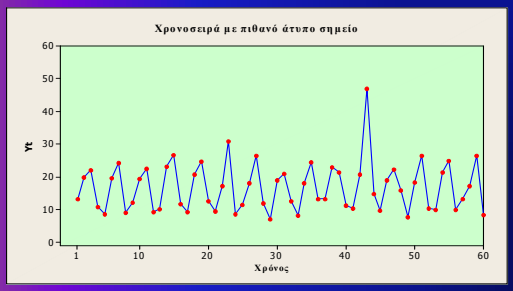
ΕΡΕΥΝΗΤΙΚΟ ΠΡΟΓΡΑΜΜΑ «ΘΑΛΗΣ»

ΑΝΑΛΥΣΗ ΑΠΟΜΟΝΩΜΕΝΩΝ ΣΗΜΕΙΩΝ ΣΕ ΣΤΑΤΙΣΤΙΚΑ ΔΕΔΟΜΕΝΑ

ΧΡΥΣΗΙΣ ΚΑΡΩΝΗ
Αναπληρώτρια Καθηγήτρια, Τομέας Μαθηματικών, ΣΕΜΦΕ
Επιστημονικός Υπεύθυνος

ΒΑΣΙΛΙΚΗ ΚΑΡΥΩΤΗ
Κύρια ερευνήτρια

ΠΟΛΥΧΡΟΝΗΣ ΟΙΚΟΝΟΜΟΥ **ΧΡΙΣΤΙΝΑ ΠΙΕΡΡΑΚΟΥ**
Νέοι ερευνητές

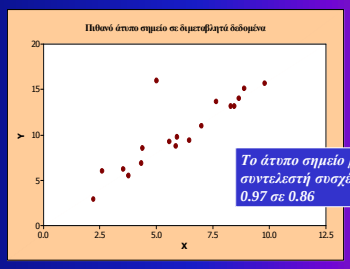


Τι είναι απομονωμένα ή άτυπα σημεία;
Γιατί είναι σημαντικά ;

Απομονωμένα σημεία (outliers) είναι στατιστικά δεδομένα που διαφέρουν πολύ από το υπόλοιπο ενός συνόλου. Η διαφορά μπορεί να εντοπιστεί στις τιμές μιας μεταβλητής (π.χ. η θράυση δοκαριού υπό εξαιρετικά μικρό φορτίο) ή στο συνδυασμό τιμών σε πολυδιάστατα δεδομένα.

Ιδιαίτερη σημασία έχει το γεγονός ότι ενδεχομένως η παρουσία απομονωμένων σημείων επηρεάζει αισθητά την εκτίμηση των παραμέτρων ενός μοντέλου, οδηγώντας σε λανθασμένα συμπεράσματα και μη-ακριβείς προβλέψεις.

Τα διαγράμματα δίνουν δύο παραδείγματα εμφανών άτυπων σημείων, σε χρονοσειρά (αριστερά) και σε ένα δισδιάστατο σύνολο δεδομένων (δεξιά).



Το άτυπο σημείο μειώνει το συντελεστή συσχέτισης από 0.97 σε 0.86

Σκοπός της έρευνας

Υπάρχει εκτεταμένη βιβλιογραφία στο θέμα των άτυπων σημείων (μόνο στους V. Barnett & T. Lewis, "Outliers in Statistical Data", 1994, υπάρχουν σχεδόν 1000 αναφορές). Παρόλα αυτά δεν έχει διερευνηθεί πολύ το πρόβλημα αυτό στην περίπτωση των *χρονοσειρών*, όπου η παρουσία ενός απομονωμένου σημείου είναι αρκετά πιθανή και οι επιπτώσεις στην εκτίμηση και ανάλυση μεγάλες (π.χ. μια *οικονομική χρονοσειρά* που επηρεάζεται από εξωτερικούς παράγοντες – απεργίες, πολέμους κ.ο.κ.).

Οι κλασικές μέθοδοι ανάλυσης χρονοσειρών εφαρμόζονται σε *μια σειρά μεγάλης διάρκειας*. Πολλές όμως είναι οι περιπτώσεις όπου προκύπτει ένα *σύνολο χρονοσειρών μικρού μεγέθους*. Η έρευνά μας ασχολείται κυρίως με την ανίχνευση άτυπης σειράς ή σημείου *σε σύνολο χρονοσειρών*.

1η περίπτωση

***Σκοπός:* η ανίχνευση απομονωμένης σειράς. Θεωρούμε το σύνολο m αυτοπαλινδρομικών σειρών AR(1)**

$$y_{it} - \mu_i = \alpha(y_{i,t-1} - \mu_i) + u_{it},\quad i=1,\ldots,m,\quad t=1,\ldots,n_i,\ u_{it} \text{ iid } N(0,\sigma^2)$$

με τις ακόλουθες δύο περιπτώσεις για τα επίπεδα μ των σειρών

I) $\mu_i = \mu, \forall i$ - όλες οι σειρές έχουν το ίδιο επίπεδο
II) $\mu_i \sim N(\mu_0, \sigma_0^2)$ - ένα μοντέλο με τυχαία κατανομή των επιπέδων των σειρών

Στην παρουσία άτυπης σειράς τα μοντέλα για τις δύο περιπτώσεις διαμορφώνονται ως εξής:

I') $\mu_i = \mu, i \neq j$
$\mu_j = \mu + \delta$, για κάποιο j

II') $\mu_i \sim N(\mu_0, \sigma_0^2),\ i \neq j$
$\mu_j \sim N(\mu_0 + \delta, \sigma_0^2)$, για κάποιο j

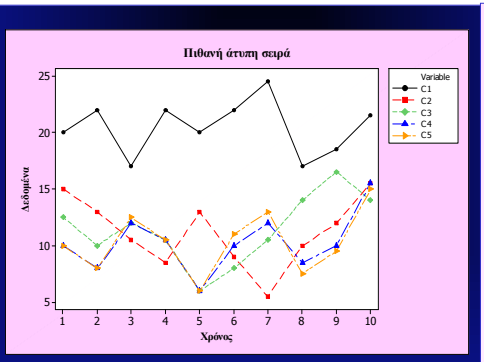
Σε αυτό το πλαίσιο,

α) κατασκευάσαμε στατιστικούς ελέγχους με την τεχνική του λόγου των πιθανοφανειών δύο σταδίων μεταξύ των υποθέσεων I-I' και II-II' και μελετήσαμε τις ιδιότητές τους,

β) εξετάσαμε την εφαρμογή απλών μεθόδων ανίχνευσης απομονωμένης τιμής σε ένα σύνολο δεδομένων, στις μέσες τιμές των σειρών (ή σε απλές συναρτήσεις αυτών).

Παρότι το άνισο μέγεθος χρονοσειρών υποδηλώνει ότι οι μέσες τιμές των σειρών έχουν άνισες διασπορές, αποδείξαμε ότι αυτή η απλή προσέγγιση λειτουργεί καλά.

(βλ. Karioti & Caroni, 2004)



ΜΟΝΤΕΛΑ ΔΙΑΡΚΕΙΑΣ ΖΩΗΣ

Τα μοντέλα διάρκειας ζωής εφαρμόζονται στη μελέτη επιβίωσης και αξιοπιστίας (χρόνος μέχρι το θάνατο ατόμου ή χρόνος μέχρι τη διακοπή λειτουργίας μηχανής ή πίεση που προκαλεί θραύση υλικού). Άτυπα σημεία είναι μονάδες με εξαιρετικά μικρή ή ασυνήθιστα μεγάλη διάρκεια ζωής, ενδεχομένως με σημαντική επίδραση στη διαμόρφωση του μοντέλου και την εκτίμηση των παραμέτρων του. Μια αρχική εξέταση αυτών των μοντέλων παρουσιάστηκε σε συνέδριο.

(βλ. Οικονόμου και Καρώνη, 2003)

2η περίπτωση

Θεωρούμε ότι εξωτερικός παράγοντας προκαλεί την εμφάνιση απομονωμένου σημείου σε κάθε σειρά του συνόλου, κατά την ίδια χρονική στιγμή. Εξετάζουμε τα *IO* (*καινοτομικά άτυπα σημεία*) καθώς και τα *AO* (*προσθετικά άτυπα σημεία*) σε αυτοπαλινδρομικές σειρές AR(p).

Τυχαίο IO:

$$u_{it} \sim N(0,\sigma^2),\ i=1,\ldots,m;\ t=1,\ldots,n_i;\ t \neq q$$
$$u_{iq} \sim N(\Delta, \sigma^2 + \sigma_\delta^2)$$

Κατασκευάσαμε ελεγχοσυνάρτηση με την τεχνική του λόγου πιθανοφανειών δύο σταδίων και προσδιορίσαμε κρίσιμα σημεία μέσω προσομοιώσεων.

IO ίδιο στο i : **Σε αυτή την περίπτωση, το μοντέλο μπορεί να γραφεί στη μορφή μιας γραμμικής παλινδρόμησης**

$$\underset{\sim}{y}_i = X_i\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}_i,\quad i=1,\ldots,m$$

με $V(\underset{\sim}{\varepsilon}) = \Sigma \otimes I$. Η παρουσία ενός άτυπου σημείου ισοδυναμεί με την πρόσθεση μιας στήλης στο X_i .

Εξετάσαμε τρία μοντέλα:

$\Sigma = \text{dia}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_m^2)$	(ετεροσκεδαστικότητα μεταξύ σειρών)
Σ χωρίς περιορισμό	
$\Sigma = \sigma^2\{(1-\rho)I + \rho J\}$	(ισοσυσχέτιση)

Με τη μέθοδο των γενικευμένων ελαχίστων τετραγώνων, έχουμε

$$\hat{\underset{\sim}{\beta}} = \left(X' V^{-1} X\right)^{-1} X' V^{-1} \underset{\sim}{y}$$

καθώς και ένα τύπο για το $\hat{V}$ διαφορετικό για τις τρεις μορφές του Σ. Οι εξισώσεις λύνονται με επαναληπτικό μέθοδο.

Κατασκευάσαμε πάλι μια ελεγχοσυνάρτηση με την τεχνική του λόγου πιθανοφανειών δύο σταδίων. Κρίσιμες τιμές προσδιορίστηκαν με την προσέγγιση X^2 και επιβεβαιώθηκαν με προσομοιώσεις.

(Βλ. Caroni & Karioti, 2004)

***AO:* η περίπτωση AO μπορεί να αναπτυχθεί στις ίδιες γενικές γραμμές με το τυχαίο IO.**

ΣΥΜΠΕΡΑΣΜΑΤΑ

Οι δύο μας δημοσιεύσεις αποτελούν τις πρώτες αναφορές στη διεθνή βιβλιογραφία στη μεθοδολογία ανίχνευσης απομονωμένων σειρών ή σημείων σε δεδομένα υπό μορφή ενός συνόλου χρονοσειρών. Ως εκ τούτου, συμβάλουν σημαντικά στη μεθοδολογία ανάλυσης τέτοιων δεδομένων, τα οποία παρουσιάζονται συχνά σε διάφορες επιστημονικές περιοχές.

ΔΗΜΟΣΙΕΥΣΕΙΣ ΣΕ ΠΕΡΙΟΔΙΚΑ

V. KARIOTI & C. CARONI (2004). "Simple detection of outlying short time series". *Statistical Papers*, **45**, 267-278.

C. CARONI & V. KARIOTI (2004). "Detecting an innovative outlier in a set of time series". *Computational Statistics and Data Analysis*, **46**, 561-570.

ΑΝΑΚΟΙΝΩΣΕΙΣ ΣΕ ΣΥΝΕΔΡΙΑ

V. KARIOTI & C. CARONI * (2002). "Detecting an outlier in a set of time series". 17th International Workshop in Statistical Modelling . Chania, Greece, 8-12 July, 2002.

V. KARIOTI & C. CARONI (2003). "Fixed and random innovative outliers in sets of time series". International Workshop on Computational Management Science, Economics, Finance and Engineering, Limassol, Cyprus, 28-30 March, 2003.

Β. ΚΑΡΥΩΤΗ & Χ. ΚΑΡΩΝΗ * (2003). «Ανίχνευση ενός απομονωμένου σημείου (ΑΟ ή ΙΟ) σε ένα σύνολο χρονοσειρών». 16ο Πανελλήνιο Συνέδριο Στατιστικής, *Ελληνικό Στατιστικό Ινστιτούτο*, Καβάλα, 30 Απριλίου-3 Μαΐου 2003.

Π. ΟΙΚΟΝΟΜΟΥ και Χ. ΚΑΡΩΝΗ * (2003). «Μελέτη και ανάπτυξη μοντέλων ανάλυσης επιβίωσης». 16ο Πανελληνίου Συνεδρίου Στατιστικής, *Ελληνικό Στατιστικό Ινστιτούτο*, Καβάλα, 30 Απριλίου-3 Μαΐου 2003.

(*) Έχουν δημοσιευθεί σε πρακτικά συνεδρίου.