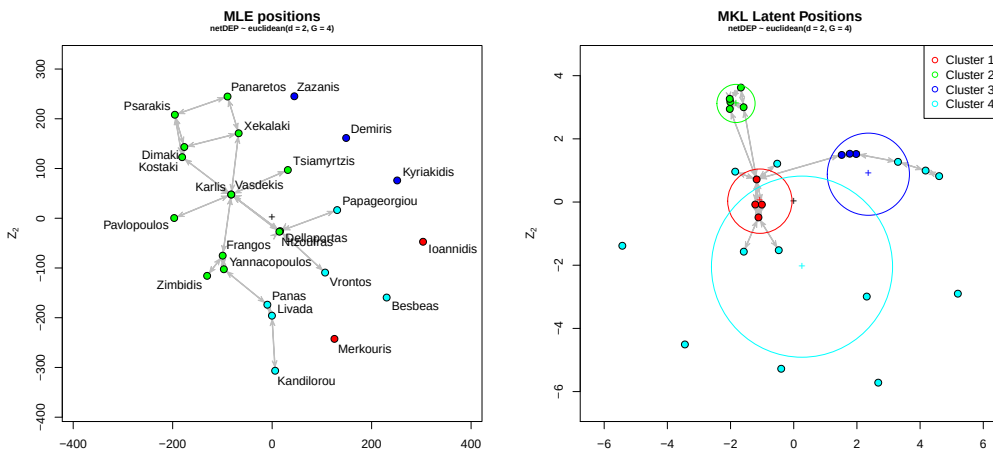


Figure 2: Left: Co-authorship between faculty members from two-stage maximum likelihood estimation of the LPCM with four groups that are shown by the colors of the nodes. Right: Estimates of clusters and latent positions for the faculty members co-authorship in Department of Statistics in AUEB from the Bayesian estimation of the LPCM.

4. CONCLUDING REMARKS

The analysis of social network data is an area that is rapidly growing, both within and outside of the discipline of statistics. Researchers from biology and bioinformatics to physics, from computer science to the information sciences and from economics to sociology are more and more engaged in the collection and statistical analysis of network data. Social network analysts focus on identifying central actors in the network, detecting communities or tracing information diffusion and opinion formation.

name	P_{cl_1}	P_{cl_2}	P_{cl_3}	P_{cl_4}
Vasdekis	0.8355	0.0080	0.215	0.1350
Vrontos	0.4105	0.0015	0.0575	0.5305
Yannacopoulos	0.0390	0.0035	0.7600	0.1975
Dellaportas	0.7880	0.0040	0.0280	0.1800
Demiris	0.0105	0.0010	0.0495	0.9390
Dimaki	0.0085	0.9770	0.0020	0.0125
Zazanis	0.0085	0.0005	0.1110	0.8800
Zimbidis	0.0470	0.0050	0.7160	0.2320
Ioannidis	0.0210	0	0.0340	0.9450
Kandilorou	0.0075	0	0.6435	0.3490
Karlis	0.8105	0.0215	0.0235	0.1445
Kyriakidis	0.0725	0.0040	0.0305	0.8930
Kostaki	0.0125	0.9675	0.0020	0.0180
Livada	0.0105	0.0005	0.7085	0.2805
Merkouri	0.0225	0.0000	0.0900	0.8875
Besbeas	0.0420	0.0005	0.0250	0.9325
Ntzoufras	0.8215	0.0055	0.0225	0.1505
Xekalaki	0.0140	0.9665	0.0050	0.0145
Panaretos	0.0065	0.9710	0.0010	0.0215
Panas	0.0165	0	0.7725	0.2110
Papageorgiou	0.4535	0.0030	0.0340	0.5095
Pavlopoulos	0.6515	0.0340	0.0705	0.2440
Tsiamyrtzis	0.7435	0.0440	0.0205	0.1920
Frangos	0.0705	0.0045	0.6630	0.2620
Psarakis	0.0075	0.9710	0.0045	0.0170

Table 1: Membership a posteriori probabilities for the 4-cluster model. With bold, the cluster that we assign each actor based on the maximum a-posteriori probabilities.

This paper discusses the use of models in order to represent social network data. We first summarize the network in a visual manner in order to draw a picture of the way actors are located in “social space”. The co-authorship network data are modeled via the LPCM of Hancock et al (2007) which extends the LSM of Hoff et al (2002) in order to take account of clustering, while the number of clusters is considered unknown a priori. Under this model, the probability of a tie between actors depends on the distance between them.

Finally, the model and methods discussed are illustrated via application concerning the co-authorship on scientific articles faculty network of department of Statistics in AUEB (Athens University of Economics and Business). More details and other related applications can be found in Oikonomopoulou (2014).

We strongly believe that such plots and approaches could be useful for comparing different departments in the level of “working together” as well as in the extend of variability among the members. Also a timeline of such approaches could be a useful tool to understand how synergies are created and whether they are supported.

Regarding further work, several interesting extensions are possible. The approach we discussed can be extended to include covariates such as area of work, position or actors’ years working in faculty revealing important factors. Another idea is the usage of a different response such as the number of papers co-authored. Moreover, creating networks based on time restrictions of the co-authorship, dynamic graphs can be produced aiming to explore the evolution of the relationship on time. Other models like a stochastic block-model can be applied to analyse a co-authorship network, while a different approach relaxing the assumption of zero covariance (spherical clusters) can be examined.

ΠΕΡΙΛΗΨΗ

Στο άρθρο αυτό παρουσιάζουμε τη χρήση στατιστικών μοντέλων για δεδομένα που αφορούν δίκτυα και που χρησιμεύουν τόσο για την οπτικοποίηση όσο και για την περιγραφή των τοπολογικών χαρακτηριστικών τους. Συγκεκριμένα χρησιμοποιούμε το Latent Position Cluster Model (LPCM) των Hancock et al (2007) για να τοποθετήσουμε το δίκτυο σε έναν λανθάνοντα “κοινωνικό χώρο” που αντιπροσωπεύει τα δεδομένα αλλά και συγχρόνως μπορεί να αναδείξει την ομαδοποίηση τους. Το μοντέλο επίσης επιτρέπει να κατατάζει κανείς τα μέλη του δικτύου σε ομάδες. Εφαρμόζουμε το μοντέλο σε ένα πραγματικό σετ δεδομένων που αφορά την συ-συγγραφή ερευνητικών εργασιών ανάμεσα στα μέλη ΔΕΠ του τμήματος Στατιστικής του ΟΠΑ σε μια προσπάθεια να μελετήσουμε την ύπαρξη ακαδημαϊκών συνεργειών ανάμεσα στα μέλη ΔΕΠ του τμήματος.

REFERENCES

- Ambroise, C. and Matias, C. New consistent and asymptotically normal parameter estimates for random-graph mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 74:3–35, 2012 2004.
- Banfield J. D. and Raftery A. E.. Model-based Gaussian and non-Gaussian clustering. *Biometrics*, 49:803–821, 1993.
- Handcock, M. S., Raftery, A.E. and Tantrum, J.M. (2007) *Model-based clustering for social network* *Journal of the Royal Statistical Society, Series A* 17049: 301–354, 2007
- Handcock, M. S., Tantrum, J. Shortreed, S. and Hoff, P. (2004) **laternet**: an R package for latent position and cluster modeling of statistical networks (Available from <http://www.csde.washington.edu/statnet/latentnet>).
- Hoff P. D., Raftery, A. E. and Handcock, M.S.(2002) *Latent space approaches to social network anlysis*. *Journal of the American Statistical Association* 97, 1090-1098.
- Kolaczyk, E. D. *Statistical Analysis of Network Data. Methods and Models*. Springer Series in Statistics. 2003
- Krivitsky, P. N. and Handcock, M. S. (2008) Fitting Position Latent Cluster Models for Social Networks with latentnet. *Journal of Statistical Software* 24 Issue 5.
- Latouche, P., Birmelo, E. and Ambroise, C. Overlapping stochastic block models with application to the French political blogosphere. *Annals of Applied Statistics*, 5 (1):309–336, 2012.
- Melnykov, V. and Maitra, R. Finite mixture models and model-based clustering. *Statistical Surveys*, 4:80–116, 2010.
- Newman, M. E. J. Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Sciences of the United States of America*, 101 (Suppl 1):5200–5205, 2004.
- Oikonomopoulou, G. Statistical modelling of Social Networks: an Application to co-authorship data. *Unpublished M.Sc. thesis, Department of Statistics, Athens University of Economics*, 2014.

HIDDEN MARKOV MODELS IN MODELLING TIME SERIES OF EARTHQUAKES

K. Orfanogiannaki^{1,2}, D. Karlis¹

¹Institute of Geodynamics, National Observatory of Athens, Greece

{korf, karlis}@aueb.gr

²Department of Statistics, Athens University of Economics and Business, Greece

ABSTRACT

Discrete valued Hidden Markov Models (HMMs) are used to model time series of event counts in several scientific fields. The model has two parts: the observed sequence of event counts and an unobserved (hidden) sequence of states that consist a Markov chain. Each state is characterized by a specific distribution and the progress of the hidden process from state to state is controlled by a transition probability matrix. In this work we aim to present an application of HMMs to a bivariate discrete valued time series occurring in seismology by extending the existing univariate models. In particular, on 26 December 2004 and 28 March 2005 occurred two of the largest earthquakes of the last 40 years between the Indo-Australian and the southeastern Eurasian plates with moment magnitudes $Mw = 9.1$ and $Mw = 8.6$, respectively. An interesting question is to examine whether the events can be correlated. To do so we examine the time series containing the daily number of events in the region of each mainshock. Our aim is firstly to identify the dynamics for each series separately by fitting univariate Poisson HMMs and secondly to account for any correlation between the two series.

While models for univariate discrete valued time series are well known we extend the HMMs to the bivariate case by assuming appropriate bivariate discrete distributions for each state. We examine properties of the model and propose inference. Maximum likelihood estimators of the models' parameters are derived using an EM algorithm.

Keywords: Hidden Markov Models; Poisson; Bivariate Poisson; earthquake counts

1. INTRODUCTION

HMMs are well known models with many applications including seismology (see MacDonald and Zucchini (1997), Ebel et al. (2007) and Orfanogiannaki et al. (2010)). They allow for overdispersion (variance larger than the mean), autocorrelation and zero inflation in the data. Characteristics that are often present in many real data. While the literature on HMM now flourishes, there is

a lack of such models for multivariate count data, i.e. when many count random variables are observed together in different time points. Such models may arise in several disciplines such as different types of earthquakes in different time periods or number of earthquakes in different areas, number of faults in parts of a large system, number of purchases of different products observed in different time points etc. They also constitute a class of time series models for multivariate counts which is less developed and has itself particular interest.

We extend HMMs to the bivariate case by assuming both the standard bivariate Poisson distribution and a bivariate discrete distribution with Poisson marginals based on the Frank Copula for each state. Copula based distributions allow for flexible dependent structure and makes them fashionable models to describe dependence. The selection of the particular copula family is based on the fact that it allows not only for positive but also negative correlation. Of course, any other copula can be used. We use the proposed models to jointly model the daily frequency of earthquakes in 2 adjacent areas in the Sumatra rupture zone. The models' parameters are estimated using an EM algorithm.

2. PROPOSED MODELS

2.1 Hidden Markov models: the general context

HMMs are discrete time stochastic processes that consist of two parts. The first part is an unobserved finite state Markov chain $\{C_i : i \in N\}$ on m states. The second part is a non-negative integer valued sequence of random variables $\{Y_i : i \in N\}$ such that, conditionally on C_i , are mutually independent. Each state is associated with a probability distribution function f from the same parametric family Λ . When C_i is known Y_i takes the value y_i with probability $f(y_i | c_i)$. If in the univariate case we assume that each observation is generated from a Poisson distribution then f takes the form $f(y_i | c_i = j) = \text{Pois}(y_i | \lambda_j) = \frac{e^{-\lambda_j} \lambda_j^{y_i}}{y_i!}$ where $\lambda_j \geq 0$ and $y_i = 0, 1, \dots$, for all $i = 1, \dots, n$ and $j = 1, \dots, m$.

The parameters of the model are the transition probabilities of the Markov chain and the parameters of the probability distributions that are associated with the states. The transition probabilities γ_{lj} are defined as: $\gamma_{lj} = P(C_i = j | C_{i-1} = l)$. This is the probability that given the hidden process was in state l at the previous time point, it will be in state j at the current. If we denote with Ψ the parameter vector to be estimated then the likelihood of an HMM is:

$$L(\Psi | y_1, \dots, y_n) \propto \sum_{c_1=1}^m \dots \sum_{c_n=1}^m P(C_1 = c_1) f(y_1 | c_1) \prod_{i=2}^n \gamma_{c_{i-1}c_i} f(y_i | c_i)$$

where n is the length of the observation sequence. For the likelihood of a HMM to be calculated, the backward, $\beta_j(i)$, and forward, $\alpha_j(i)$, probabilities were introduced by Baum et al. (1970): $\beta_j(i) = P(y_{i+1}, \dots, y_n | C_i = j)$, $\alpha_j(i) =$

$P(y_1, \dots, y_i, C_i = j)$. The likelihood can then be calculated in terms of the forward probabilities as: $L = \sum_{l=1}^m \alpha_l(n)$.

We will formulate 2 different models. The Bivariate Poisson Hidden Markov model (BPHMM) and a Hidden Markov model based on the Frank copula (HMMC). What differentiates the models is the selection of the distribution family. In the BPHMMs each state corresponds to a standard bivariate Poisson distribution while in the HMMCs each state corresponds to a bivariate distribution with Poisson marginals defined via a Frank copula.

2.1.1 Bivariate Poisson Model

In the BPHMM the distribution associated with each state is a bivariate Poisson distribution. Assume that X_i , are independent Poisson distributions with parameters λ_i , respectively where $i = 0, 1, 2$. The random variables Y_1, Y_2 defined as: $Y_1 = X_1 + X_0$ and $Y_2 = X_2 + X_0$ follow the bivariate Poisson distribution with parameters $(\lambda_0, \lambda_1, \lambda_2)$. In BPHMM each state is associated with a different bivariate Poisson distribution. Consider, that $\boldsymbol{\lambda}^j = (\lambda_0^j, \lambda_1^j, \lambda_2^j)'$ is the vector of parameters of the bivariate Poisson distribution that corresponds to state j and $\mathbf{y}_i = (y_{1i}, y_{2i})'$ for $i = 1, \dots, n$ is the vector of observed data that corresponds to the i -th observation. The joint probability distribution of Y_1, Y_2 that corresponds to state j is given by the formula:

$$f(y_{1i}, y_{2i}) = BP(\mathbf{y}_i | \boldsymbol{\lambda}^j) = e^{-(\lambda_1^j + \lambda_2^j + \lambda_0^j)} \frac{\lambda_1^{j y_{1i}}}{y_{1i}!} \frac{\lambda_2^{j y_{2i}}}{y_{2i}!} \sum_{r=0}^s \binom{y_{1i}}{r} \binom{y_{2i}}{r} r! \left(\frac{\lambda_0^j}{\lambda_1^j \lambda_2^j} \right)^r,$$

$s = \min(y_{1i}, y_{2i})$, $y_1, y_2 = 0, 1, \dots$, and $\lambda_0^j, \lambda_1^j, \lambda_2^j$, while $Cov(Y_{1t}, Y_{2t}) = \lambda_{0j} \geq 0$ is the covariance between the random variables Y_1 and Y_2 and since it is always equal to a non negative number, only positive correlation is allowed. When the covariance is equal to zero independence is implied.

2.1.2 Hidden Markov Models with Copula

Copulas are bivariate (multivariate) distributions with uniform marginals. Copulas are currently fashionable models to describe dependence. For a more formal definition see Nelsen (2006).

In the discrete case in order to derive the joint probability mass function (pmf) we need to take differences i.e.: for the bivariate case with marginals $F(x)$ and $G(y)$ the joint pmf is given by the formula:

$$\begin{aligned} P(x, y) &= C(F(x), G(y)) - C(F(x-1), G(y)) - \\ &\quad C(F(x), G(y-1)) + C(F(x-1), G(y-1)) \end{aligned}$$

where $F(\cdot)$ and $G(\cdot)$ are the marginal cdfs. This can be generalized to larger dimensions, but as dimensions increase excessive summation is needed. This creates a challenge on selecting copulas that can be efficient for the calculations. So, far we have worked with a bivariate Frank copula given by

$$C(u, v) = -\frac{1}{\tau} \log \left\{ 1 + \frac{(\exp^{-\tau u} - 1)(\exp^{-\tau v} - 1)}{(\exp^{-\tau} - 1)} \right\}$$

where τ is the copula parameter representing the dependence implied. Frank copula allows for both negative and positive correlation. Of course any other copula can be used. In our case we have selected Poisson marginals.

3. PARAMETER ESTIMATION

Due to the underlying structure of HMMs that allow for a missing data representation of the model an EM algorithm (Dempster et al. (1997)) is adopted for Maximum Likelihood estimation of the parameters of interest both in the univariate and in the bivariate case. In the univariate case (for details see MacDonald and Zucchini (1997)) and in the case of bivariate Poisson the parameters can be estimated by closed form equations. We will provide full details for the latter case. In a BPHMM with m states, the parameter vector to be estimated $\Psi = (\gamma, \lambda^j)$ consists of the transition probabilities between the states of the Markov chain $\gamma = (\gamma_{11}, \gamma_{12}, \dots, \gamma_{m-1,m}, \gamma_{m,m})'$ and the parameters of the Bivariate Poisson distribution for each state i.e. for state j : $\lambda^j = (\lambda_0^j, \lambda_1^j, \lambda_2^j)'$, $j = 1, \dots, m$. We define the indicator random variables $U_j(i)$ and $V_{jk}(i)$ respectively, where $U_j(i) = 1$, if $C_i = j$ and 0 otherwise and $V_{jk}(i) = 1$, if $C_{i-1} = j$ and $C_i = k$ and 0 otherwise. We also define a vector of latent variables $\mathbf{X}_i^j = (X_{0i}^j, X_{1i}^j, X_{2i}^j)'$ that correspond to state j and the i -th observation as: each random variable follows a Poisson distribution i.e. $X_{qi}^j \sim \text{Poisson}(\lambda_q^j)$. The complete data log-likelihood, in terms of the latent and the indicator random variables is given by:

$$l(\Psi) = \log \pi_{c_1} + \sum_{i=2}^n \sum_{j=1}^m \sum_{l=1}^m v_{lj}(i) \log \gamma_{lj} + \sum_{i=2}^n \sum_{j=1}^m u_j(i) \sum_{r=0}^2 \log \text{Pois}(x_{ri}^j | \lambda_r^j)$$

where π_{c_1} is the initial state probability, i.e. the initial state of the model is C_1 . At the E-step of the EM algorithm we compute the conditional expectations of the indicator and the latent variables using the observed data. At the M-step we maximize the complete data log-likelihood to update the parameter estimates. A detailed description of the EM algorithm is the following:

- E-step: Using the observed data and the current estimates of the parameters compute the backward and forward probabilities: $\beta_j(n) = 1, j = 1, \dots, m$,
 $\beta_j(i) = \sum_{l=1}^m \gamma_{jl} \cdot BP(\mathbf{y}_{i+1} | \lambda^l) \cdot \beta_l(i+1)$ and $\alpha_j(i) = \sum_{l=1}^m \alpha_l(i-1) \cdot \gamma_{lj} \cdot$

$BP(\mathbf{y}_i|\boldsymbol{\lambda}^j)$, $\alpha_j(1) = \pi_{c_1=j} \cdot BP(\mathbf{y}_1|\boldsymbol{\lambda}^j)$, $j = 1, \dots, m$. Then using the backward and forward probabilities calculate the indicator variables: $v_{lj}(i) = \frac{\gamma_{lj} \cdot BP(\mathbf{y}_i|\boldsymbol{\lambda}^j) \cdot \alpha_l(i-1) \cdot \beta_j(i)}{\sum_{l=1}^m \alpha_l(n)}$ and $u_j(i) = \frac{\alpha_j(i) \cdot \beta_j(i)}{\sum_{l=1}^m \alpha_l(n)}$. The final part of the E-step

is to compute the latent variables $\mathbf{x}_i^j = (x_{0i}^j, x_{1i}^j, x_{2i}^j)'$ for $i = 1, \dots, n$:

$$x_{0i}^j = \frac{\sum_{r=0}^{\min(y_{1i}, y_{2i})} r \text{Pois}(y_{1i} - r | \lambda_1^j) \text{Pois}(y_{2i} - r | \lambda_2^j) \text{Pois}(r | \lambda_0^j)}{BP(\mathbf{y}_i|\boldsymbol{\lambda}^j)}$$

$$x_{1i}^j = y_{1i} - x_{0i}^j \text{ and } x_{2i}^j = y_{2i} - x_{0i}^j$$

- Update the parameter estimates by: $\gamma_{lj}^{(new)} = \frac{\sum_{i=2}^n v_{lj}(i)}{\sum_{i=2}^n \sum_{h=1}^m v_{lh}(i)}$ and in vector

$$\text{form: } \lambda^{j(new)} = \frac{\sum_{i=1}^n u_j(i) \mathbf{x}_i^j}{\sum_{i=1}^n u_j(i)}$$

- If some convergence criterion applies stop iterating, otherwise return to the E-step for one more iteration.

In the model with copula there are no close equations so numerical maximization technics are used (we implemented them in R). The problem is equivalent to one of maximizing the log-likelihood for a copula based discrete model. Numerical optimization is found easy to use if good initial estimates are available.

4. APPLICATION

4.1 Data

Two large earthquakes occurred in the region of Sumatra on 26 December 2004 ($Mw = 9.1$) and 28 March 2005 ($Mw = 8.6$). The data sources are the USGS and ISC earthquake data files for the region defined by the rectangle E with coordinates $-1.00N - 15.00N$ and $91.00E - 100.00E$. This region was divided into two sub-regions, North (N) and South (S), corresponding to the rupture zones of the two big earthquakes of 26.12.04 and 28.03.05, respectively (see Lay et al (2005)). Data completeness analysis based on the magnitude-frequency relationship showed that the data in all regions are complete for $mb \geq 4.2$ for the time interval from 01.01.2000 to 25.12.2004. Each time series consists of earthquake counts, in daily time units, in the regions determined by the aftershock zones of the two mainshocks. The time series of the number of earthquakes that occurred in each sub region versus time appear in Figure 1. The mean number (variances) of earthquakes per day is 0.14 (0.43) in the North sub-region and 0.10 (0.22) for the South. The variance in the North sub-region is almost 3 times the mean while in the South it is double than the mean. There is evidence of overdispersion, in

both time series. 248 earthquakes have occurred in the North sub-region and 185 in the South, in the time period examined. The maximum number of earthquakes observed in the North sub-region is 12, while in the South is 14. The correlation between the two regions is 0.03. Hidden Markov models are adequate models to describe for both overdispersion and serial correlation that appear in the univariate series of earthquake counts. The extended HMMs to the bivariate case allow us to jointly model the frequency of earthquakes in the 2 sub-regions and estimate the correlation between them.

Figure 1: Time series plots (top and middle) for the daily number of earthquakes occurring in the 2 sub-regions examined. The zero point in time corresponds to 01.01.2000 while the end point in time is 25.12.2004. At the bottom is the estimated sequence of Hidden States.

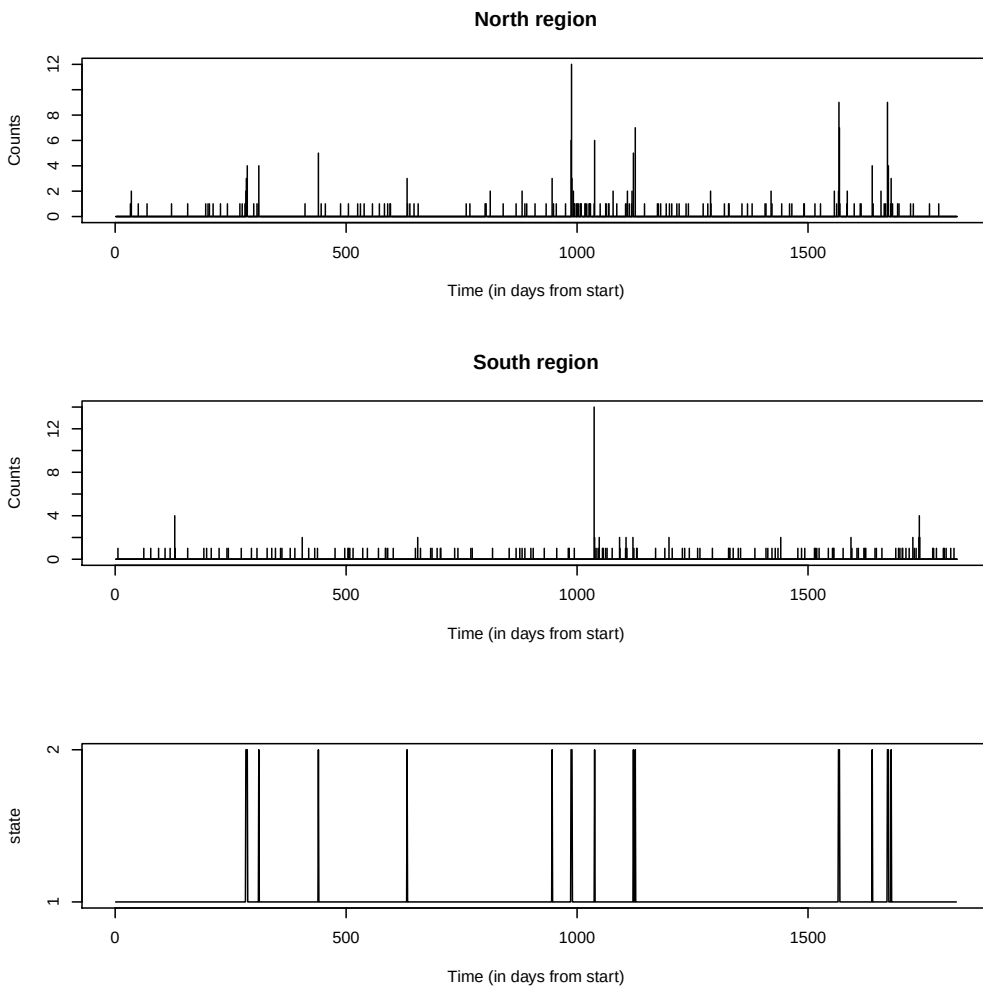


Table 1: Comparison of the fitted models on the basis of BIC. Key: m is the number of states

m	Independent PHMMs		Bivariate Poisson		Frank Copula Biv Poisson	
	Loglik	BIC	Loglik	BIC	Loglik	BIC
2	-1291.183	2627.409	-1291.085	2642.226	-1254.663	2569.383
3	-1250.091	2590.268	-1248.038	2608.683	-1235.939	2584.486
4	-1239.323	2628.789	-1237.453	2655.076	-1225.423	2631.018

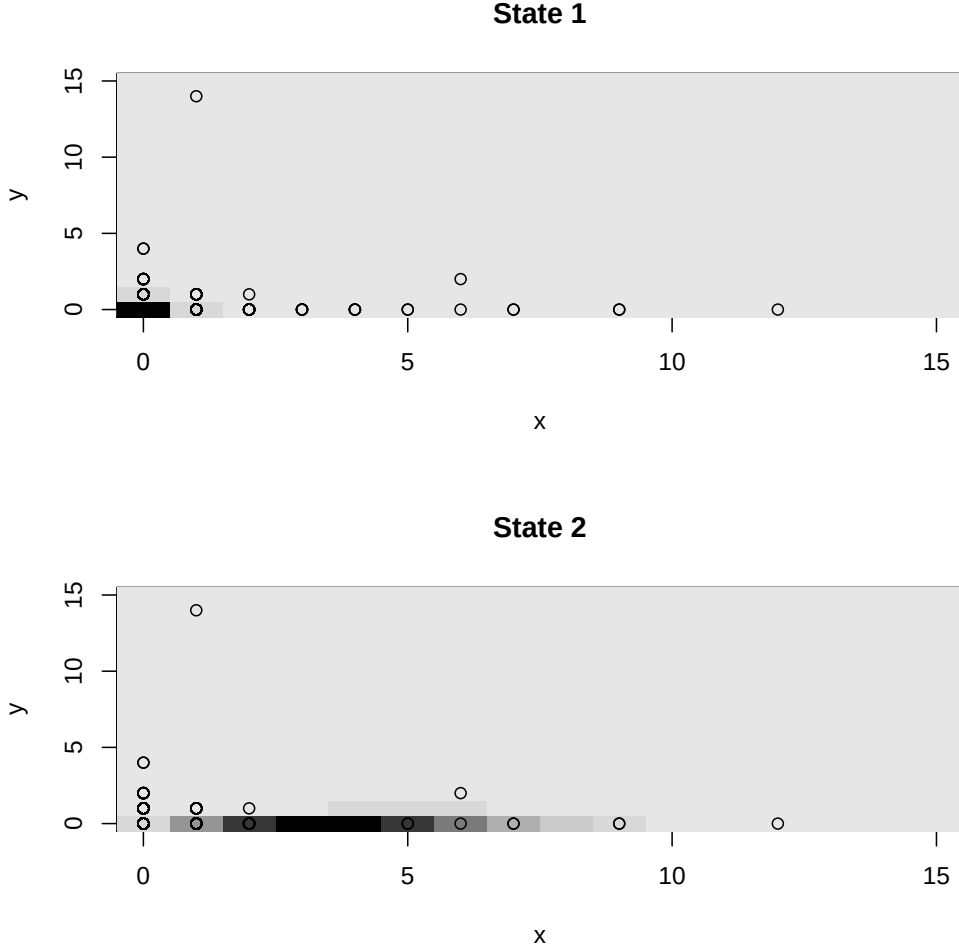
4.2 Results

We applied PHMMs to the 2 univariate time series of earthquake counts for different number of states. We also modeled jointly the two time series considering both BPHMMs and HMMC. The comparison between the different models was based on the values of the Bayesian information criterion (BIC) defined as: $BIC(m) = -2l(k) + \ln(n)d_f$, where $l(k)$ is the maximized log-likelihood for the model with m states, d_f is the number of free parameters of that model and n is the size of the sample. The values of the maximized log-likelihood and of BIC for the different models with different number of states are summarized in Table 1. The model that best describes the data is the HMMC with 2 states. The parameter estimates for the HMMC model with 2 states are shown in Table 2. Each column of the table corresponds to a different state. The first 2 rows correspond to the seismicity rates for the 2 regions while the last row corresponds to the parameter of the Frank copula. The estimated transition probability matrix determined for the HMMC model with 2 states is

$$\hat{\Gamma} = \begin{bmatrix} 0.99 & 0.01 \\ 0.61 & 0.39 \end{bmatrix}.$$

State 1 corresponds to a state of seismic quiescence. When the two regions are in state 1 they remain in that state with probability 0.99. This state tries to describe the excessive number of zeros present in the two time series. This is also evident in Figure 2. At the top diagram of Figure 2 where the colored rectangles represent the probability density values of the Frank Copula that correspond to state 1, the probability of observing zero counts in both regions is very high. At the bottom diagram of the same Figure, high probabilities are observed for moderate counts in the North and very small counts in the South sub-region. The North sub-region is the region where the 26 December 2004 large earthquake occurred. There is some moderate seismic activity before the earthquake occurrence that is captured by state 2. During this period the seismicity in the South subregion remains low. State 2 is a very active state for the North sub-region. This is also evident by the bottom diagram of Figure 1 where the sequence of Hidden States that underlie

Figure 2: Scatter plots of observations. x and y correspond to the earthquake counts observed in the North and South sub-regions respectively. The colored rectangles represent the joint probability density values of the Frank Copula that correspond to state 1 (top) and state 2 (bottom). The closer the probability is to unity the darker is the color.



the data is plotted versus time. The picks correspond to state 2 and they are associated with large counts observed in the North sub-region.

5. CONCLUDING REMARKS

We developed 2 different models BPHMMs and HMMC by extending the existing univariate models. In BPHMMs each state is associated with a Bivariate

Table 2: *Parameter estimates for the HMMC model with 2 states.*

	1st state	2st state
North	0.076	4.020
South	0.094	0.077
copula parameter	0.825	2.844

Poisson distribution. In HMMC a Frank copula corresponds to each state. In the case of the Bivariate Poisson distribution the parameters are the parameters of the univariate Poisson distributions and the covariance between them for each state. In the case of the Frank copula the parameters are the parameters of two univariate Poisson distributions and the parameter of the Frank Copula for each state. The parameter of the Frank Copula is different from the covariance between two univariate Poisson distributions. It is associated with the two well-established measures of correlation which are Spearman's ρ and Kendall's τ but the functions that determine the relationship between them are not simple. The improvement of the likelihood when copula are included in the model is substantial which highlights the need for models that allow for more flexible dependence structure.

The usefulness of the proposed models is illustrated with real data application. We applied all the different models with different number of states to jointly model daily earthquake counts in the rupture zones of the two of the largest earthquakes of the last 40 years that occurred between the Indo-Australian and the southeastern Eurasian plates in Sumatra. As the model that best fits the data was selected the HMMC with 2 states. From the results obtained the characteristics of the two sub-regions are revealed. Such an approach would help to improve knowledge on the interrelationships between different areas. Of course, further applications from areas with different seismotectonic characteristics would be enlightening.

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.

Περίληψη

Εφαρμογές των διακριτών Λανθανόντων Μοντέλων Μάρκοβ σε χρονολογικές σειρές που αποτελούνται από συχνότητες γεγονότων γίνονται σε διάφορα επιστημονικά πεδία. Τα μοντέλα αποτελούνται από δύο μέρη: το πρώτο μέρος είναι μία παρατηρούμενη ακολουθία από συχνότητες γεγονότων και το δεύτερο μέρος είναι μία μη παρατηρούμενη ακολουθία καταστάσεων που αποτελούν μία Μαρκοβιανή αλυσίδα. Κάθε κατάσταση χαρακτηρίζεται από μία συγκεκριμένη κατανομή και η μετάβαση της Λανθάνουσας ακολουθίας από μία κατάσταση σε άλλη καθορίζεται από τον πί-

νακα πιθανοτήτων μεταπήδησης της Μαρκοβιανής αλυσίδας. Σκοπός μας είναι να παρουσιάσουμε μία εφαρμογή των Λανθανόντων μοντέλων Μάρκοβ σε μία διακριτή δι-μεταβλητή χρονολογική σειρά από σεισμολογικά δεδομένα, επεκτείνοντας την υπάρχουσα μονο-μεταβλητή περίπτωση. Συγκεκριμένα στις 26 Δεκεμβρίου 2004 και στις 28 Μαρτίου 2005, σημειώθηκαν δύο από τις ισχυρότερες σεισμικές δονήσεις των τελευταίων 40 χρόνων με μεγέθη ροπής $Mw = 9.1$ και $Mw = 8.6$, αντίστοιχα. Παρουσιάζει ενδιαφέρον να εξετάσουμε κατά πόσο τα γεγονότα αυτά μπορεί να είναι συσχετισμένα. Για το σκοπό αυτό μελετάμε τις χρονολογικές σειρές που περιέχουν τον ημερήσιο αριθμό σεισμών που σημειώθηκαν στη ζώνη διάρρηξης του κάθε κύριου σεισμού. Επεκτείνουμε τα γνωστά μοντέλα για διακριτές μονο-μεταβλητές χρονολογικές σειρές στη δι-μεταβλητή περίπτωση χρησιμοποιώντας κατάλληλες δι-μεταβλητές διακριτές κατανομές για κάθε κατάσταση. Στη συνέχεια εξετάζουμε ιδιότητες των μοντέλων και προτείνουμε συμπερασματολογία. Εκτιμητές μέγιστης πιθανοφάνειας των παραμέτρων του μοντέλου προκύπτουν με τη χρήση του αλγορίθμου EM.

REFERENCES

- Baum, L.E., Petrie, T., Soules, G. and Weiss, N. (1970). A maximization technique in the Statistical analysis of probabilistic functions of Markov chains. *The Annals of Mathematical Statistics*, **41**, 164-171.
- Dempster, A.P., Laird, N.M. and Rubin, D.B (1997). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society B*, **39**, 1-38.
- 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. *Geophysical Research Letters*, **78**, 57-65.
- Lay, T., Kanamori, H., Ammon, C.J., Nettles, M., Ward, S.N., Aster, R.C., Beck, S.L., Bilek, S.L., Brudzinski, M.R., Butler, R., DeShon, H.R., Ekstrom, G., Satake, K., Sipkin, S. (2005). The great Sumatra-Andaman earthquake of 26 December 2004. *Science*, **308**, 1127-1133
- MacDonald, I.L. and Zucchini, W. (1997). *Hidden Markov models and other models for discrete valued time series*, London: Chapman & Hall.
- Nelsen, R. (2006). *An Introduction to copulas, 2nd Edition* New York: Springer.
- Orfanogiannaki, K., Karlis, D. and Papadopoulos, G.A. (2010). Identifying seismicity patterns using Poisson Hidden Markov Models. *Pure Applied Geophysics*, **167**, 919-931.



PRICE DISCOVERY UNDER CRISIS: UNCOVERING THE DETERMINANT FACTORS OF PRICES USING EFFICIENT BAYESIAN MODEL SELECTION METHODS

Athanassios Petralias¹, Prodromos Prodromidis^{2,3}

¹Department of Statistics, Athens University of Economics and Business

²Center for Planning and Economic Research (KEPE)

³Department of Management Science and Technology, Athens University of
Economics and Business

petralia@aueb.gr; pjprodr@kepe.gr

ABSTRACT

We seek to uncover the determinants of consumer prices in Greece by considering a large set of potential factors and accounting for delayed effects. To accomplish this we rely on recently presented Bayesian model selection methods which are efficiently adapted in the present context. Consumer prices drop significantly during sales periods, especially in the last two years; VAT changes are mostly absorbed (above 50%) by the producers; while a drop in retail sales affects prices with a significant lag (of 6 months). Evidence on the ways several other factors affect prices is also obtained.

Keywords: prices; taxes; sales period; retail sales; model selection; population MCMC; Subspace Carlin and Chib.

1. INTRODUCTION

Of all the countries that have been adversely affected by the recent economic crisis, the one that stands out is Greece. In the course of three years, from June 2009 to June 2012, the country's GDP (per capita) has fallen by 13.2%, and employee compensation by 27.4%, while prices, as measured by the domestic Consumer Price Index (CPI), have risen by 10.2%. This results to a drastic reduction in real incomes and thus purchasing power, feeding a central debate currently held in Greece and other European countries under crisis: Since demand falls, why prices are still rising? In an effort to shed light on this debate, we consider a large set of possible predictors and try to identify the main factors that drive prices in the Greek market.

2. THE MODEL

In the international literature, national CPIs are often explained empirically in terms of international food price indices, crude oil, other commodity price indices, producer price indices, output, the money supply, exchange and interest rates (e.g., Askari and Krichene, 2010; Jail and Tamayo Zea, 2011; Alemu, 2012; Kamenik et al, 2013; and other authors cited therein).

We consider a linear regression model for the first differences (i.e. monthly changes) of CPI

$$Y_t = c + \sum_{k=1}^K a_k S_{k,t} + \sum_{i=1}^I \sum_{l_i=0}^{12} \theta_{i,l_i} Z_{i,t-l_i}, \quad e_t \sim N(0, \sigma^2), \quad (1)$$

where $S_{k,t}$ correspond to K possible factors affecting the CPI's monthly change (Y_t) at the same time t , and $Z_{i,t-l_i}$ to I possible factors affecting the CPI's monthly change, each with lags (l_i) up to 12 months (see Table 1).

For a given model $m \in M$, the parameter space is defined as (suppressing index m)

$$\omega_m = \{k, i, l_i, c, a_k, \theta_{i,l_i}, \sigma^2\}, \quad k=1, \dots, K, \quad i=1, \dots, I, \quad l_i=0, \dots, 12.$$

That is, we jointly estimate parameters $\{c, a_k, \theta_{i,l_i}, \sigma^2\}$ and perform model selection with respect to factors k and i that may affect the CPI, as well as for each factor i , the specific lags l_i at which interacts (i.e. a factor may affect the CPI only at a given lag).

This way, the model space M includes $2^{(K+I \times 13)}$ possible models, which corresponds to a vast of 2^{356} models in our case, each with parameters ω_m to be estimated.

Note in the above formulation that one could also incorporate other technical factors (besides a trend and monthly dummies) such as ARMA-GARCH effects. However (a) a preliminary analysis showed that the first differences of the CPI (as well as the square of these differences) do not exhibit statistical significant autocorrelations or partial autocorrelations and (b) we are mainly interested in analyzing the way in which the factors considered affect the CPI and not in predictions.

3. THE DATA

The dataset consists of monthly observations of the domestic CPI from January 2001 up to August 2012 (140 observations). The potential determinant factors (observed on a monthly basis) considered to affect CPI are displayed in **Table 1**. These include: (a) technical factors such as monthly dummies, the trend and dummies capturing the CPI revisions by the statistical authority; (b) the VAT and sales periods, which directly affect the CPI level; (c) a dummy for the election of a new government, which may affect prices either because elections take place in the particular month or because measures taken by newly elected governments may affect CPI several months later; (d) prices of agricultural inputs such as seeds, fertilizers, pesticides, veterinary me-

Table 1. Determinant factors considered to affect prices

Factors ($K=18$) affecting prices at the same time instance $S_{k,t}$, $k = 1, \dots, K$	Factors ($I=26$) affecting prices with possible lags up to 12 months, $Z_{i,t-l_i}$, $i = 1, \dots, I$, $l_i = 0, \dots, 12$
1-11. Monthly Dummies (D1, D2,...,D11) capturing abnormal prices changes in the corresponding months (January-November) 12. Trend 13. VAT: The series was constructed as a weighted average of the VAT rates corresponding to each subcategory (4-digit) of the CPI at each month. Subgroup weights are those used by the National statistical service to construct the CPI 14. Sales: The percentage of each month (i.e. days out of total month days) corresponding to a sales period (based on formal sales periods as announced by the Ministry of Development from 2001 to 2012) 15. Revision of the CPI from the National Statistical Service in January 2005: Dummy variable with values equal to the unity in the period Jan. 2005-Dec. 2010 and zero otherwise 16. Revision of the CPI from the National Statistical Service in January 2011: Dummy variable with values equal to the unity in the period Jan. 2011 onwards and zero otherwise 17. Sales times January 2005 CPI revision (so as to capture different effects in the CPI during the sales period after the revision) 18. Sales times January 2011 CPI revision (so as to capture different effects in the CPI during the sales period after the revision)	1. New government: Dummy for the month a new government comes to power 2. Domestic price index of seeds 3. Domestic price index of fertilizers 4. Domestic price index of pesticides 5. Domestic price index of veterinary medicines 6. Domestic price index of animal feed 7. Domestic price index of agricultural equipment and machinery 8. Domestic price index of buildings and building materials 9. Domestic price index of electricity 10. Domestic price index of heating oil 11. Domestic price index of gasoline 12. Domestic price index of purchase prices of cars 13. Domestic price index of trains transports 14. Domestic price index of airplanes transports 15. Domestic price index of ship transports 16. Domestic Industrial Production index 17. Domestic Retail Sales index 18. Price index of Imports from the EU 19. Price index of Imports from non EU countries 20. International Meat Price Index 21. International Dairy Price Index 22. International Cereals Price Index 23. International Oils Price Index 24. International Sugar Price Index 25. Euro-dollar exchange rate 26. Spreads: 10 year Greek government bond yields versus German bond yields
Sources: Domestic indices and import prices: National Statistical Service of Greece; International price indices: FAO; Spreads and Euro-dollar exchange rate: Eurostat.	

dicines, animal feed and agricultural equipment of machinery that indirectly affect retail prices through the supply chain; (e) prices of buildings and building materials; (f) energy prices, including electricity and heating oil (g) transportation costs, namely gasoline prices, car prices, train, airline and ship fares; (h) indices capturing demand, supply and economic climate, namely industrial production, retail sales and interest rate spreads, and, finally (i) international prices of commodities (meat, dairy products, cereals, oil, sugar), import prices from EU and non-EU countries and euro-dollar exchange rate. Note that in the selection process of potential factors, there were limitations with respect to data availability. That is, several indicators of consumer demand and growth such as GDP and its proxies (wages, personal income) are available only on a quarterly basis. Considering the large number of potential predictors, any analysis based on data observed on lower frequencies (than monthly) would not be robust. Thus, besides technical and other factors which directly affect prices (i.e. VAT changes, sales periods), we consider those factors related to production cost, international prices, economic climate and consumer demand, that are available on a monthly basis.

All variables except for the dummies, trend and sales periods, are converted to indices the base value of which (100) has been set at the 2009 average. This is done to simplify the inference procedure and prior assignment, as well as the interpretation of model parameters. Then the first differences (i.e. monthly changes) of these indices are taken to construct the final dataset used in the analysis, to reduce the correlation among the 356 candidate variables.

4. INFERENCE

In such a complex model space we need a powerful inference tool to obtain the posterior quantities of interest, i.e. model probabilities and parameters ω_m . So we rely on Bayesian inference and construct a Population Markov Chain Monte Carlo (MCMC) algorithm that efficiently samples from both the model and parameter space.

4.1 Prior specification

Considering that the data have been converted to monthly changes of indices, we make use of typical vague proper priors, $\{c, a_k, \theta_{i,l_i}\} \sim N(0, 10^2)$, $\sigma^2 \sim IG(10^{-6}, 10^{-6})$.

For the model m , we place a prior on the model size $(k + i \times l_i)$ and given the model size, the prior is taken uniform among models with the same number of parameters. Since we want to identify a limited number of variables affecting CPI, we set $p(m) \sim \text{Poisson}(1)$. Furthermore, we present results (prior sensitivity analysis) under two alternative priors, i.e. Poisson with mean 5, and Beta-Binomial (Kohn et al., 2001) with mean and standard deviation 5.

4.2 The model selection algorithm

Given the large number of explanatory variables and the incumbent correlation structure, we need an effective inference tool to obtain the posterior summaries of interest. It is important to note that a typical Reversible Jump (RJ) algorithm (Green, 1995) we tested, fails to sample efficiently from the posterior, even after 10 million iterations (see Section 5 for further evidence). In this respect, we make use of recent techniques, namely the Subspace Carlin & Chib (SCC) algorithm (Petralias and Dellaportas, 2013), which is designed to sample efficiently under high dimensional spaces. Furthermore, we augment this algorithm to a population framework, so that several parallel chains of the SCC algorithm run under different temperatures, exchanging information in an efficient way.

In order to simplify the notation we: (a) Set each $Z_{i,t-l_i}$, for each lag l_i as a different variable, and denote the set of explanatory variables $X_{m,t} = \{S_{k,t}, Z_{i,t-l_i}\}$ for a model m , including κ out of p variables, with $p=356$ in our application; and $b_m = \{c, a_k, \theta_{i,l_i}\}$ the associated parameters to estimate. (b) Denote with $f(y|b_m, m)$ the (normal) likelihood and $f(b_m, m) = f(b_m|m)f(m)$ the prior densities. (c) Define pseudoprior densities $f(b_j | j \neq m)$ for parameters b_j not belonging to model m . These are not prior densities, but conveniently chosen linking densities required to generate parameters for variables not belonging to model m , that could be added in the next step of the algorithm. Thus, these are like proposal densities used in a RJ algorithm. In our application, considering all data are taken as monthly changes of indices, and after inspecting descriptive summaries of the variables, pseudopriors (proposals) are taken $f(b_j | j \neq m) \underset{iid}{\sim} N(0,1)$. We tested the algorithm using pseudoprior densities obtained from pilot Gibbs runs on each explanatory variable, and inflated their standard deviation by a factor of 2 or 3. The strategy turned out to provide equally well mixing ability of the algorithm, as the plain proposals (pseudopriors) suggested above.

We denote with $S_m = \{S_m^-, S_m^0, S_m^+\}$ the neighborhood of model m , including all models that can be formed by deleting, replacing or adding a variable to those present in the model, where $\{S_m^-, S_m^0, S_m^+\}$ represent the respective deletion, replacement and addition subspace. Denote with $Q_m = \{q_m^-, q_m^0, q_m^+\}$ the respective proposal probabilities for selecting each subspace. These are taken uniform in our application. Then the SCC algorithm proceeds as follows.

The SCC algorithm

- Generate parameters b_j from the full conditionals

$$f(b_j | b_{-j}, m, y) \sim \begin{cases} f(y | b_m, m) f(b_m, m), & j = m \\ f(b_j | j \neq m), & j \neq m \end{cases} \quad (2)$$

- Choose a subspace, say S_m^+ , with probability q_m^+
- Propose a new model m' with probability

$$g(m, m') = \frac{A_{m'}}{\sum_{s \in S_m^+} A_s} \quad (3)$$

where A_m is the posterior up to a constant

$$A_m = f(y | b_m, m) f(b_m, m) \prod_{j \neq m} f(b_j | j \neq m) \quad (4)$$

- Accept the move with probability

$$\alpha = \min \left(1, \frac{q_{m'}^- \sum_{s \in S_m^+} A_s}{q_m^+ \sum_{s \in S_{m'}^-} A_s} \right), \quad (5)$$

which is the ratio of the posterior mass of the proposed (addition) subspace, to the inverse (deletion) subspace. Thus via (3) we propose easily more probable models within a subspace, and the move is accepted if the subspace we are moving to has higher posterior mass than the inverse subspace; see (5). For the replacement move, we suggest using a simple RJ step, in view of the large number of densities A_m that require evaluation at every iteration (Petralias and Dellaportas, 2013).

The above algorithm is enriched by applying a population algorithm (see Jasra et al., 2007, for applications in the transdimensional space). We denote with π the SCC invariant distribution with states (b_m, m) , and construct tempered auxiliary distributions π^{ζ_λ} with ζ_λ denoting the inverse temperature parameter of chain λ . Considering the efficiency of the underlying SCC algorithm, we propose to use just three parallel chains, as in (Fouskakis et al., 2009): One with $\zeta=1$ being the central untempered sampling chain; one with $\zeta>1$ that jumps among models that have higher posterior density; and one with $\zeta<1$ (flatter density) that explores less probable models and improves the mixing ability of the algorithm. In our application we set $\zeta_1 = 0.8$, $\zeta_3 = 1.2$, so as for the acceptance rate of the exchange move to be about 0.5 (Liu, 2001).

The Population SCC algorithm

- Run $\lambda=3$ parallel Markov chains with target densities π^{ζ_λ} where $\zeta_1 < 1, \zeta_2 = 1, \zeta_3 > 1$
- At every iteration choose randomly between:
 - An exchange move which changes the states between a randomly chosen tempered chain with the central untempered chain. The Metropolis-Hastings acceptance rate used to swap the values of chain 1 with chain 2, is of the form

$$\alpha = \min \left(1, \frac{\pi^{\zeta_1}(\beta_{m_2}, m_2) \pi^{\zeta_2}(\beta_{m_1}, m_1)}{\pi^{\zeta_1}(\beta_{m_1}, m_1) \pi^{\zeta_2}(\beta_{m_2}, m_2)} \right) \quad (6)$$

A mutation move which updates all chains according to the SCC algorithm.

The Population SCC algorithm runs for 300,000 iterations with an additional 100,000 iterations being used for the burn-in period. It needed about 16 hours to run in an Intel Core i7 CPU. It is coded in Matlab and is available upon request from the first author.

5. RESULTS

Figure 1 displays the number of variables included in the posterior sample of the Population SCC algorithm under the different prior specifications. We observe that under the Poisson (1) prior the algorithm by and large visited models that include 5 to 10 variables, under the Beta-Binomial (5) prior models that (again) include 5 to 10 variables (but with higher frequency 6 variables), and under the Poisson (5) prior models with 8 to 17 variables (with higher frequency models that include 11 variables). This is typical in high dimensional models spaces with incumbent correlation structures, so that the actual model size is affected by prior beliefs. Nevertheless, this figure is associated only with the probability of the model size, which differs from the probability of specific models (shown in Table 2). That is, it is a usual phenomenon for model selection algorithms to visit often higher dimension spaces, but most probable models to include a relatively small number of variables.

Figure 1. Number of variables visited by the Population SCC algorithm under each prior specification

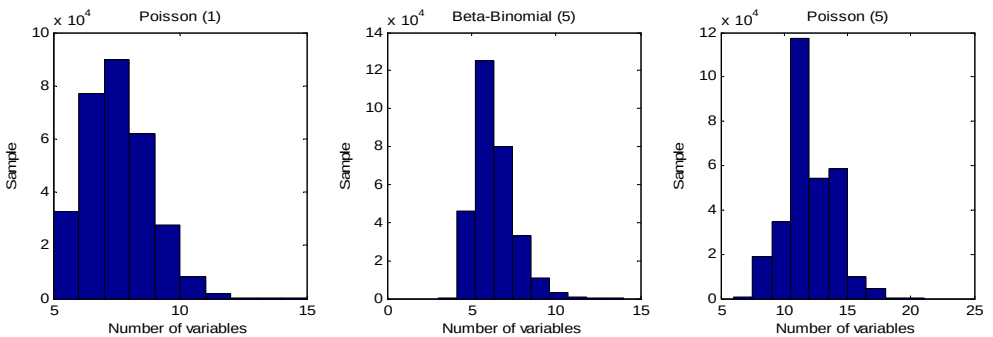


Table 2 presents the four most probable models visited under each prior specification, along with their parameter estimates, the posterior probability, and the Monte Carlo standard error. The Monte Carlo standard error, which is used to estimate the variability due to simulation, is calculated based on the batch mean method (Geyer , 1992), dividing the sample to $\sqrt{N}$ batches (each with size $\sqrt{N}$), where N denotes the total number of the iterations of the Population SCC algorithm.

Table 2: Posterior model probabilities (Monte Carlo standard error in parenthesis and mean values (*standard deviation in parenthesis*) of the posterior distribution of the parameters (*in the four best (highest posterior probability) models under each prior specification*

Prior distribution	Poisson (1)				Beta-Binomial (5)				Poisson (5)			
Model	First best	Second best	Third best	Fourth best	First best	Second best	Third best	Fourth best	First best	Second best	Third best	Fourth best
Probability % (Monte Carlo standard error)	6.10 (0.22)	4.71 (0.16)	1.23 (0.08)	1.12 (0.06)	25.98 (0.65)	11.90 (0.46)	3.86 (0.22)	3.31 (0.22)	0.069 (0.011)	0.065 (0.011)	0.047 (0.011)	0.046 (0.012)
Dummy February		-0.295 (0.079)	-0.339 (0.077)				-0.296 (0.079)		-0.344 (0.077)	-0.240 (0.075)	-0.287 (0.072)	-0.323 (0.071)
Dummy June	-0.800 (0.096)	-0.470 (0.072)	-0.484 (0.070)	-0.437 (0.077)	-0.800 (0.096)	-0.777 (0.103)	-0.470 (0.073)	-0.450 (0.082)	-0.493 (0.067)	-0.456 (0.071)	-0.467 (0.070)	-0.463 (0.064)
Dummy July		-0.521 (0.080)	-0.531 (0.078)	-0.449 (0.084)			-0.520 (0.080)	-0.400 (0.087)	-0.533 (0.077)	-0.512 (0.076)	-0.510 (0.076)	-0.503 (0.074)
Dummy October										0.230 (0.073)		
VAT	0.059 (0.013)	0.056 (0.013)	0.059 (0.013)	0.058 (0.014)	0.059 (0.013)		0.056 (0.013)		0.059 (0.013)	0.058 (0.013)	0.056 (0.013)	0.052 (0.012)
Sales	-1.634 (0.049)	-1.618 (0.049)	-1.598 (0.048)	-1.685 (0.048)	-1.634 (0.049)	-1.647 (0.052)	-1.617 (0.049)	-1.694 (0.052)	-1.597 (0.049)	-1.617 (0.045)	-1.618 (0.052)	-1.576 (0.043)
Sales after January 2011	-1.275 (0.114)	-1.243 (0.113)	-1.250 (0.111)	-1.236 (0.120)	-1.275 (0.114)	-1.345 (0.120)	-1.243 (0.111)	-1.307 (0.125)	-1.240 (0.111)	-1.197 (0.103)	-1.244 (0.098)	-1.359 (0.103)
Gasoline price index	0.039 (0.005)	0.044 (0.005)	0.044 (0.005)	0.041 (0.005)	0.039 (0.005)	0.043 (0.006)	0.044 (0.005)	0.045 (0.006)	0.045 (0.005)	0.045 (0.005)	0.044 (0.005)	0.047 (0.005)
Imports from the EU price index (lag 2)										0.295 (0.102)		
Dummy new government (lag 4)												-0.329 (0.108)
International Dairy price index (lag5)												0.013 (0.003)
Domestic Retail sales index (lag 6)	0.017 (0.003)				0.017 (0.003)	0.016 (0.003)						
Agricultural equipment and machinery price index (lag 9)			0.163 (0.053)						0.170 (0.055)			0.194 (0.051)
Constant	0.251 (0.023)	0.292 (0.025)	0.268 (0.025)	0.259 (0.024)	0.251 (0.023)	0.265 (0.024)	0.292 (0.025)	0.272 (0.026)	0.266 (0.024)	0.222 (0.030)	0.291 (0.023)	0.260 (0.022)
Error std.dev.	0.231	0.228	0.220	0.241	0.231	0.249	0.228	0.257	0.220	0.214	0.228	0.202

The first best model visited under the Poisson (1) and Beta-Binomial (5) priors is the same, featuring a rather large posterior probability (6.1% and 26% respectively)

considering the dimension of the model space (2^{356}). The model includes (i) a monthly dummy variable which is associated with a seasonal CPI drop in June, (ii) changes in the VAT, (iii) the discount sales periods, the coefficient of which suggests a CPI drop of about 1.6, and an even larger drop (about 2.9%) after the CPI's revision of January 2011 (iv) changes in the gasoline price index, which appear to affect the CPI by about 4%, and (v) changes in the domestic index of retail sales, which appear to affect the CPI with a 6-month lag by about 1.7%.

In view of these results the following comments are in order. The coefficient associated to VAT is up to 0.59 in most probable models. Note that the VAT series is also converted to an index the base period of which is set to 2009, and the average VAT rate weighted across all products, services and regions, was equal to 12.2% in 2009. (The corresponding rate in 2012 was 16.6%). Thus, an increase to the VAT rate by 1%, from 12.2% to 13.2% in terms of the 2009 average, leads to an increase of 0.48% in the CPI. Likewise, a potential increase from 16.6% to 17.6%, is expected to raise the CPI by 0.44%. This suggests that, *ceteris paribus*, a significant portion (over 50%) of VAT rate changes is absorbed by the producers.

Furthermore, it seems that the discount sales period effect is larger after January 2011. However, it is unclear whether this is due the CPI's revision or the deepening of the economic crisis, or both. The relatively small effect associated with domestic retail sales reflects considerable price pressures with respect to changes in demand. For example, a 10% fall in retail sales is expected to reduce the CPI only by 0.17%, six months later.

Considering the variables included in subsequent models, it is interesting to note the large effect of EU-imports prices on the CPI (by 29.5%, according to the second best model under Poisson (5) prior), and the negative impact of (new) government formation on the CPI four months after swearing in (by 0.33% according to the fourth best model under the Poisson (5) prior).

A measure usually employed in order to help identify important variables in high dimensional spaces is the marginal probability of inclusion (see Barbieri and Berger, 2004). This is calculated as the percentage of times a variable is observed in the transdimensional MCMC sample. If we denote with $\hat{f}(m|y)$ the estimated posterior probability of a model m , then the probability of inclusion of a variable i is given as

$$p_i = \sum_{m \in M^i} \hat{f}(m|y)$$

where M^i denotes all models visited by the algorithm in which variable i is present.

Table 3 lists the variables according to their marginal probability of inclusion under each prior specification. The fourteen variables with the highest probability of inclusion are the same under the Poisson (1) and Beta-Binomial (5) priors, and also feature among the top 20 variables visited under the Poisson (5) prior and in the four best models of Table 2 alongside the index of fertilizer prices (lag 5). Therefore, there

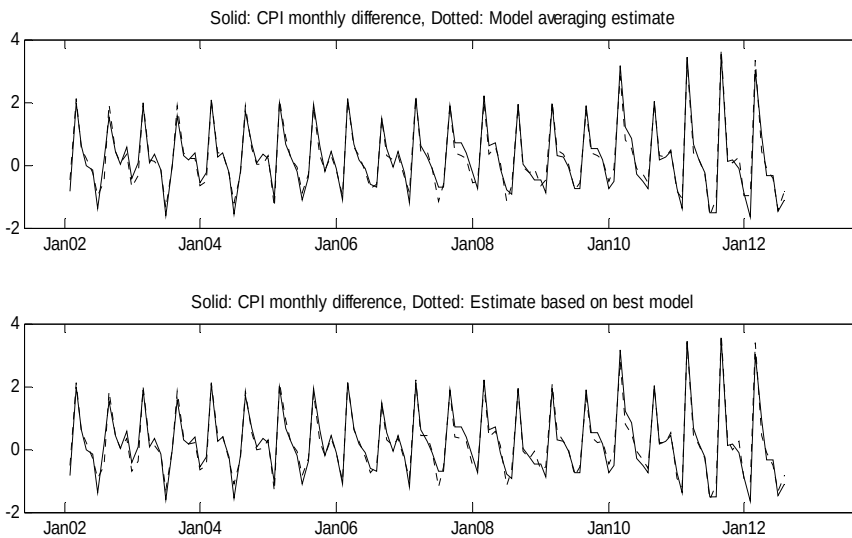
is considerable confidence that they are indeed among the most probable factors likely to affect the CPI, with the discount sales periods, the gasoline price index, and the VAT appearing as the primary factors.

Table 3: *Marginal probability of inclusion of variables under each prior specification. Presented are the top twenty variables visited under each prior specification, irrespectively of model*

Poisson (1)			Beta-Binomial (5)			Poisson (5)		
Variable	Probability	Monte Carlo standard error	Variable	Probability	Monte Carlo standard error	Variable	Probability	Monte Carlo standard error
Discount sales (DS)	1.000	0.000	Discount sales (DS)	1.000	0.000	Discount sales (DS)	1.000	0.000
DS after January 2011	1.000	0.000	DS after January 2011	1.000	0.000	DS after January 2011	1.000	0.000
Index of gasoline prices	1.000	0.000	Index of gasoline prices	1.000	0.000	Index of gasoline prices	1.000	0.000
June dummy	0.991	0.002	June dummy	1.000	0.000	June dummy	0.997	0.001
VAT	0.824	0.009	Index retail sales (lag 6)	0.691	0.012	VAT	0.908	0.007
July dummy	0.615	0.009	VAT	0.640	0.010	July dummy	0.891	0.009
Index retail sales (lag 6)	0.391	0.009	July dummy	0.310	0.011	February dummy	0.608	0.011
February dummy	0.389	0.008	February dummy	0.156	0.006	Ind. agric. eq. etc. (lag 9)	0.359	0.010
Index EU imports (lag 2)	0.137	0.005	Index EU imports (lag 2)	0.050	0.002	October dummy	0.348	0.010
October dummy	0.130	0.005	New gov. dummy (lag 4)	0.049	0.001	New gov. dummy (lag 4)	0.296	0.007
New gov. dummy (lag 4)	0.097	0.003	October dummy	0.035	0.002	Index EU imports (lag 2)	0.269	0.009
Ind. agric. eq. etc. (lag 9)	0.088	0.004	Index fertilizer pr. (lag 5)	0.030	0.002	August dummy	0.161	0.007
Int'l dairy index (lag5)	0.054	0.004	Int'l dairy index (lag5)	0.021	0.002	Index retail sales (lag 6)	0.118	0.009
Index fertilizer pr. (lag 5)	0.042	0.003	Ind. agric. eq. etc. (lag 9)	0.021	0.001	DS after January 2005	0.109	0.005
New gov. dummy (lag 7)	0.038	0.001	Ind. train fares (lag 7)	0.020	0.002	Int'l dairy index (lag5)	0.108	0.007
Ind. buildings etc. (lag 3)	0.038	0.002	Int'l cereal pr. Ind. (lag5)	0.019	0.001	Ind. seed prices (lag 8)	0.104	0.006
Ind. seed prices (lag 8)	0.037	0.002	New gov. dummy (lag 7)	0.016	0.001	New gov. dummy (lag 7)	0.103	0.003
New gov. dummy (lag 3)	0.033	0.001	Ind. buildings etc. (lag 3)	0.014	0.001	Ind. buildings etc. (lag 3)	0.101	0.005
Ind. train fares (lag 7)	0.033	0.003	Ind. seed prices (lag 8)	0.013	0.001	April dummy	0.098	0.005
Int'l cereal pr. Ind. (lag5)	0.032	0.003	New gov. dummy (lag 5)	0.013	0.001	Index fertilizer pr. (lag 5)	0.096	0.006

Figure 2 displays the CPI monthly changes estimates based (a) on all models visited by the algorithm, through model averaging, and (b) on the best model, under the Poisson (1) prior. In both cases the estimates seem to lie close to the observed CPI monthly changes, with similar Root Mean Squared Error (RMSE) (equal to 0.2273 under the model averaging approach and to 0.2286 when estimates are based only on the best model). Similar results are obtained under the other prior specifications.

Figure 2. *Model fit: CPI monthly changes estimates based on all models (model averaging) and the best model; Poisson (1) prior; the high seasonality is primarily related to sales periods*

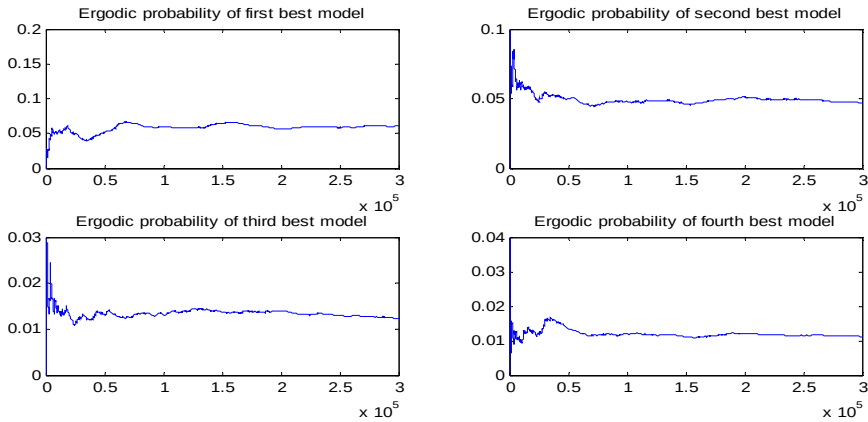


The effectiveness and accuracy of the SCC algorithm, compared to several widely used model selection methods, have been demonstrated in Petralias and Dellaportas (2013). Thus, here we only comment on the need to use such a technique compared to a standard sampler. As depicted in Figure 3, a standard Reversible Jump (RJ) sampler fails completely to converge even after 10 million iterations and displays significant drift in the ergodic probability estimates of most probable models, whereas the Population SCC algorithm seems to sample efficiently within only 300 thousand iterations. Furthermore, the Monte Carlo standard error for the four best models was significantly higher for RJ, i.e. equal to 0.0031, 0.0023, 0.0010 and 0.0011, as opposed to 0.0022, 0.0016, 0.0008 and 0.0006 respectively, for Population SCC (Poisson (1) prior). Similar differences observed for other prior specifications signal higher accuracy in model probability estimates. The enhanced mixing ability of the Population SCC algorithm is also depicted in the multiple times higher

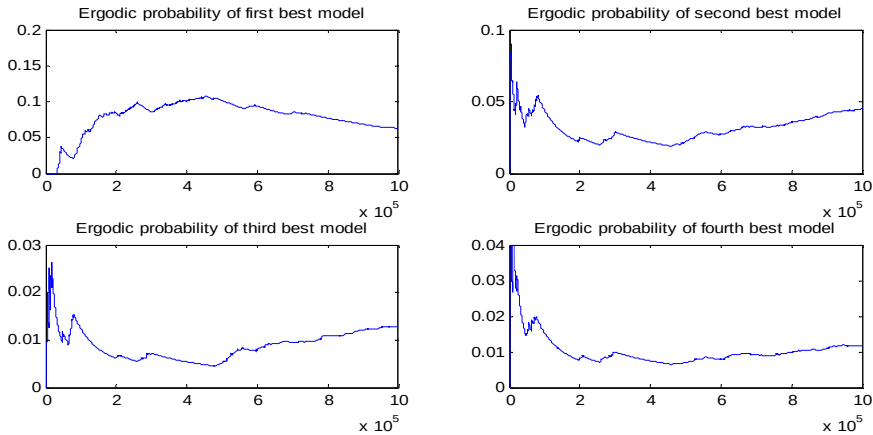
(addition/deletion) acceptance rate which was equal to 0.569 (central chain; Poisson (1) prior), as opposed to 0.037 for RJ.

Figure 3. Ergodic probability plots of the four best models estimated under the Population SCC and the RJ algorithm; displayed is the ergodic probability of each model across the sample obtained from each algorithm; Poisson (1) prior

Population SCC (300 thousand iterations without thinning)



Reversible Jump (RJ) (10 million iterations; thinning per 10 iterations)



6. CONCLUSIONS AND POLICY IMPLICATIONS

Two of the most important factors identified as likely to affect prices are the VAT rates and the gasoline price index. Note that the weighted average (across products and regions) VAT rate has increased from 12.2% in 2009 to 16.6% in September 2011 and remains stable thereafter. The Gasoline price index has increased by approximately 80% since 2009, mostly because of taxes, so that in 2012 more than

60% of the retail gasoline price was related to taxes. By isolating the effect of the VAT and gasoline price changes, one can easily infer that the CPI in August 2012 lies close to the 2009 average. Furthermore, prices seem to drop considerably during the discount sales periods, more so in recent years, thus affecting increased seasonal variability in the CPI around sales periods. The CPI seems to be marginally affected by changes in retail sales, with a significant lag (6 months), while the prices of imports appear to have a significant effect, which suggests that the Greek market depends extensively on imports and less on domestic production.

In view of the above, the primary measures that could help reduce prices include (a) the expansion of sales periods; (b) the reduction of the VAT; (c) lowering indirect fuel taxes while stepping up fuel market checks; (d) bringing down dependence on imports while raising domestic production.

ΠΕΡΙΛΗΨΗ

Επιδιώκουμε να προσδιορίσουμε τους παράγοντες που επιδρούν στον Δείκτη Τιμών Καταναλωτού της Ελλάδος, λαμβάνοντας υπ' όψιν ένα μεγάλο σύνολο δυνητικών παραγόντων και τις χρονικές υστερήσεις αυτών. Προς τούτο στηρίζομαστε σε πρόσφατες μεθόδους μπεϋζιανής επιλογής υποδειγμάτων, οι οποίες προσαρμόζονται αποτελεσματικά στο υπόδειγμα. Οι τιμές μειώνονται σημαντικά κατά τις περιόδους των εκπτώσεων, ειδικά κατά τα τελευταία δύο έτη, οι αλλαγές του ΦΠΑ απορροφούνται σε μεγάλο ποσοστό (άνω του 50%) από τους πωλητές, ενώ μια μείωση στις λιανικές πωλήσεις επηρεάζει τις τιμές με σημαντική υστέρηση (6 μηνών). Επίσης παρατίθενται στοιχεία για τον τρόπο με τον οποίο αρκετοί παράγοντες επηρεάζουν τις τιμές.

REFERENCES

- Alemou Z.G. (2012). Causality links between consumer and producer price inflation in South Africa. *Applied Economic Letters*, **19**: 13-18.
- Askari H. and N. Krichene (2010). Monetary policy and world commodity markets: 2000-2007. *PSI Quarterly Review*, **63**: 145-177.
- Barbieri, M. M., and Berger, J. O. (2004). Optimal predictive model selection. *The Annals of Statistics*, **32**, 870–897.
- Fouskakakis, D., Ntzoufras, I. and Draper, D. (2009). Population-Based Reversible Jump MCMC for Bayesian Variable Selection and Evaluation Under Cost-Limit Restrictions. *Journal of the Royal Statistical Society C (Applied Statistics)*, **58**, 383-403.
- Geyer C.J. (1992). Practical Markov Chain Monte Carlo (with discussion). *Statistical Science*, **7**, 473-511.
- Green P.J. (1995). Reversible jump MCMC computation and Bayesian model determination, *Biometrika*, **82**, 711–732.

- Jail M. and E. Tamayo Zea (2011). Pass-through of International Food Prices to Domestic Inflation During and After the Great Recession: Evidence from a Set of Latin American Economies. *Desarrollo y Sociedad*, **67**: 135-179.
- Jasra A., Stephens D.A., and Holmes C.C. (2007). Population-based reversible jump Markov chain Monte Carlo, *Biometrika*, **94**, 787–807.
- Kamenik O., H. Kiem, V. Klyuev and D. Laxton (2013). Why Is Canada's Price Level So Predictable? *Journal of Money, Credit and Banking*, **45**: 71-85.
- Kohn R, Smith M., and Chan D. (2001). Nonparametric regression using linear combinations of basis functions, *Statistics and Computing*, **11**, 313–322.
- Liu, J. S. (2001). *Monte Carlo Strategies in Scientific Computing*, New York: Springer.
- Petralias, A. and Dellaportas, P. (2013). An MCMC model search algorithm for regression problems. *Journal of Statistical Computation & Simulation*, **83**: 1722-1740.



ON THE JOINT ESTIMATION OF INFORMATION BETAS

V. Polimenis, I. Papantonis

Aristotle University of Thessaloniki
{polimenis, papantonis}@econ.auth.gr

ABSTRACT

This paper enhances a co-skew based risk measurement methodology initially introduced in Polimenis (2012), by extending it for the joint estimation of the information betas for two stocks. We introduce the possibility of idiosyncratic jumps and analyze the robustness of the estimated sensitivities when two stocks are jointly fit to the same set of latent jump factors. We argue that, when individual stock skews substantially differ from those of the market, the requirement that the individual skew is exactly matched is placing a strain on the single stock estimation system. Once we relax this exact match requirement, and essentially allow for the presence of idiosyncratic stock skew, the system moves to a more robust solution.

Keywords: information betas; asymmetric betas; price jumps; Lévy process; return skewness.

1. INTRODUCTION

The classical beta arises as the proper measure of risk only under very strict conditions that are now under heavy questioning. Yet, it is well accepted that investors care not only about variance, but also about return skewness. Since Kraus and Litzenberger (1976), Friend and Westerfield (1980) and, more recently, Harvey and Siddique (2000), co-skewness with market returns is proven to be a significant factor driving asset prices. Moreover, there is evidence of asymmetric correlations among financial returns that has triggered the focus on asymmetric response betas and downside risk. Markowitz (1959), Hogan and Warren (1974), Bawa and Lindenberg (1977), Fishburn (1977), Harlow and Rao (1989), Ang, Chen, and Xing (2006), Post and Van Vliet (2006), Estrada (2007) and others have suggested the use of lower-partial moments and, as a special case, semivariance, in order to proxy downside risk. A more detailed presentation of downside risk measures can be found in Sortino and Satchell (2001).

Following Merton (1976), academics have argued in favor of allowing for discontinuities in the price paths of financial assets, casting doubts on the validity of asset pricing models (such as the classical model of Black & Scholes, 1973) that assume continuity for the return dynamics. Investors are exposed to jump-risks and hence demand higher compensation for bearing the possibility of an unfavorable extreme event (Pan, 2002). A

recent empirical finding by Carr et al. (2002) uncovers the existence of jumps in financial markets and indicates that these jumps may not only be large and infrequent Poisson jumps, but also smaller jumps that arrive at an infinite rate. Aït-Sahalia and Jacod (2009) draw a similar conclusion by analyzing the degree of jump activity in high-frequency data. Jumps are indeed prevalent in financial asset prices and these jumps are also responsible for introducing asymmetries in the distributions of returns (Carr & Wu, 2004).

A natural and parsimonious model, that is also general enough so that it encompasses most empirical findings, is one where (a) stock returns are exposed to both diffusive and jump risks, (b) jump arrivals of various sizes may occur at low and high frequencies, (c) stock returns may exhibit differential sensitivities to diffuse vs. jump market information and (d) jump sensitivities are potentially asymmetrical. In this framework, systematic jumps are random points in time where market-wide information arrival occurs, while price continuity is preserved by very frequent noise trading (Black, 1986).

Market returns include both continuous and discontinuous components. If the market treats asymmetrically price discontinuities from smooth variation, such a decomposition is of utmost importance for asset pricing. There is still no simple statistical procedure for estimating jump betas. A recent strand of the literature is concerned with analyzing high-frequency data (see Aït-Sahalia (2004), Aït-Sahalia and Jacod (2009), Barndorff-Nielsen and Shephard (2004a)). These procedures aim to measure realized power and bipower variation (introduced by Barndorff-Nielsen & Shephard (2004b)), in order to decompose total quadratic variation into its continuous and jump components. Todorov and Bollerslev (2010) utilize this framework in order to disentangle the systematic continuous and discontinuous parts and estimate the sensitivities of individual stocks towards these risk factors. The problem with the above techniques is that they rely on the availability of extremely fine data, i.e. an almost continuous time record, which is likely to contaminate data with noise (such noise arises from bid-ask bounce, price discreteness, non-synchronicity in trades etc). Skewness and jumps are generally related to illiquidity and market frictions and hence high-frequency techniques will introduce complications in exactly the situations where higher moments are expected to heavily impact asset prices.

In order to fill this void and develop a practically and academically sound framework, Polimenis (2012) proposes the use of a set of co-skew based moment conditions for disentangling and estimating asymmetric sensitivity towards systematic jump risks. This approach boosts the “signal-to-noise” ratio by utilizing co-skew moments, so that the diffusive component is filtered out through higher-order cumulants. Since cumulants in a Lévy process evolve linearly in time, this approach is horizon independent and hence can be deployed at all frequencies.

In this paper we build on this approach by focusing on a joint Generalized Method of Moments (GMM) estimation, at the same time introducing the possibility of idiosyncratic stock skewness. A joint estimation improves the estimation robustness of the true size and impact of the latent market jump risk factors. Allowing for idiosyncratic skews

relaxes the demands placed on the estimation system, and hence improves its fit by focusing on matching market skew that is more informational. Besides, allowing for stock jumps that are not related to the market is a realistic assumption. In the next section we present the model framework. We then explain about the estimation procedures for the single-stock and the joint cases. Finally, we discuss the results regarding the latent skew parameters and their sensitivities, highlighting the importance of the idiosyncratic stock skews captured by the system.

2. FRAMEWORK

Let $y(t)$ and $z(t)$ denote two latent market information factors (i -factors), respectively for positive and negative market news, that are significant enough to produce price jumps. Also, define γ and δ to be the sensitivities to these jumps, which are incorporated in an individual asset's returns through:

$$R_i(dt) = \alpha_i dt + \beta_i \sigma_m dW_t + \gamma_i y(dt) - \delta_i z(dt) + \eta_i dW_{i,t} + u_i(dt),$$

where α is a deterministic drift, β a classical (CAPM-like) systematic diffuse coefficient and $(W, W_1, \dots, W_N)$ a standard Brownian motion vector with independent components. The stock price includes jumps introduced either systematically by the pure jump processes y and z or individually by the idiosyncratic corporate jump risk u . Jumps in y are counted via the Poisson random measure μ_y and arrive at a rate $\lambda_y(dy)$, similarly for jumps in z and u . All jump arrivals are independent. Our interest here is primarily in the jump coefficients γ and δ , that regulate the transmission of the systematic jump risk factors y and z . Market returns lack an idiosyncratic component and are driven by:

$$R_m(dt) = \alpha_m dt + \sigma_m dW_t + y(dt) - z(dt).$$

Total risk decomposition into its three latent components is not directly observable. Jumps cannot be safely disentangled from the continuous counterpart, thus properly quantifying the magnitude and significance of the various systematic risk components is not a trivial task, especially if we are unwilling to impose a heavy distributional structure on the jump dynamics. Jumps induce skewness in the return distribution, and hence one may naturally proxy jumps with skewness. Since all higher-order cumulants of the diffuse part are zero, one can easily filter out the diffusive component of the individual stock price variation (Polimenis, 2012). By utilizing the affine-transformation property of cumulants we derive the following higher-order moment conditions:

$$\begin{bmatrix} \sigma_{mmm} \\ \sigma_{imm} \\ \sigma_{iim} \\ \sigma_{iii} \end{bmatrix} = \begin{bmatrix} \sigma_{yyy} - \sigma_{zzz} \\ \gamma_i \sigma_{yyy} - \delta_i \sigma_{zzz} \\ \gamma_i^2 \sigma_{yyy} - \delta_i^2 \sigma_{zzz} \\ \gamma_i^3 \sigma_{yyy} - \delta_i^3 \sigma_{zzz} \end{bmatrix} \quad (1)$$

where σ_{yyy} and σ_{zzz} are the skew cumulants of the latent positive and negative jump factors, respectively. We use σ_{ijk} for the 3rd cumulant, where $\sigma_{ijk} = \mathbb{E}[(R_i - \mu_i)(R_j - \mu_j)(R_k - \mu_k)]$. The notation σ_{imm} and σ_{iim} refers to the two co-skews that relate stock returns with market returns. Finally, σ_{iii} is the individual stock skew and σ_{mmm} denotes

market skew. By expanding this moment condition framework we are able to incorporate more stocks, but this is only possible at the expense of increasing the dimensionality of the problem. Yet, these moment conditions are highly non linear, and thus the estimation method is easy to implement only for the single-stock case, where a single solution may be found that exactly matches all moment conditions. We aim to build on this approach by expanding it to a two-asset framework, focusing on a joint GMM estimation.

The motivation of our analysis stems from the fact that the implied upward and downward market skews may vary when estimated solely from one stock and the market. We confirm this intuition by utilizing the exact single-stock closed-form solution, applying it separately for Google and Yahoo. We observe that there is a wedge between the implied latent skew factors estimated by the two different stocks. This spurs the necessity of a joint multi-asset estimation in order to extract information about the level of the latent market skew factors that is less dependent on the data selection.

3. A JOINT GMM ESTIMATION FRAMEWORK

We may naturally enhance the system to incorporate multiple assets. By including an additional stock to the optimization scheme, we are increasing the dimensionality of the problem. Let us consider a joint system of two stocks and the market. Expanding and rearranging the terms in (1) yields the following non-linear sample moment conditions:

$$\mathbb{E}[\mathbf{h}_t(\theta_0)] = \mathbb{E} \left[\begin{array}{c} (R_{m,t} - \mu_m)^3 - (\sigma_{yyy} - \sigma_{zzz}) \\ (R_{i,t} - \mu_i)(R_{m,t} - \mu_m)^2 - (\gamma_i \sigma_{yyy} - \delta_i \sigma_{zzz}) \\ (R_{i,t} - \mu_i)^2(R_{m,t} - \mu_m) - (\gamma_i^2 \sigma_{yyy} - \delta_i^2 \sigma_{zzz}) \\ (R_{i,t} - \mu_i)^3 - (\gamma_i^3 \sigma_{yyy} - \delta_i^3 \sigma_{zzz}) \\ (R_{j,t} - \mu_j)(R_{m,t} - \mu_m)^2 - (\gamma_j \sigma_{yyy} - \delta_j \sigma_{zzz}) \\ (R_{j,t} - \mu_j)^2(R_{m,t} - \mu_m) - (\gamma_j^2 \sigma_{yyy} - \delta_j^2 \sigma_{zzz}) \\ (R_{j,t} - \mu_j)^3 - (\gamma_j^3 \sigma_{yyy} - \delta_j^3 \sigma_{zzz}) \end{array} \right] = 0 \quad (2)$$

where $\theta = (\sigma_{yyy}, \sigma_{zzz}, \gamma_i, \delta_i, \gamma_j, \delta_j)$ is an L -parameter vector that satisfies the orthogonality conditions and $\mathbf{h}_t(\theta)$ is a $K \times T$ matrix that represents the sample moment errors for each of the K moment conditions. Denote by $\mathbf{g}_T(\theta)$ the $K \times 1$ vector of the average sample moment errors for an arbitrary vector θ of parameters, i.e. $\mathbf{g}_T(\theta) = 1/T \sum_{t=1}^T \mathbf{h}_t(\theta)$, with $\mathbf{h}_t(\theta)$ being a stationary and ergodic process. The key concept of the GMM lies within the minimization of the objective function $\hat{\theta}(\mathbf{W}) = \arg \min_{\theta} J(\theta, \mathbf{W})$, where $J(\theta, \mathbf{W}) = T \mathbf{g}_T(\theta)' \mathbf{W} \mathbf{g}_T(\theta)$ and $\mathbf{W}$ is a $K \times K$ positive definite weighting matrix.

According to Hansen (1982) an efficient GMM estimator results if we set $\mathbf{W} = \hat{\mathbf{S}}^{-1}$, where $\hat{\mathbf{S}}^{-1}$ represents the inverse of a consistent estimate of the asymptotic variance-covariance matrix $\mathbf{S} = \text{avar}(\mathbf{g})$ of the average sample moment error vector $\mathbf{g}_T(\theta)$. If the specified moment conditions are valid, then as $T \rightarrow \infty$, the objective value J should follow a χ^2 distribution $J \sim \chi^2(K - L)$, the degrees of freedom being equal to the number of overidentifying restrictions $K - L$. Note that, unlike the single-stock case, system (2) in the joint GMM estimation will be overidentified since $K > L$.

To account for the possibility of heteroscedasticity and serial correlation in the sample moment error vector we estimate the long-run variance-covariance matrix with the HAC estimation procedure suggested by Newey and West (1987). The estimator of $\mathbf{S}$ has the form $\hat{\mathbf{S}} = \hat{\mathbf{\Omega}}_0 + \sum_{j=1}^m w_j(\hat{\mathbf{\Omega}}_j(\hat{\theta}) + \hat{\mathbf{\Omega}}_j'(\hat{\theta}))$, with $\hat{\mathbf{\Omega}}_j(\hat{\theta}) = 1/T \sum_{t=j+1}^T \hat{\mathbf{h}}_t(\hat{\theta})\hat{\mathbf{h}}_{t-j}'(\hat{\theta})$, where w_j is a continuous weighting function (*kernel*) $k(\cdot)$, used to weight lagged variance-covariance matrices and m is the smoothing parameter (*bandwidth*). According to the Bartlett kernel specification the weights are defined as $k(j, m) = 1 - [j/(m + 1)]$. We use the Andrews (1991) methodology to set the optimal level of bandwidth m^* for the Bartlett kernel.

4. DISCUSSION ON THE RESULTS

4.1 Single-stock results

In order to compare with Polimenis (2012) we use the same 3-year sample (2006-2008) of adjusted daily returns for Google, Yahoo and Nasdaq. We start by estimating the exact-identified system for each individual stock separately. As expected, the estimated parameters perfectly fit the moment conditions. For Google, the estimated skews for the upside and downside jumps are $\hat{\sigma}_{yyy} = 0.108 \times 10^{-6}$ and $\hat{\sigma}_{zzz} = 1.021 \times 10^{-6}$, respectively. The estimated coefficients to these factors are $\hat{\gamma} = 4.13$ and $\hat{\delta} = 0.98$, which suggests significantly asymmetrical exposure to systematic jumps. These asymmetries are even more pronounced for Yahoo, where we estimate $\hat{\gamma} = 15.75$ and $\hat{\delta} = 1.85$, the corresponding implied skew parameters being $\hat{\sigma}_{yyy} = 0.011 \times 10^{-6}$ and $\hat{\sigma}_{zzz} = 0.924 \times 10^{-6}$. Despite their negative returns, both stocks (and especially Yahoo) are highly exposed to upward market jumps, even if these jumps are less powerful than the negative ones.

Table 1: Descriptive Statistics.

	$\mu_i^{(**)}$	σ_i	$\sigma_{iii}^{(*)}$	σ_{iii}/σ_i^3	$\sigma_{imm}^{(*)}$	$\sigma_{iim}^{(*)}$
Market	-4.67 bp	1.71%	-0.9132	-0.1826		
Google	-4.60 bp	2.50%	6.6406	0.4249	-0.5556	0.8585
Yahoo	-16.05 bp	3.40%	36.0750	0.9220	-1.5379	-0.4926

(*) All 3rd raw cumulants have been scaled up by 10^6 .

(**) 1 basis point (bp) = 10^{-4}

The basic idea of our approach is to relax the assumption underlying (1) that all stock skew is systematic, and allow for idiosyncratic stock skews. In this context, our analysis can be thought as a minimization of the unexplained *idiosyncratic* stock skewness. Thus, we maintain the exact fit for the market skew and the co-skew moments, since market jumps are by definition systematic, but relax the fit for the individual stock skew to include both a systematic and an idiosyncratic component. We allow the jump cumulants to vary and plot the resulting (sub-optimal) J -values. On every iteration we fix the value of the positive jump cumulant σ_{yyy} . But since we demand all market skew to be systematic, the negative jump cumulant can be directly estimated through $\sigma_{zzz} = \sigma_{yyy} - \sigma_{mmm}$. Technically, by imposing these constraints, we damage the overall fit of the system by

cancelling 3 degrees of freedom. Yet, and since the original system in (1) is exactly identified, the optimal results of the restricted estimation are identical to those of the unrestricted model that we mentioned above. However, this exercise provides a better insight to the sensitivity of the coefficients to changes in the underlying systematic skew parameters. Figures reveal that the results are very sensitive to changes in σ_{yyy} , which is due to the highly nonlinear nature of the system equations.

Figure 1: Google *i*-gamma

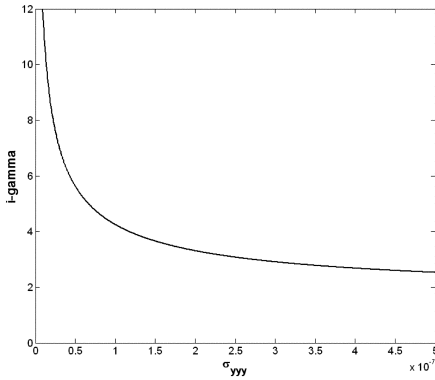
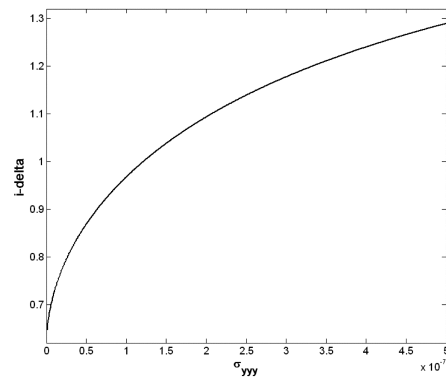


Figure 2: Google *i*-delta



The graphs above reveal that both *i*-factors are strictly positive, with *i*-gammas being negatively sloped and convex. On the other hand *i*-deltas exhibit a positive and concave relation with the upward skew parameter. The explanation of these plots is not trivial; the key to understand them is to realize that we need to calibrate the system to values that turn a negative market skew into a strongly positive stock skew. In the absence of idiosyncratic skew, the strain of switching the skew sign is placed on the informational sensitivities γ and δ . This means that for small upside jump risks, we need high *i*-gamma sensitivities to match the strong positive stock skews, given the fact that the market is negatively skewed.

The analysis of Yahoo reveals a higher sensitivity to the level of jump risks, since its coefficients to the *i*-factors change sharply with changes of the latent skew parameters. The objective *J*-value for Yahoo is minimized at $\hat{\sigma}_{yyy} = 0.011 \times 10^{-6}$, the remaining implied parameter values also being the same as in the unrestricted estimation. Again, this result is statistically significant when compared to the critical value of the χ^2 -distribution at the 1% level of significance and 1 degree of freedom. Note that the error in fitting individual stock skewness is a positively sloped and concave function with respect to σ_{yyy} . This is important because by underestimating the magnitude of positive jump risks in the market, we may end up severely overestimating individual stock skewness.

4.2 Joint estimation for two stocks

We observe a wedge between the values of the implied market skew parameters when estimated by the two different stocks. When fitting Yahoo, the positive skew cumulant is approximately $\hat{\sigma}_{yyy} = 0.011 \times 10^{-6}$, smaller than the respective quantity we get by fitting Google. This becomes clear in Figure 3, where we see that the Yahoo plot attains its minimal value far to the left of the Google plot. The magnitude of this difference is the reason why a joint estimation is required. Results from an unrestricted joint GMM estimation suggest that the upward jump skew factor is equal to $\hat{\sigma}_{yyy} = 0.10 \times 10^{-6}$. Similarly, the estimated optimal skew for the downward jumps is $\hat{\sigma}_{zzz} = 0.73 \times 10^{-6}$. As expected, the unrestricted joint model is unable to capture total market skewness. The implied market skewness based on the estimated skew parameters is $\hat{\sigma}_{mmm} = -0.63 \times 10^{-6}$. This means that a level of -0.28×10^{-6} of systematic skew remains unexplained. It is reasonable to find a more pronounced risk factor associated with negative news than for positive news, since during the period that we are studying the market was dominated by downward jumps that induced significantly negative skewness in its return distribution.

We perform a GMM using a Bartlett kernel in order to compute a weighting matrix that is robust to heteroskedasticity and autocorrelation (HAC). The J -statistic is 0.048, with a p -value of 83% at 1 degree of freedom (the J -statistic is quite robust to the selection of the smoothing kernel method used during the estimation of the asymptotic variance-covariance matrix of the sample error vector). The estimated exposures of Google (stock i) to the upward and downward jumps are $\hat{\gamma}_i = 4.25$ and $\hat{\delta}_i = 0.84$. These values are close to those estimated in the exactly identified single-stock case. This is not the case for Yahoo (stock j), since we find $\hat{\gamma}_j = 7.49$ and $\hat{\delta}_j = 2.89$, due to the implied upward skew parameter being higher than when estimated solely by Yahoo. The fitted skew of Google is 7.31×10^{-6} , which means that the model slightly overestimates the skewness of Google by 0.66×10^{-6} , implying a negative idiosyncratic skew for Google. For Yahoo, the skew fitted by the system is 25×10^{-6} , hence a level of 10.5×10^{-6} of positive skew remains unexplained by the systematic factors. This suggests a significantly positive idiosyncratic jump risk factor for this stock.

In order to capture total systematic market skew and further focus on idiosyncratic skew, we apply a restricted estimation. We require an exact match for the market skew, and all the co-skews. The remaining moments refer to the individual stock skews σ_{iii} and σ_{jjj} . We have a 6-parameter vector but require 5 exact relations, hence there is only 1 d.f. to match the two remaining moments. The restricted optimization of the system shows that the upward skew slightly drops to 0.084×10^{-6} , hence demanding a higher level for the downward jump factor of approximately 0.997×10^{-6} . At 1 d.f., the corresponding p -value for the minimized objective J -statistic (0.0657) is 80%. Results for Google are close to those estimated in the single-stock case, and specifically the values for $\hat{\gamma}$ and $\hat{\delta}$ are 4.56 and 0.94. The optimized parameters for Yahoo are 6.91 and 2.12, slightly lower compared to their unrestricted estimates. Results are less sensitive for Google, which is

visible if we compare the J -statistic functions for both individual and joint cases.

Figure 3: Joint case; J -values*

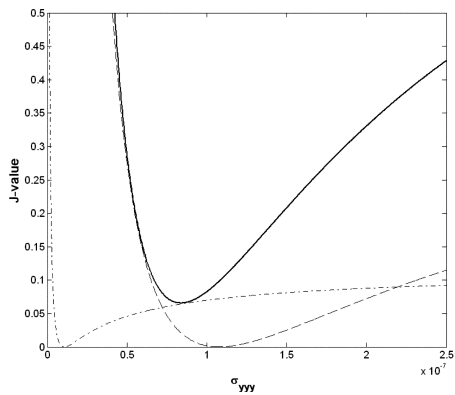
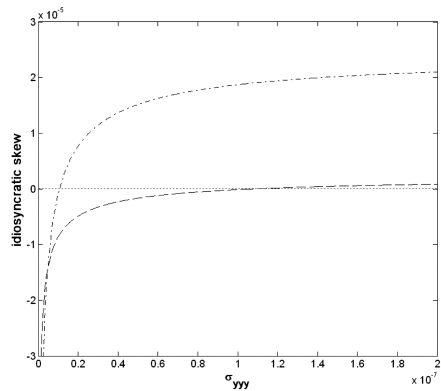


Figure 4: Joint case; idiosyncratic skews*



*The solid line is for the joint estimation; the dashed line for Google; and the dash-dot line for Yahoo.

The requirement that the extremely high positive Yahoo skew is exactly matched, together with a small but negative market skew, places an exorbitant strain on the single stock system for Yahoo. This requirement can only be met at a very high γ value for Yahoo and a very low positive market skew (σ_{yyy}). Once we relax this exact match requirement, and essentially allow for the presence of idiosyncratic stock skew, the system calibrates to a more robust solution. The model still overestimates the skewness of Google by 0.46×10^{-6} (lower than the overestimation found in the unrestricted case), but this is only in the expense of capturing an even lower percentage of the total skewness of Yahoo in order to improve the fit for Google. Generally, the idiosyncratic jump risk factor is much more intense for the case of Yahoo, since there is a significantly non-zero skew residual term that cannot be explained by the two systematic jump risk factors.

Our results also reveal that some stocks with more volatile and more negatively skewed returns, like Yahoo, may exhibit increased exposure to jumps, and especially to positive ones. During the 3-year period of our study the index was loosing approximately 4.67 basis points on average per day. The stock price of Yahoo was declining rapidly in value, shrinking by a daily average of 16.05 basis points, with a daily volatility of 3.4%, which was twice the volatility of the market returns. Theoretically, in a standard mean-variance world, it would be suboptimal for a risk-averse investor to hold a stock with low risk-adjusted returns. Yet, we find that even in a strongly bearish market a stock like Yahoo may possess a higher exposure to an unexpected upward market jump. This indicates that theories about skewness preference (Kraus & Litzenberger, 1976) may provide an alternative insight to the reasons why investors are willing to hold stocks of poor mean-variance profile in their portfolios. Our newly proposed framework is likely to provide an alternative explanation about investors' behavior. Implied non-linearities in stochastic discount factor may allow for a potentially risk-seeking domain (Post & Levy,

2005) in the utility optimization scheme faced by investors.

5. CONCLUSIONS

We expand on the Polimenis (2012) technique by applying it jointly for two stocks, and allowing for idiosyncratic stock skew. This paper is part of a broad research program that aims to study the transmission mechanism of market jumps into individual returns and see whether these jump risk-factors are properly priced. Broadly speaking, there are three underlying hypotheses that need to be tested: a) whether systematic information is indeed transmitted with jumps and/or via infinitesimally small movements, b) whether jump sensitivities are asymmetric, and c) that jump risk is properly priced.

ΠΕΡΙΛΗΨΗ

Στην εργασία αυτή επεκτείνουμε μία μεθοδολογία μέτρησης κινδύνου η οποία βασίζεται στη λοξότητα (Polimenis, 2012), παρουσιάζοντας τον τρόπο με τον οποίο αυτή θα έπρεπε να εφαρμοστεί σε ένα πλαίσιο με περισσότερες από μία μετοχές. Εισάγουμε τη δυνατότητα για ιδιοσυγκρατικά άλματα στις τιμές των μετοχών και αναλύουμε την ευαισθησία των συντελεστών σε ένα σύστημα στο οποίο δύο μετοχές εκτίθενται ταυτόχρονα σε μη παρατηρήσιμα θετικά και αρνητικά άλματα της αγοράς. Εφόσον οι λοξότητες μεταξύ των μετοχών και της αγοράς μπορεί να διαφέρουν σημαντικά, η απαίτηση το σύστημα να εξηγήει πλήρως τη λοξότητα των μετοχών είναι μη ρεαλιστική. Βλέπουμε πως όταν επιτρέψουμε την ύπαρξη ιδιοσυγκρατικών αλμάτων στις μετοχές, το σύστημα οδηγείται σε πιο ευσταθείς λύσεις.

REFERENCES

- Aït-Sahalia, Y. (2004). Disentangling diffusion from jumps. *Journal of Financial Economics*, 74(3), 487–528.
- Aït-Sahalia, Y., & Jacod, J. (2009). Estimating the degree of activity of jumps in high frequency data. *The Annals of Statistics*, 2202–2244.
- Andrews, D. W. (1991). Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica*, 817–858.
- Ang, A., Chen, J., & Xing, Y. (2006). Downside risk. *Review of Financial Studies*, 19(4), 1191–1239.
- Barndorff-Nielsen, O. E., & Shephard, N. (2004a). Econometric analysis of realized covariation: High frequency based covariance, regression, and correlation in financial economics. *Econometrica*, 72(3), 885–925.
- Barndorff-Nielsen, O. E., & Shephard, N. (2004b). Power and bipower variation with stochastic volatility and jumps. *Journal of Financial Econometrics*, 2(1), 1–37.
- Bawa, V. S., & Lindenberg, E. B. (1977). Capital market equilibrium in a mean-lower partial moment framework. *Journal of Financial Economics*, 5(2), 189–200.
- Black, F. (1986). Noise. *The Journal of Finance*, 41(3), 529–543.
- Black, F., & Scholes, M. (1973). The pricing of options and corporate liabilities. *The Journal of Political Economy*, 637–654.
- Carr, P., Geman, H., Madan, D. B., & Yor, M. (2002). The fine structure of asset returns: An empirical investigation*. *The Journal of Business*, 75(2), 305–333.

- Carr, P., & Wu, L. (2004). Time-changed lévy processes and option pricing. *Journal of Financial Economics*, 71(1), 113–141.
- Estrada, J. (2007). Mean-semivariance behavior: Downside risk and capital asset pricing. *International Review of Economics & Finance*, 16(2), 169–185.
- Fishburn, P. C. (1977). Mean-risk analysis with risk associated with below-target returns. *The American Economic Review*, 67(2), 116–126.
- Friend, I., & Westerfield, R. (1980). Co-skewness and capital asset pricing. *The Journal of Finance*, 35(4), 897–913.
- Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the Econometric Society*, 1029–1054.
- Harlow, W. V., & Rao, R. K. (1989). Asset pricing in a generalized mean-lower partial moment framework: Theory and evidence. *Journal of Financial and Quantitative Analysis*, 24(3), 285–311.
- Harvey, C. R., & Siddique, A. (2000). Conditional skewness in asset pricing tests. *The Journal of Finance*, 55(3), 1263–1295.
- Hogan, W. W., & Warren, J. M. (1974). Toward the development of an equilibrium capital-market model based on semivariance. *Journal of Financial and Quantitative Analysis*, 9(1), 1–11.
- Kraus, A., & Litzenberger, R. H. (1976). Skewness preference and the valuation of risk assets*. *The Journal of Finance*, 31(4), 1085–1100.
- Markowitz, H. (1959). Portfolio selection: Efficient diversification of investments. *John Wiley and Sons*, 12, 26–31.
- Merton, R. C. (1976). Option pricing when underlying stock returns are discontinuous. *Journal of Financial Economics*, 3(1), 125–144.
- Newey, W. K., & West, K. D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 703–708.
- Pan, J. (2002). The jump-risk premia implicit in options: Evidence from an integrated time-series study. *Journal of Financial Economics*, 63(1), 3–50.
- Polimenis, V. (2012). Information arrival as price jumps. *Optimization*, 61(10), 1179–1190.
- Post, T., & Levy, H. (2005). Does risk seeking drive stock prices? a stochastic dominance analysis of aggregate investor preferences and beliefs. *Review of Financial Studies*, 18(3), 925–953.
- Post, T., & Van Vliet, P. (2006). Downside risk and asset pricing. *Journal of Banking & Finance*, 30(3), 823–849.
- Sortino, F. A., & Satchell, S. (2001). *Managing downside risk in financial markets*. Butterworth-Heinemann.
- Todorov, V., & Bollerslev, T. (2010). Jumps and betas: A new framework for disentangling and estimating systematic risks. *Journal of Econometrics*, 157(2), 220–235.



A NEW TWO-STAGE, SAMPLE SIZE RE-ESTIMATION DESIGN FOR CLASSICAL PHASE II CLINICAL TRIALS

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

Recently, there has been a rapid increase in the research and exploitation of new modern molecular agents, which makes necessary the development of more flexible and adaptive designs for the implementation of clinical trials. The objective of adaptive designs is to ensure direct and dynamic control of the effectiveness and the safety of a new treatment by allowing the adjustment of the elements of the study (i.e sample size), during the study, in such a way that we will not sacrifice elements which are associated with the credibility of the study (i.e. statistical power) and also issues which concern ethical characteristics of the clinical trials.

In practice, when designing a clinical trial, investigators often estimate the sample size from limited information about the variance of response rate and the size of treatment effect. Thus, it is not uncommon that the initial estimation of effect size and its variability turns to be erroneous, which could lead to an underpowered or overpowered design. Therefore, it is crucial to have the flexibility to adjust or re-estimate the sample size during the trial.

In this work we developed a new two-stage sample size re-estimation process, for classical single-arm phase II clinical trials, which allows to adjust the required sample size of the second stage n_2 and the decision cut-points a_2 and r_2 based on the information about the number of responses S_1 observed over the n_1 patients accrued at the first stage, and at the same time to maintain the targeted type I and II errors.

Keywords: adaptive designs; sample size re-estimation designs; clinical trials; phase II; multi-stage;

1. INTRODUCTION

In phase II clinical trials the investigator's purpose is to screen out therapeutic regimens that are unlikely to be effective, as well as to promote more promising therapeutic regimens to further phase III testing as quickly as possible with the

minimum number of patients. In order to minimize the total number of patients required for a phase II trial, multi-stage designs are widely used. In multi-stage designs, the trial is terminated early when the treatment shows unacceptably low therapeutic activity or acceptable enough for further testing evaluation in a phase III study. However these "classical" multi-stage designs have fixed number of interim looks and number of patients per stage, based on certain assumptions about the treatment effect and its variability at the beginning of the trial. Recently more flexible adaptive designs for the implementation of clinical trials have been developed, which allow the adjustment of the elements of the study, like the sample size, during the course of the study.

In the case of single-arm phase II designs, Banerjee and Tsiatis (2006) proposed an adaptive two-stage design which allowed the sample size at the second stage to depend on the results at the first stage. They used a Bayesian decision theory approach to minimize the average sample size (ASN) under the null hypothesis. Banerjee and Tsiatis (2006) consider only designs where the maximum sample size does not exceed the maximum sample size (N) of the corresponding Simon(1989)'s two-stage design by more than 10%. Their optimal designs always have expected sample sizes under the null hypothesis 3–5% smaller from Simon(1989)'s designs. However, as they have pointed out, the main disadvantage of the method is that, when there is an increase of the response in the first stage, there is an increase of the sample size in the second stage up to a certain point and after this point the proposed size for the second stage is zero.

Jin and Wei (2012) based on frequentist framework proposed a new adaptive design. In their design, they choose as adaptive sample size for the second stage the minimum sample size that is smaller than the corresponding second stage sample of Simon(1989)'s design and satisfies the two constraints: The conditional type I error is smaller than that of Simon(1989)'s two-stage design, and the conditional power remains at the pre-specified level. With these constraints, they guarantee that as the responses of the first stage increase the number of additional patients could be much less. However, although their design can definitely control the overall unconditional type I error under the pre-specified level, in many cases it can't control the overall unconditional type II error and as a result present 3–7% lower overall unconditional power than the corresponding designs of Simon(1989) and Banerjee and Tsiatis (2006).

In this paper, we propose a new frequentist sample size re-estimation procedure for single-arm phase II clinical trials, which is based on the overall unconditional error rates. In case of adaptive procedures, designs with more than two stages show many difficulties in setting up the design. Thus, we propose a two-stage sample size re-estimation design, in which using the information given from the first stage (i.e. the number of responses), we modify the sample size and the decision point in the second stage, while the overall unconditional type I and type II error rates are controlled.

The remaining of the paper is organized as follows: in Section 2 the proposed sample size re-estimation design is described, then in Section 3 the optimality criteria described and optimal designs are derived by minimizing ASN, subject to constraints on the unconditional errors. Finally, performance of the proposed design and concluding remarks are presented in Sections 4 and 5 respectively.

2. SAMPLE SIZE RE-ESTIMATION DESIGN

The proposed sample size re-estimation design is a modification of "classical" optimal phase II design in which, in contrast to Simon (1989)'s designs, stopping for futility or efficacy is allowed at each stage similar to Mander and Thompson (2010)'s designs. In this procedure the therapeutic efficacy is evaluate principally on the basis of response rate p and we test the classical hypothesis of single-arm phase II clinical trials ($H_0 : p \leq p_0$ versus $H_A : p \geq p_A$) by using a two-stage design. In two-stage designs, patients are entered in two stages with sample sizes n_1 and n_2 , where n_i is the size of the group added in the i^{th} stage. In the proposed sample size re-estimation two-stage procedure, the sample size of the second stage n_2 and the decision cut-points a_2, r_2 depend on the results of the first stage and more specific on the number of responses S_1 observed over the n_1 patients accrued at the first stage. Thus, hereafter we will denote the sample size and the decision cut-points for the second stage as $n_2(S_1)$ and $a_2(S_1), r_2(S_1)$, respectively.

Generally, the multi-stage phase II clinical trials, after enrolling n_i patients in the i^{th} stage a decision is made based on the observation of cumulative number of subjects with response to treatment S_i , and to the decision cut-points at the the specific stage, a_i and r_i . In the proposed two-stage sample size re-estimation design the decision rule can be described as follows:

For the first stage:

- If $S_1 \leq a_1$, do not reject $H_0 : p \leq p_0$, i.e., consider the therapy ineffective and stop sampling.
- If $S_1 \geq r_1$, reject $H_0 : p \leq p_0$, i.e., consider the therapy effective enough to warrant further testing in a phase III study and stop sampling.
- If $a_1 < S_1 < r_1$, based on the first stage number of responses S_1 continue to the second stage by taken $n_2(S_1)$ additional patients.

For the second and last stage, stop sampling and:

- If $S_2 \leq a_2(S_1)$, do not reject $H_0 : p \leq p_0$, i.e., consider the therapy ineffective.
- Otherwise, if $S_2 \geq r_2(S_1)$, reject $H_0 : p \leq p_0$, i.e., consider the therapy effective enough to warrant further testing in a phase III study,

where $a_2(S_1) = r_2(S_1) - 1$, as in the "classical" multi-stage phase II design (Fleming, 1982).

Similar to the "classical" phase II design, in the proposed design we have two possible stopping probabilities: The probability of futility, which is the probability of stopping to claim inefficacy by not rejecting the null hypothesis H_0 , and

the probability of stopping to claim efficacy by rejecting the null hypothesis H_0 . These two refer to the efficacy stopping probability (ESP) and futility stopping probability (FSP), respectively. In this adaptive design, based on the observed number of responses S_1 which occur at the first stage, we have different design parameter sets $(n_2(S_1), a_2(S_1), r_2(S_1))$ for each $S_1 \in (a_1, r_1)$, that means that we have more than one contingency to reject (or accept) the therapy at the second stage. Thus, we must calculate ESP and FSP based on the overall unconditional probabilities of stopping due to futility or efficacy of the new therapy, respectively. Based on Schultz et al (1973)'s calculation of the probability to have m responses observed up to the i^{th} stage the futility stopping probability for the two-stage procedure, which is the overall unconditional probability of rejecting the new therapy as ineffective, is given by: $FSP(p) = \sum_{i=1}^2 P(\text{not reject } H_0 \text{ at } i^{th} \text{ stage} | p) = B(a_1; p, n_1) + \sum_{S_1=a_1+1}^{r_1-1} b(S_1; p, n_1) \cdot B(a_2(S_1) - S_1; p, n_2(S_1))$. While the efficacy stopping probability that means the overall unconditional probability of accepting the new therapy as effective enough to warrant further testing in a phase III study, is given by: $ESP(p) = \sum_{i=1}^2 P(\text{reject } H_0 \text{ at } i^{th} \text{ stage} | p) = 1 - B(r_1 - 1; p, n_1) + \sum_{S_1=a_1+1}^{r_1-1} b(S_1; p, n_1) \cdot (1 - B(a_2(S_1) - S_1; p, n_2(S_1)))$. In all above $b(s; p, n)$ denotes the probability mass function of the binomial distribution evaluated at s with success probability p and number of trials n . Similarly $B(s; p, n)$ denotes the cumulative function of the binomial distribution.

Based on the above stopping probabilities the corresponding error probabilities for the new sample size re-estimation procedure are the probability of an erroneous rejection of the therapy as ineffective, when in fact the response probability is $p = p_A$ and the maximum probability of erroneous acceptance of the therapy as effective, when in fact the response probability is $p \leq p_0$. These two error probabilities are in fact the overall unconditional type I and II errors and denoted as FEP and EEP , respectively. Thus, the average sample number (ASN) of the total proposed re-estimation procedure is given by: $ASN(p) = n_1 + \sum_{S_1=a_1+1}^{r_1-1} n_2(S_1) \cdot b(S_1; p; n_1)$, where the sample size of the second stage $n_2(S_1)$ depends on the number of responses S_1 observed over the n_1 patients have been accrued at the first stage.

In the proposed sample size re-estimation design, as in the Fleming (1982)'s procedure and many others that proposed multi-stage designs, we can terminate the study at any stage both for futility and efficacy results. Thus we based on Mander and Thompson (2010)'s optimal two-stage design instead of Simon's optimal two-stage design. Since, in the proposed design the sample size and decision cut-points of the second stage are based on the interim data, the designs' parameters for the first stage must be the same as those of the corresponding non adaptive designs. That means, for specific p_0 and p_A and under the same error probability constraints the proposed adaptive design for the first stage has sample size n_1 and decision cut-off points a_1 and r_1 the same with the equivalent "classical" optimal phase II design (Mander and Thompson, 2010).

3. OPTIMAL SAMPLE SIZE RE-ESTIMATION DESIGN

Naturally, there are many combinations of $(n_2(S_1), a_2(S_1))$ for every $S_1 \in (a_1, r_1)$ which satisfy the constraints for error levels. Thus, for specific values of p_A , p_0 , and error constraints for α and β in order to find the design parameters of the proposed adaptive procedure we choose these parameters which minimized the average sample size under the null hypothesis ($H_0 : p \leq p_0$). These optimal parameters can be found based to the following strategy.

First, for the design parameters of the first stage (n_1 , a_1 and r_1) we use the same values as in the equivalent Mander and Thompson (2010)'s optimal two-stage procedure under the same error probability constraints. Thus, for these specific values n_1 , a_1 and r_1 we follow the next procedure: For maximum value of N_{max} equal to the corresponding one in case of non adaptive optimal Mander and Thompson (2010)'s design, we determine for every $S_1 \in (a_1, r_1)$ all combinations of $(n_2(S_1), a_2(S_1))$ such that $n_2(S_1) \in (n_1, N_{max})$ and $a_2(S_1) \in (a_1 + 1, n_2(S_1))$ and satisfy the constraints of the overall unconditional type I and II errors. Then from all these feasible solutions of $(n_2(S_1), a_2(S_1))$ we select the optimal solution in respect to the minimum average sample size under the null hypothesis ($ASN(p_0)$).

Based on this algorithm, we developed a SAS macro to perform the new optimal two-stage sample size re-estimation designs for single-arm phase II clinical trials. This section presents numerical examples for various values of uninteresting futility level p_0 and optimal target efficacy level p_A . We consider combinations of p_0 and p_A for targeted improvement $p_A - p_0$ being 20% (Table 1), with error probabilities constraints: $\alpha \leq 0.05$ and $\beta \leq 0.20$. For completeness we also present the corresponding design parameters in case of classical optimal Mander and Thompson (2010)'s design (using the subscript "C").

Table 1: Optimal sample size re-estimation design with corresponding optimal classical phase II design (Mander and Thompson, 2010) (denoted as MT) for $p_A - p_0 = 20\%$

$p_0 = 0.05$ and $p_A = 0.25$						
$n_1 = 9, a_1 = 0, r_1 = 3$						
Our Design				MT design		
S_1	$n_2(S_1)$	$N(S_1)$	$a_2(S_1)$	n_{2c}	N_c	a_{2c}
0	0	9		0	9	
1	8	17	2	8	17	2
2	7	16	2	8	17	2
≥ 3	0	9		0	9	
$ASN_{Re-est} = 11.83$				$ASN_c = 11.89$		

Table 1 – <i>Continued</i>						
$p_0 = 0.10$ and $p_A = 0.30$						
$n_1 = 10, a_1 = 1, r_1 = 5$						
Our Design				MT design		
S_1	$n_2(S_1)$	$N(S_1)$	$a_2(S_1)$	n_{2c}	N_c	a_{2c}
≤ 1	0	10		0	10	
2	19	29	5	19	29	5
3	18	28	5	19	29	5
4	10	22	4	19	29	5
≥ 5	0	10		0	10	
$ASN_{Re-est} = 14.85$				$ASN_c = 14.98$		
$p_0 = 0.20$ and $p_A = 0.40$						
$n_1 = 13, a_1 = 3, r_1 = 8$						
Our Design				MT design		
S_1	$n_2(S_1)$	$N(S_1)$	$a_2(S_1)$	n_{2c}	N_c	a_{2c}
≤ 3	0	13		0	13	
4	30	43	12	30	43	12
5	30	43	12	30	43	12
6	30	43	12	30	43	12
7	42	42	12	30	43	12
≥ 8	0	13		0	13	
$ASN_{Re-est} = 20.54$				$ASN_c = 20.54$		
$p_0 = 0.30$ and $p_A = 0.50$						
$n_1 = 15, a_1 = 5, r_1 = 12$						
Our Design				MT design		
S_1	$n_2(S_1)$	$N(S_1)$	$a_2(S_1)$	n_{2c}	N_c	a_{2c}
≤ 5	0	15		0	15	
6	31	46	18	31	46	18
7	31	46	18	31	46	18
8	30	45	18	31	46	18
9	28	43	17	31	46	18
10	1	16	6	31	46	18
11	1	16	6	31	46	18
≥ 12	0	15		0	15	
$ASN_{Re-est} = 23.45$				$ASN_c = 23.62$		

As it can be observed at Table 1 the proposed optimal sample size re-estimation designs with maximum sample size N_{max} the same with the corresponding classical designs (Mander and Thompson, 2010) has in most cases smaller ASN . In few cases that the two procedures present the same ASN , the proposed adaptive procedure gives the same design parameters with the corresponding classical non

adaptive procedure. This reduction of ASN with the proposed design can become greater if we allow the maximum sample size N_{max} to become larger than that we have in case of classical non adaptive design. In fact, this property was used by Banerjee and Tsiatis (2006) to produce the Bayesian sample size re-estimation designs. They considered designs where the maximum sample size can exceed the corresponding maximum sample size for Simon's designs up to 10%.

4. PERFORMANCE OF THE PROPOSED DESIGN

In order to demonstrate the performance of our frequentist sample size re-estimation phase II procedure, we compare our sample size re-estimation designs not only with the relevant non adaptive classical phase II designs, but also with the correspond frequentist sample size re-estimation designs of Jin and Wei (2012). For the specific hypothesis values $p_0 = 0.10$, $p_A = 0.30$, Figure 1, we present the probability of rejecting H_0 , probability of accepting H_0 and the average sample number ASN for each of the three procedures. In order to have fair comparisons, we evaluate these designs under the same error probability constraints: $\alpha \leq 0.05$ and $\beta \leq 0.20$.

All three procedures present equal probabilities of correct rejection of the treatment as ineffective when the response rate values are smaller than the uninterest level p_0 (Figure 1). Also, due to the same restrictions in error levels for all designs, they present equal probabilities of erroneous acceptance of the therapy as effective enough to warrant further study, when it has response probability smaller than the standard treatment. For hypothetical true response rate values greater than the uninterest level p_0 , the proposed sample size re-estimation design present smaller probability to reject the investigation treatment from Jin and Wei (2012)'s adaptive design and "classical" not adaptive design. In contrast, the proposed procedure presents greater probability to accept the treatment as effective enough to warrant further study (Phase III trial), when the response rate is greater than the uninterest level p_0 , compared with the two other procedures. Final, regarding the average sample size (ASN) the proposed adaptive design presents smaller ASN from "classical" non adaptive design, but Jin and Wei (2012)'s adaptive design presents even smaller ASN by presenting designs with overall unconditional type II error greater from the predefine constrains.

5. CONCLUDING REMARKS

In this paper a new two-stage sample size re-estimation designs proposed for single-arm phase II trials, in which based on constraints for the overall unconditional type I and II errors, and the information given from the first stage we modify the sample size and the decision point of the second stage. This fact distinguishes from Jin and Wei (2012)'s adaptive single-arm phase II design, in which they used the condition type I and II errors to find the adaptive parameters for

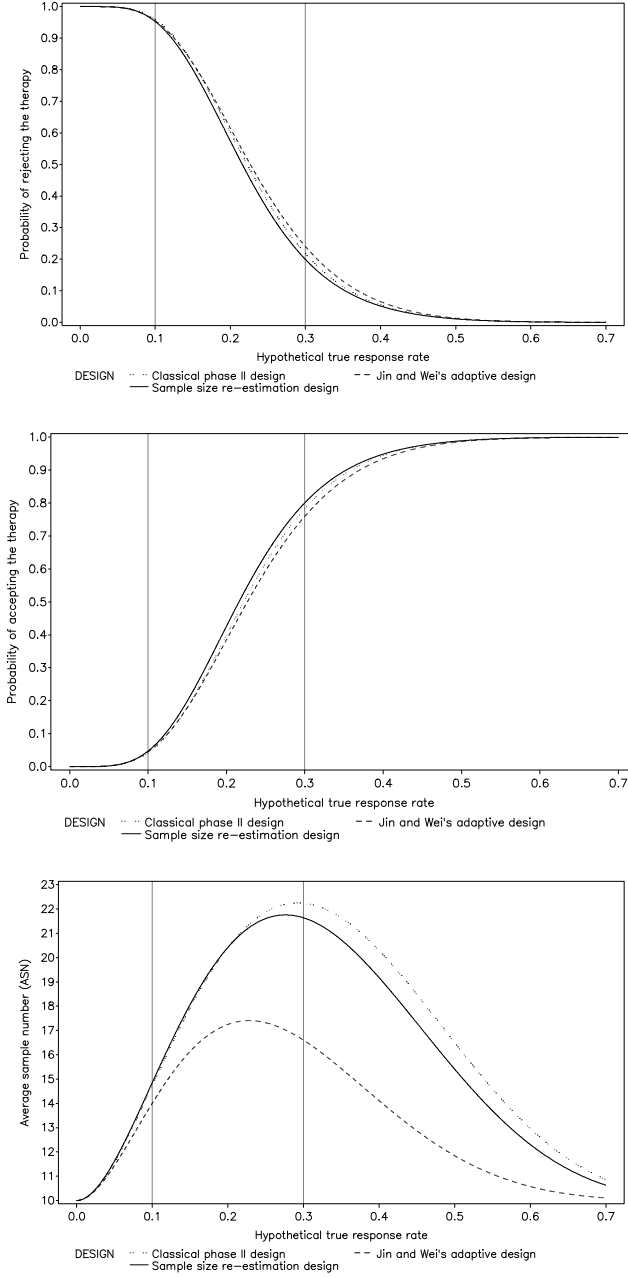


Figure 1: Comparisons of proposed sample size re-estimation designs with classical two-stage phase II designs (Mander and Thompson, 2010) and Jin and Wei (2012)’s adaptive design, for $p_0 = 0.10$ and $p_A = 0.30$.

the second stage and as a result in many cases they failed to maintain the overall unconditional type II error under the corresponding constraint. Also in the proposed adaptive procedure we have the opportunity to stop at each stage also

for futility and efficacy reasons, in contrast with the existing adaptive single-arm phase II designs in which at the first stage we can stop the trial only if the new therapy proves to be ineffective.

The proposed sample size re-estimation procedure gives us the opportunity to reject erroneously an investigation treatment with less probability compared to the the classical non adaptive procedure and Jin and Wei (2012)'s adaptive procedure. While, with this we have greater certainty when we accept correctly a treatment for further trial. Now, regarding the average sample size (*ASN*) the sample size re-estimation designs present smaller values compared to the classical non adaptive design, for hypothetical response rates greater from the standard treatment. While, the Jin and Wei's design present even smaller average sample size (*ASN*), fact that occurs due to the inability of Jin and Wei (2012)'s procedure to maintain the constrains for the overall unconditional type II error.

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

Τα τελευταία χρόνια έχει υπάρξει μια ταχύτατη αύξηση στην έρευνα και την εκμετάλλευση νέων σύγχρονων μοριακών θεραπευτικών αγωγών, γεγονός το οποίο καθιστά αναγκαία την ανάπτυξη περισσότερο ευέλικτων και προσαρμόσιμων σχεδιασμών στην υλοποίηση κλινικών μελετών. Ο σκοπός των προσαρμόσιμων σχεδιασμών είναι να διασφαλίσουν τον άμεσο και δυναμικό έλεγχο της αποτελεσματικότητας και της ασφάλειας μιας νέας θεραπευτικής αγωγής, επιτρέποντας την προσαρμογή των στοιχείων της μελέτης (π.χ μέγεθος δείγματος) κατά τη διάρκεια της, με τέτοιο τρόπο ώστε να μην θυσιάζονται στοιχεία που να συσχετίζονται με την αξιοπιστία της μελέτης (π.χ στατιστική ισχύς) καθώς και θέματα που σχετίζονται με τα ηθικά χαρακτηριστικά των κλινικών μελετών.

Στην πράξη, όταν σχεδιάζουν μια κλινική μελέτη, οι ερευνητές συχνά εκτιμούν το μέγεθος του δείγματος βασιζόμενοι σε περιορισμένες πληροφορίες για την διακύμανση του ποσοστού ανταπόκρισης και του μεγέθους της θεραπευτικής επίδρασης. Για αυτό το λόγο, κάποιες φορές η αρχική εκτίμηση της θεραπευτικής επίδρασης και της διακύμανσής της καταλήγουν να είναι εσφαλμένες, γεγονός το οποίο θα μπορούσε να οδηγήσει σε έναν σχεδιασμό με μεγαλύτερη ή μικρότερη ισχύ. Ως εκ τούτου, είναι κρίσιμο να έχουμε την ευελιξία ώστε να προσαρμόζουμε ή να επανεκτιμούμε το μέγεθος του δείγματος κατά την διάρκεια της κλινικής μελέτης.

Στην εργασία αυτή, αναπτύξαμε μία νέα διαδικασία, δύο σταδίων, επανεκτίμησης για κλασσικές κλινικές μελέτης φάσης II (single-arm), η οποία μας επιτρέπει την προσαρμογή του απαραίτητου μεγέθους δείγματος του δεύτερου σταδίου n_2 και των αντίστοιχων σημείων απόφασης a_2 και r_2 της μελέτης, βασισμένο στο συνολικό αριθμό των αποκρίσεων S_1 που παρατηρήθηκαν μεταξύ των n_1 ασθενών που συσσωρευτήκαν στο πρώτο στάδιο της μελέτης, και ταυτοχρόνως διατηρεί τα σφάλματα τύπου I και II αμετάβλητα.

REFERENCES

- Banerjee, A., Tsiatis A.A. (2006). Adaptive two-stage designs in phase II clinical trials *Stat in Med* **25**, 3382-3395.
- Fleming, TR. (1982). One-Sample Multiple Testing Procedure for Phase II Clinical Trials. *Biometrics* **38**, 143-151.
- Jin, H., Wei, Z. (2012). A new adaptive design based on Simon's two-stage optimal design for phase II clinical trials. *Contemporary Clinical Trials* **33**(6), 1255-1260.
- Mander, AP., Thompson, SG. (2010) Two-stage designs optimal under the alternative hypothesis for phase II cancer clinical trials. *Contemporary Clinical Trials* **31**, 572-578.
- Schultz, JR., Nichol, FR., Elfring, GL. and Weed, SD. (1973). Multiple stage procedures for drug screening. *Biometrics* **29**, 293-300.
- Simon, R. (1989). Optimal two-stage designs for phase II clinical trials. *Controlled Clinical Trials* **10**, 1-10.



GRANGER CAUSALITY IN MULTIVARIATE SYSTEMS USING RESTRICTED VECTOR AUTOREGRESSIVE MODELS

E.Siggiridou, D.Kugiumtzis

Department of Mathematical, Physical and Computational Sciences, Faculty of Engineering, Aristotle University of Thessaloniki
elsa_siggiridou@hotmail.com, dkugiu@gen.auth.gr,

ABSTRACT

Granger causality stems from econometrics and regards the effect of one variable on the time evolution of another variable. Measures of Granger causality have been employed for the investigation of the inter-dependence structure of complex systems. Though most complex systems are inherently nonlinear and thus only nonlinear Granger Causality measures are in principle appropriate, often the limitations of noise and time series length suggest the use of linear Granger causality measures. In the presence of many observed variables, even linear Granger causality measures may collapse due to the instability of the estimation of vector autoregressive models (VAR) of many variables on time series of limited length (large number of coefficients of lagged variables to be estimated). Here, we investigate VAR models with constraints, also referred to as sparse VAR models or dynamic regression models (DR), derived under different selection methods for the lagged variables, i.e. the top-down strategy, the bottom-up strategy, the so-called LASSO method, and a recently developed technique called backward-in-time selection (BTS). For the Granger causality from a driving variable X_i to a response variable X_j , $X_i \rightarrow X_j$, we concentrate on the presence of X_i in the representation of the DR model for X_j estimated on the time series of K observed variables $X_1, X_2, \dots, X_K$. If X_i is present, then there is Granger Causality and it is quantified by the conditional Granger causality index (CGCI), otherwise CGCI is exactly zero. The main advantage of this approach is that the DR representation involves only a subset of the set of all Kp lagged variables, where p is the maximum lag, and thus it can be applied even in problems where $Kp > N$. In this context, the top down strategy for the selection of DR is inferior to the other methods. We assess the DR representation from the four selection methods and the complete VAR representation in terms of the sensitivity and specificity of CGCI. For this, a simulation study is conducted using linear stochastic multivariate systems and different lengths of generated time series.

Keywords: Granger causality, conditional Granger causality index (CGCI), constrained or sparse VAR models

1. INTRODUCTION

The Granger causality measure has been widely applied, first in econometrics (e.g. see Hoover, 2008) and then to other sciences such as neuroscience (e.g. see the special issue in Faes et al, 2012), to reveal inter-dependence structure in multivariate time series. The measure relies on vector autoregressive (VAR) models and their estimation may be problematic when there are many observed variables and short time series. Even for systems that are inherently nonlinear and complex, such as the financial markets and the brain function, when studied at small time intervals the nonlinear methods are not successfully applicable. The same holds for linear methods that involve a large number of free parameters, such as the coefficients of a VAR model in many variables and/or high order.

The problem of reducing the number of explanatory variables has been addressed in linear regression, and many techniques like the top-down and the bottom-up strategy were called in to form VAR models with parameter constraints (Lütkepohl, 2005, Chp 5). To this respect, the LASSO method (Tibshirani, 1996), a powerful dimension reduction method in regression combined also with a recently introduced significance test called the covariance test (Tibshirani et al, 2013), can also be used to constrain VAR models. However, these methods do not take into account the lag dependence structure that typically exists in multivariate time series. This is addressed in a very recently developed method known as backward-in-time-selection (BTS) (Vlachos and Kugiumtzis, 2013).

So far there has been little focus on the problem of estimating the linear Granger causality index in short time series of many variables, where standard VAR model estimation may be problematic. We are aware only of few works attempting to estimate Granger causality after constraining the VAR models, namely the use of LASSO in electroencephalograms (Haufe et al, 2010) and microarrays (Fujita et al, 2007; Shojaie and Michailidis, 2010), and the use of principle component analysis on functional magnetic resonance imaging (Zhou et al, 2009). In this work, we compare four abovementioned techniques for VAR models with constraints (top-down, bottom-up, LASSO, BTS) in estimating best the conditional Granger causality index.

The structure of the paper is as follows. In Section 2 the methods for constraining VAR models and the estimation of Granger causality are presented, in Section 3 the simulation study is described and the results are presented, and conclusions are drawn in Section 4.

2. METHODOLOGY

Let $\mathbf{X}_t = \{X_{1,t}, X_{2,t}, \dots, X_{K,t}\}$ be a K -dimensional stationary time series. For the definition of the conditional Granger causality index (CGCI) from a driving variable X_i to a response variable X_j we consider the unrestricted vector autoregressive (VAR) model (U-model) for X_j (Brandt and Williams, 2007)

$$X_{j,t} = \sum_{k=1}^K (a_{jk,1}X_{k,t-1} + \dots + a_{jk,p}X_{k,t-p}) + u_{j,t}, \quad (1)$$

where p is the model order. The restricted VAR model (R-model) for X_j is

$$X_{j,t} = \sum_{k=1, k \neq i}^K (a_{jk,1}X_{k,t-1} + \dots + a_{jk,p}X_{k,t-p}) + e_{j,t}, \quad (2)$$

which does not include variable X_i . The terms $u_{j,t}$ and $e_{j,t}$ are white noise with mean zero and variances σ_U^2 and σ_R^2 , respectively. Fitting the U-model and R-model we get the estimates of the residual variances $\hat{\sigma}_U^2$ and $\hat{\sigma}_R^2$ and define

$$\text{CGCI}_{X_i \rightarrow X_j} = \ln \frac{\hat{\sigma}_R^2}{\hat{\sigma}_U^2}. \quad (3)$$

To assess the statistical significance of CGCI we used a significance test on the coefficients of the lagged driving variable in the U-model (Brandt and Williams, 2007). The null hypothesis is $H_0: \alpha_{ji,l}=0$ for all $l=1, \dots, p$ and the statistic is $F=((\text{SSE}^R - \text{SSE}^U)/p)/(\text{SSE}^U/ndf)$ where SSE is the sum of squared errors, ndf is the number of degrees of freedom being $(N-p)-Kp$, $N-p$ is the number of equations and Kp is the number of coefficients of the U-model. When we have many observed variables (K large) and short time series (N small) so that Kp is large compared to N the estimation of the model coefficients may not be accurate and then the CGCI cannot be trusted.

The proposed idea is to calculate the CGCI from VAR models with constraints. First we build a U-model for X_j with constraints (dynamic regression (DR) models) using one of the methods for selecting the lagged variables of $\{X_1, X_2, \dots, X_K\}$. Then we check if the lagged variables of X_i are present in the U-model. If not then the CGCI is equal to zero. Otherwise we consider as R-model the structure of the constrained U-model but without the lagged X_i components. We compute the CGCI as before from the residuals of the U-model and the R-model.

2.1. METHODS FOR SELECTING THE LAGGED VARIABLES

The methods for selecting the lagged variables are the top-down strategy, the bottom-up strategy, the LASSO method and the BTS method. Several variations of the top-down and bottom-up strategies were attempted as well. For K variables, maximum lag $p_{\max}$ (order of VAR), and a response variable X_j , all methods attempt to constrain the dynamic regression model in (1). For all dynamic regression models the parameters are estimated by least squares. For model order estimation we used either the Bayesian Information Criterion (BIC) or the Final Prediction Error (FPE). All methods are described below as algorithms.

Top-down strategy (bicTDlag, fpeTDlag) (Lütkepohl, 2005, Chp 5)

Below we show the algorithm making use the criterion BIC, denoted bicTDlag. Similarly the method is defined for the criterion FPE and is denoted fpeTDlag.

1. We start from the full DR model (1) for X_j and calculate the criterion BIC1.
2. We omit the last term $X_{K,t-p_{\max}}$ and compute the criterion BIC2 on this model.
3. If $\text{BIC1} > \text{BIC2}$ the term $X_{K,t-p_{\max}}$ is removed, otherwise it is kept in the model.

- Steps 2-3 are repeated for the second last term $X_{K-1,t-p_{\max}}$ and so on up to the last term. At each iteration, the examined lagged variable is removed from the model if the criterion BIC for this model is smaller than that for the current model with the lagged variable.

Bottom-up strategy (bicBUlag, fpeBUlag) (Lütkepohl, 2005, Chp 5)

Accordingly, bicBUlag denotes the method using BIC and fpeBUlag the method using FPE (the algorithm shown below is for BIC).

- We start with the model for X_j considering only lags of the first variable X_1 , and select the model with optimal order p_1 (estimated by BIC for $0, \dots, p_{\max}$)

$$X_{j,t} = a_{j1,1}X_{1,t-1} + \dots + a_{j1,p_1}X_{1,t-p_1} + u_{j,t}$$

- In the next step, the model of p_1 terms of X_1 is augmented to include lags of X_2 for an optimal order p_2 :

$$X_{j,t} = a_{j1,1}X_{1,t-1} + \dots + a_{j1,p_1}X_{1,t-p_1} + a_{j2,1}X_{2,t-1} + \dots + a_{j2,p_2}X_{2,t-p_2} + u_{j,t}$$

If $p_2=0$ then X_2 does not enter the model.

- Step 2 applies sequentially to variables $X_3, \dots, X_K$.
- To account for over-estimation of lagged variables, the top-down procedure is applied to the model derived after the completion of all the steps in 3.

LASSO method (Tibshirani, 1996)

The method of least absolute shrinkage and selection operator (LASSO) is a least squares method with L_1 constraints (Tibshirani, 1996) on the regression parameters. The LASSO is used here for the estimation of the coefficients of the lagged variables in the DR model (1) for X_j .

The LASSO estimate of the coefficients of the DR model stacked in vector form, $\tilde{\alpha} = \{a_{ji,l}\}$, is obtained as

$$\tilde{\alpha} = \arg \min \left\{ \sum_{t=p_{\max}+1}^N \left(X_{j,t} - \sum_{l=1}^{p_{\max}} \sum_{k=1}^K a_{jk,l} X_{k,t-l} \right)^2 \right\} \text{ subject to } \sum_{l=1}^{p_{\max}} \sum_{k=1}^K |a_{jk,l}| < s$$

where s is a tuning parameter. This method gives non-zero values for all $Kp_{\max}$ coefficients $a_{jk,l}$. To find which of $a_{jk,l}$ are statistically significant we use a significant test for the LASSO known as covariance test (Tibshirani et al, 2013).

Backward-in-time-selection technique (BTS) (Vlachos and Kugiumtzis, 2013)

The BTS method searches from smaller to larger lags and in this way it is supposed to deal with feedback and multi-collinearity, often present in multivariate time series.

- Begin with the set of orders $(0,0,\dots,0)$ for the K -variables (zero-order model) and compute BIC (equal to the variance of input noise).
- Increase the order by one for each variable separately and have thus K candidate sets of orders: $(1,0,\dots,0)$, $(0,1,\dots,0)$, ..., $(0,0,\dots,1)$. Compute the BIC for the K dynamic regression models, one for each set of orders.
- Find the set of orders for which the BIC value is smaller than the BIC value of the previous step. If there are more than one such set or orders select the one with the

- smallest BIC value and go to the next step. If no set of order has smaller BIC than the BIC from the previous step, increase the lag by one and go to the next step.
4. If the current set of order is $(p_{max}, \dots, p_{max})$ then terminate, otherwise go to step 2.

Several variations of top-down and bottom-up strategies

For the top-down strategy, the check for removal of terms in the full DR model in (1) starts from the largest lag and checks across all variables and continues to the next largest lag and so on. An alternative order of checks is with respect to variables instead of lags, i.e. start from the last variable and check across all lags and continue to the next variable and so on. We denote this method bicTDvar (based on BIC) and fpeTDvar (based on FPE). We assume the same alternative top-down scheme at the last step of the bottom-up approach and denote it bicBUvar (based on BIC) and fpeBUvar (based on FPE).

2.2. STATISTICAL EVALUATION OF METHOD ACCURACY

For a system of K variables there are $K(K-1)$ ordered pairs of variables to search for causality. On these pairs one can compute the specificity, sensitivity and Matthews correlation coefficient, which are used here as statistical measures for rating the above methods. The sensitivity is the proportion of true causal effects (true positives, TP) correctly identified as such, given as $TP/(TP+FN)$, where FN (false negatives) denotes the number of pairs having causal effects but have gone undetected. The specificity is the proportion of pairs correctly not identified having causal effects (true negatives, TN), given as $TN/(TN+FP)$, where FP (false positives) denotes the number of pairs found falsely to have causal effects. An ideal method would give values of sensitivity and specificity at one. However, this is seldom attainable and in order to weigh sensitivity and specificity collectively we consider the Matthews correlation coefficient (MCC) (Matthews, 1975) given as

$$\frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP) \cdot (TP+FN) \cdot (TN+FP) \cdot (TN+FN)}}$$

which ranges from -1 to 1. If $MCC=1$ there is perfect identification of the pairs of true causality, if it $MCC=-1$ there is total disagreement (we identify pairs of no causality as pairs of causality and vice versa), whereas MCC at the zero level indicates random identification.

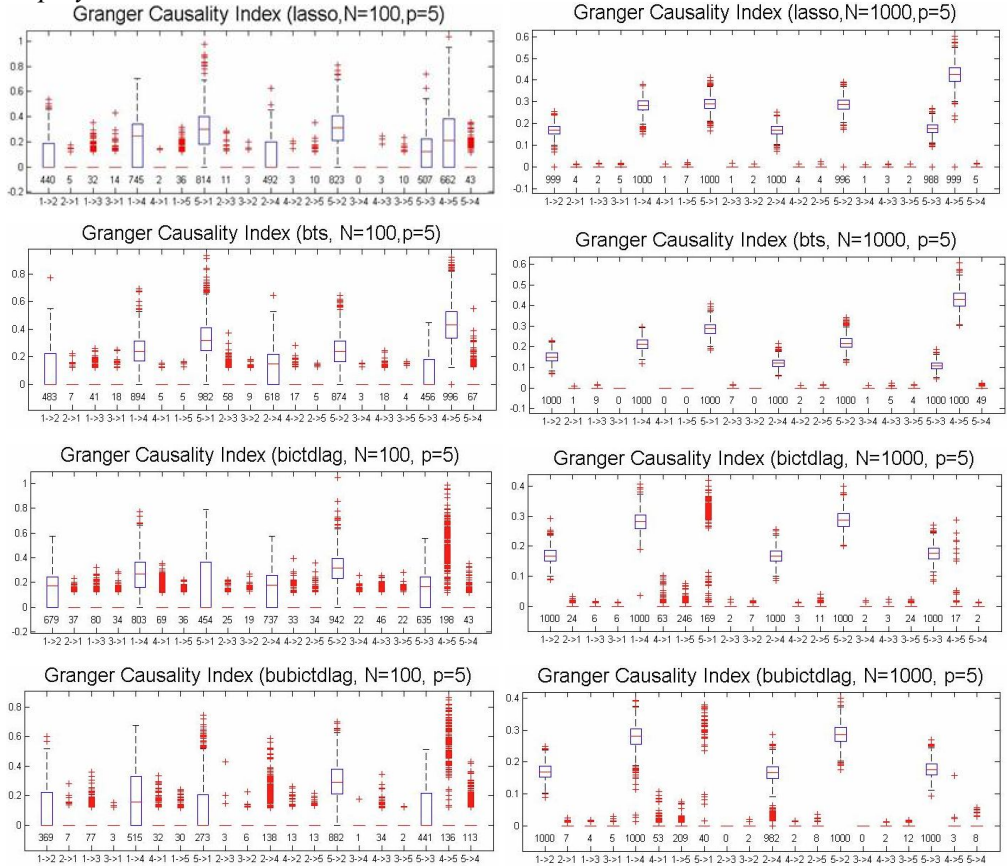
3. SIMULATIONS

In the simulations, we considered all 10 methods (BTS, LASSO, and 4 variations for each of the top-down and bottom-up strategies for ordering with respect to lags and variables and for BIC and FPE) and evaluated them in three multi-dimensional systems. The first system is an independent time series and is used to control the specificity of the methods at the total absence of Granger causality. The second system, VAR(4) and $K=5$ (model 1 in Schelter et al (2006)) and the third system, VAR(5) and $K=4$ (model 1 in Winterhalder et al (2005)), have different causality structure and are used to test both sensitivity and specificity of the methods. We made

1000 realizations for each system for different orders $p_{\max}$ and two time series lengths, $N=100, 1000$.

Specific results for all causality pairs are given as boxplots in Figure 1 for the second system. For small N , LASSO and BTS find the true causality effects at large percentages of the 1000 realizations, but not for all couplings (e.g. see causal effect from the first to the second variable denoted $1 \rightarrow 2$), whereas false causality effects are detected at a percentage smaller than the significance level 5% of the significance test (see the number of rejections displayed below the boxplots). Note that BTS tends to identify better than LASSO the true causality effects, e.g. see the true causal effect $4 \rightarrow 5$. When N increases to 1000 both methods converge and identify correctly the true and false causality effects. The top-down and bottom-up strategies (based on lag ordering) seem to perform somehow worse and cannot identify all the causality effects even for $N=1000$, e.g. see true couplings $5 \rightarrow 1$ and $4 \rightarrow 5$.

Figure 1. Boxplots of CGCI for the second system with true causality effects $\square_1 \rightarrow \square_2$, $\square_1 \rightarrow \square_4$, $\square_2 \rightarrow \square_4$, $\square_4 \rightarrow \square_5$, $\square_5 \rightarrow \square_1$, $\square_5 \rightarrow \square_2$, $\square_5 \rightarrow \square_3$, and $p_{\max}=5$ for LASSO, BTS, top-down and bottom-up across the four rows, and $N=100$ and 1000 across the two columns. Below each boxplot the number of statistically significant CGCI are displayed.



For each realization we calculate the measures sensitivity, specificity and MCC and report their mean and standard deviation (SD) from 1000 realizations. We also rank the 10 methods with respect to each of the three measures. The results on specificity for the independent time series in Table 1 and the other two systems in Table 2 show that all methods score high in specificity and their differences are overall slight as all values are very close to one. The largest differences observed are all for $N=100$ and read that the top-down strategies score lowest and BTS and LASSO highest in all systems. Interestingly, the scores of specificity are not affected by changes in the VAR order $p_{\max}$.

Table 1. *Specificity of all methods for the first system (independent time series).*

K	N	$P_{\max}$	bicTDlag	bicTDvar	BTS	bicBULag	bicBUvar	fpeBULag	fpeBUvar	fpeTDlag	fpeTDvar	LASSO
5	100	3	0.98 (0.02)	0.979 (0.02)	0.99 (0.02)	0.99 (0.01)	0.989 (0.02)	0.968 (0.03)	0.967 (0.03)	0.960 (0.03)	0.967 (0.03)	0.990 (0.01)
Rank			3	4	1	1	2	5	6	7	6	1
5	100	5	0.982 (0.02)	0.984 (0.02)	0.996 (0.01)	0.995 (0.01)	0.995 (0.01)	0.97 (0.03)	0.967 (0.03)	0.953 (0.04)	0.926 (0.03)	0.995 (0.01)
Rank			4	3	1	2	2	5	6	7	8	2
5	1000	4	0.992 (0.01)	0.993 (0.01)	0.998 (0.08)	0.997 (0.00)	0.997 (0.00)	0.978 (0.02)	0.971 (0.03)	0.973 (0.03)	0.98 (0.02)	0.995 (0.01)
Rank			5	4	1	2	2	7	9	8	6	3

Table 2. *Specificity of all methods for the second and third system.*

K	N	$P_{\max}$	bicTDlag	bicTDvar	BTS	bicBULag	bicBUvar	fpeBULag	fpeBUvar	fpeTDlag	fpeTDvar	LASSO
5	100	4	0.970 (0.04)	0.962 (0.04)	0.98 (0.03)	0.979 (0.03)	0.974 (0.03)	0.964 (0.04)	0.965 (0.04)	0.949 (0.05)	0.951 (0.05)	0.988 (0.02)
Rank			5	8	2	3	4	7	6	10	9	1
5	100	5	0.972 (0.04)	0.965 (0.04)	0.985 (0.02)	0.981 (0.03)	0.980 (0.03)	0.964 (0.04)	0.972 (0.03)	0.943 (0.05)	0.948 (0.05)	0.99 (0.02)
Rank			5	6	2	3	4	7	5	9	8	1
5	1000	5	0.977 (0.03)	0.952 (0.03)	0.995 (0.01)	0.982 (0.03)	0.929 (0.03)	0.932 (0.05)	0.957 (0.04)	0.921 (0.05)	0.927 (0.05)	0.997 (0.01)
Rank			4	6	2	3	8	7	5	10	9	1
4	100	6	0.993 (0.01)	0.993 (0.02)	0.999 (0.00)	0.998 (0.00)	0.998 (0.00)	0.991 (0.02)	0.983 (0.02)	0.981 (0.03)	0.98 (0.03)	0.998 (0.00)
Rank			3	3	1	2	2	3	4	5	6	2
4	1000	6	0.999 (0.00)	0.999 (0.00)	0.999 (0.00)	0.999 (0.00)	0.999 (0.00)	0.993 (0.01)	0.986 (0.02)	0.989 (0.02)	0.99 (0.02)	0.999 (0.00)
Rank			1	1	1	1	1	2	5	4	3	1

Regarding sensitivity, there are larger differences among the methods and more evidently for $N=100$, as shown in Table 3. The BTS, bicTDvar and fpeTDvar score highest and LASSO is ranked much lower and catch up when $N=1000$, where all methods improve and most of them reach the highest sensitivity level of one. The

results on the Matthews correlation coefficient (MCC) are similar to the results on sensitivity, mainly because MCC balances sensitivity and specificity and the main differences of the methods lie in sensitivity (see Table 4).

Dealing with systems of many variables, high order and long time series one has often to consider the computation cost of the dimension reduction method and it is known that a main shortcoming of LASSO is the computation time. We confirmed this in our simulation study and found that LASSO is about 50 times slower than BTS, which in turn is more than 5 times slower than the top-down and bottom-up strategies, the latter giving about the same computation times.

Table 3. Sensitivity of all methods for the second and third system.

K	N	P _{max}	bicTDlag	bicTDvar	BTS	bicBUlag	bicBUvar	fpeBUlag	fpeBUvar	fpeTDlag	fpeTDvar	LASSO
5	100	4	0.660 (0.15)	0.807 (0.11)	0.785 (0.14)	0.399 (0.15)	0.614 (0.13)	0.602 (0.14)	0.774 (0.14)	0.712 (0.14)	0.831 (0.12)	0.685 (0.17)
Rank			7	2	3	10	8	9	4	5	1	6
5	100	5	0.635 (0.15)	0.831 (0.12)	0.986 (0.13)	0.393 (0.16)	0.625 (0.13)	0.581 (0.15)	0.743 (0.15)	0.699 (0.15)	0.855 (0.12)	0.640 (0.17)
Rank			7	3	1	10	8	9	4	5	2	6
5	1000	5	0.741 (0.05)	1 (0)	1 (0)	0.718 (0.03)	0.997 (0.02)	0.777 (0.07)	1 (0)	0.818 (0.09)	1 (0)	0.997 (0.01)
Rank			5	1	1	6	2	4	1	3	1	2
4	100	6	0.995 (0.03)	0.998 (0.02)	0.994 (0.03)	0.940 (0.10)	0.945 (0.10)	0.993 (0.04)	0.998 (0.02)	0.997 (0.02)	0.999 (0.01)	0.893 (0.13)
Rank			4	2	5	8	7	6	2	3	1	9
4	1000	6	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	1 (0)	0.999 (0.01)
Rank			1	1	1	1	1	1	1	1	1	2

4. CONCLUSIONS

We evaluated four methods (bottom-up, top-down, LASSO and BTS) for the estimation of sparse vector autoregressive (VAR) models in their ability to identify the correct causal relationships. For causality estimation we used the conditional Granger causality index. The simulation study on a white noise process and two VAR processes showed that LASSO and BTS had the overall best performance. Differences among the methods could be detected at small time series length N , whereas for larger N all methods converge in estimating the correct causal effects. Specifically, for small N , BTS and LASSO obtained highest specificity, whereas the highest sensitivity and Matthews correlation coefficient were observed for the methods BTS and top-down (best when removing terms with respect to variable).

Our simulation study confirmed the computational shortcoming of LASSO of being much slower than all other methods (though it was implemented in R and all other methods in Matlab). The top-down strategy requires the estimation of the full model in the initial step and thus bears the same shortcoming as the full VAR for many variables and short time series. On the other hand, the bottom-up strategy and

BTS starts from few (actually none) degrees of freedom adding one at a time and thus they are both computationally and operationally effective for short time series even when the number of time series is very large. LASSO can also be used in this setting but not the covariance test. However, BTS is the only method that is designed for time-series making use of the time order of the lagged variables, and this may explain its strength over the other methods.

Table 4. Matthews correlation coefficient of all methods for second and third system.

K	N	P_{max}	bicTDlag	bicTDvar	BTS	bicBULag	bicBUvar	fpeBULag	fpeBUvar	fpeTDlag	fpeTDvar	LASSO
5	100	4	0.702 (0.14)	0.8 (0.12)	0.812 (0.13)	0.506 (0.17)	0.674 (0.13)	0.647 (0.14)	0.779 (0.13)	0.706 (0.14)	0.800 (0.12)	0.754 (0.14)
Rank			6	2	1	9	7	8	3	5	2	4
5	100	5	0.687 (0.14)	0.82 (0.12)	0.803 (0.11)	0.506 (0.16)	0.695 (0.13)	0.628 (0.15)	0.767 (0.13)	0.685 (0.16)	0.812 (0.13)	0.725 (0.13)
Rank			7	1	3	10	6	9	4	8	2	5
5	1000	5	0.778 (0.07)	0.924 (0.05)	0.992 (0.02)	0.771 (0.06)	0.887 (0.05)	0.727 (0.10)	0.932 (0.06)	0.743 (0.11)	0.889 (0.07)	0.994 (0.02)
Rank			7	4	2	8	6	10	3	9	5	1
4	100	6	0.979 (0.05)	0.98 (0.05)	0.993 (0.02)	0.959 (0.07)	0.926 (0.06)	0.974 (0.05)	0.971 (0.06)	0.951 (0.07)	0.95 (0.07)	0.93 (0.08)
Rank			3	2	1	6	10	4	5	7	8	9
4	1000	6	0.998 (0.01)	0.997 (0.01)	0.999 (0.00)	0.999 (0.00)	0.999 (0.00)	0.983 (0.04)	0.964 (0.06)	0.971 (0.05)	0.974 (0.05)	0.998 (0.01)
Rank			2	3	1	1	1	4	7	6	5	2

ΠΕΡΙΛΗΨΗ

Η αιτιότητα κατά Granger πηγάζει από την οικονομετρία και μετρά την επίδραση μιας μεταβλητής στη χρονική εξέλιξη μιας άλλης μεταβλητής. Τα μέτρα της αιτιότητας κατά Granger έχουν χρησιμοποιηθεί για τη μελέτη της δομής της αλληλεξάρτησης πολύπλοκων συστημάτων, όπου κάθε παρατηρούμενη μεταβλητή μπορεί επίσης να αναπαριστά ένα υποσύστημα. Αν και τα πιο πολύπλοκα συστήματα είναι μη-γραμμικά και επομένως μόνο μη-γραμμικά μέτρα αιτιότητας κατά Granger είναι κατ'αρχήν κατάλληλα, συχνά οι περιορισμοί του θορύβου και του μήκους των χρονοσειρών συνιστούν τη χρήση γραμμικών μέτρων. Σε περιπτώσεις πολλών παρατηρούμενων μεταβλητών, ακόμα και τα γραμμικά μέτρα αιτιότητας κατά Granger μπορεί να καταρρεύσουν λόγω της αστάθειας της εκτίμησης των αυτοπαλινδρομούμενων μοντέλων (VAR) πολλών μεταβλητών σε χρονοσειρές μικρού μήκους (μεγάλος αριθμός συντελεστών συστήματος). Εδώ διερευνούμε VAR μοντέλα με περιορισμούς, που αναφέρονται και ως δυναμικά μοντέλα παλινδρόμησης (DR), που σχηματίζονται με μεθόδους επιλογής των μεταβλητών υστέρησης, όπως η προσέγγιση από πάνω προς τα κάτω και από κάτω προς τα πάνω, η μέθοδος LASSO, καθώς και την πρόσφατα προτεινόμενη μέθοδο της χρονικά προς τα πίσω επιλογής (BTS). Για την αιτιότητα κατά Granger από μια μεταβλητή X_i σε μια μεταβλητή X_j , $X_i \rightarrow X_j$, ελέγχουμε αν η X_i υπάρχει στο DR μοντέλο για τη X_j που εκτιμήθηκε από τις χρονοσειρές K μεταβλητών $X_1, X_2, \dots, X_K$. Εάν η X_i είναι παρούσα στο μοντέλο, τότε υπάρχει αιτιότητα κατά Granger και εκτιμάται από το δείκτη της δεσμευμένης αιτιότητας κατά Granger (CGCI), διαφορετικά η CGCI είναι μηδέν. Το κύριο πλεονέκτημα αυτής της προσέγγισης είναι ότι το DR μοντέλο περιέχει μόνο ένα

υποσύνολο των KP μεταβλητών με υστέρηση, όπου P είναι η μέγιστη υστέρηση, και επομένως μπορεί να εφαρμοσθεί και σε προβλήματα όπου $Kp > N$. Αξιολογούμε τα DR μοντέλα που σχηματίζονται με τις τέσσερις μεθόδους επιλογής μεταβλητών εξετάζοντας την ευαισθησία και την ειδικότητα τους ως προς τον εντοπισμό των σωστών σχέσεων αιτιότητας σε γραμμικά πολύ-μεταβλητά στοχαστικά συστήματα για διαφορετικά μήκη παραγόμενων χρονοσειρών.

REFERENCES

- Brandt PT and Williams JT (2007). *Multiple Time Series Models*, Sage Publications
- Faes L, Andrzejak RG, Ding M and Kugiumtzis D (2012). Methodological Advances in Brain Connectivity, *Computational and Mathematical Methods in Medicine*, 2012: Article ID 492902
- Fujita A, Sato J, Garay-Malpartida H, Yamaguchi R, Miyano S, Sogayar M and Ferreira C (2007). Modeling gene expression regulatory networks with sparse autoregressive model, *BioMed Central*, 1:39
- Hoover KD (2008). Causality in economics and econometrics, *The New Palgrave Dictionary of Economics*. 2nd Edition, Eds. Durlauf SN and Blume LE, Palgrave Macmillan, 2008
- Haufe S, Mueller K-R, Nolte G and Krämer N (2010). Sparse Causal Discovery in Multivariate Time Series, *JMLR Workshop and Conference Proceedings*, 6: 97-106
- Lütkepohl H (2005). *New Introduction to Multiple Time Series Analysis*, Springer
- Matthews, B.W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme, *Biochimica et biophysica acta*, Vol. 405, No. 2. pp. 442-451
- Schelter B, Winterhalder M, Hellwig B, Guschlbauer B, Lücking CH and Timmer J (2006). Direct or indirect? Graphical models for neural oscillators, *Journal of Physiology*, 99: 37-36
- Shojaie A and Michailidis G (2010). Discovering graphical Granger causality using the truncating lasso penalty, *Bioinformatics*, Vol. 26: 517-523, doi: 10.1093/bioinformatics/btq377
- Tibshirani R (1996). Regression shrinkage and selection via the Lasso, *Journal of the Royal Statistical Society Series B*, 58(1): 267-288
- Tibshirani R, Tibshirani RJ, Taylor J and Lockhart R (2013). A significance test for the lasso, *arXiv:1301.7161*
- Vlachos I and Kugiumtzis D (2013). Backward-in-time selection of the order of dynamic regression prediction model, *Journal of Forecasting*, to appear
- 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
- Zhou Z, Chen Y, Ding M, Wright P, Lu Z and Liu Y (2009). Analyzing brain networks with PCA and conditional Granger causality, *Human Brain mapping*, 30:2197-2206



ΕΥΡΕΤΗΡΙΟ ΣΥΓΓΡΑΦΕΩΝ (INDEX)

BALAKRISHNAN N.	108	ΓΕΩΡΓΙΟΥ Σ.	65
CHARITIDOU E.	253	ΔΗΜΗΤΡΑΚΟΣ Θ.	35
DONATOS G.S.	261	ΔΟΝΑΤΟΥ Α.	45
DRAKOS K.	271	ΔΡΙΤΣΑΚΗΣ Ν.	55
FOUSKAKIS D.	253	ΕΥΘΥΜΙΟΥ Ι.	65
KARAGRIGORIOU A.	291	ΘΕΟΔΟΣΙΑΔΟΥ Ο.	76
KARLIS D.	302, 312, 346	ΚΑΛΑΜΑΡΑΣ Δ.	87
KITSOS C	281	ΚΑΛΑΜΑΤΙΑΝΟΥ Α.	87
KOUTRAS V.	271	ΚΑΡΑΚΑΣΙΔΟΥ Σ.	98
KUGIUMTZIS D.	356	ΚΙΤΣΟΥ Δ.	242
MAKRIDES A.	291	ΚΟΛΥΒΑ-ΜΑΧΑΙΡΑ Φ.	180
MICHAILIDIS G.C.	261	ΚΟΥΓΙΟΥΜΤΖΗΣ Δ.	158
NTZOUFRAS I.	253	ΚΟΥΚΟΥΒΙΝΟΣ Χ.	170
OIKONOMOPOULOU G.	302	ΚΟΥΝΙΑΣ Σ.	170, 180
ORFANOGIANNAKI K.	312	ΚΟΥΤΡΑΣ Μ.Β.	108
PAPANTONIS I.	336	ΚΟΥΤΡΟΥΒΕΛΗΣ Ι.Α.	118
PETRALIAS A.	322	ΚΥΡΙΑΚΙΔΗΣ Ε.	35
POLIMENIS V.	336	ΛΑΓΟΥ Β.	128
POULOPOULOU S.	346	ΛΕΒΕΝΤΙΔΗΣ Ι.	45
PRODROMIDIS P.	322	ΜΗΛΙΕΝΟΣ Φ.Σ.	108
SIGGIRIDOU E.	356	ΜΠΑΡΤΖΗΣ Γ.	128
ANTZOYΛΑΚΟΣ Δ.Α.	201	ΜΠΕΡΤΟΛΗΣ Ι.	128
ΒΑΣΙΛΕΙΑΔΗΣ Β.Γ.	25	ΜΠΕΡΣΙΜΗΣ Σ.	211

ΜΠΟΥΤΣΙΚΑΣ Μ.Β.	138, 201
ΜΩΥΣΙΑΔΗΣ Π.	148
ΠΑΠΑΪΩΑΝΝΟΥ Τ.	211
ΠΑΠΑΣΠΥΡΟΠΟΥΛΟΣ Κ.Γ.	158
ΠΕΡΙΚΛΕΟΥΣ Κ.	170, 180
ΠΕΤΡΟΣ Σ.	190
ΠΟΛΥΜΕΝΗΣ Β.	76
ΠΡΟΔΡΟΜΙΔΗΣ Π.	190
ΡΑΚΙΤΖΗΣ Α.Χ.	201
ΡΗΓΑΣ Α.Γ.	25
ΣΑΧΛΑΣ Α.	211
ΣΕΛΗΝΙΔΟΥ Α.	128
ΣΤΑΜΑΤΟΠΟΥΛΟΥ Ε.	128
ΤΡΙΚΑΛΙΩΤΟΥ Α.	128
ΤΣΑΚΛΙΔΗΣ Γ.	76
ΤΣΙΜΠΟΣ Κ.	222
ΦΑΡΜΑΚΗΣ Ν.	231
ΧΑΛΚΟΣ Γ.	242
ΧΑΤΖΟΠΟΥΛΟΣ Σ.	180