

# Αλγόριθμος στατιστικής επεξεργασίας δεδομένων συνδρομικής επιτήρησης σε χώρους φροντίδας μεταναστών και προσφύγων

## Σκοπός

Ο προτεινόμενος αλγόριθμος σκοπό έχει τον υπολογισμό της **εκτιμώμενης αναλογικής νοσηρότητας** για κάθε επιτηρούμενο σύνδρομο, τόσο συνολικά όσο και ανά κέντρο φιλοξενίας. Επίσης σκοπό έχει τον υπολογισμό ενός **επιπέδου επιφυλακής** (alert level), το οποίο αντικειμενικά διαχωρίζει τα συνήθη επίπεδα νοσηρότητας από ασυνήθιστες επιδημικές εξάρσεις.

Ο αλγόριθμος βασίζεται στη μέθοδο κατά Farrington<sup>1</sup>, η οποία είναι απλή, ευέλικτη και χρησιμοποιείται ευρύτατα για σκοπούς επιδημιολογικής επιτήρησης<sup>2</sup>.

## Δεδομένα εισόδου

- $Y(t)$ : αριθμός περιστατικών του επιτηρούμενου συνδρόμου στον χρόνο  $t$  (ημέρα ή εβδομάδα δήλωσης).
- $N(t)$ : συνολικός αριθμός επισκέψεων στο ιατρείο στον χρόνο  $t$  (ημέρα ή εβδομάδα δήλωσης).

## Βήμα 1: Μοντέλο παλινδρόμησης

Στα ανωτέρω δεδομένα εισόδου προσαρμόζεται ένα μοντέλο παλινδρόμησης quasi-Poisson με υπερδιασπορά (quasi-Poisson regression with overdispersion), από το οποίο υπολογίζεται ο αναμενόμενος αριθμός περιστατικών  $\mu(t)$  για ορισμένο συνολικό αριθμό επισκέψεων.

Στην απλούστερη περίπτωση, το μοντέλο είναι intercept-only, και επομένως η αναμενόμενη αναλογική νοσηρότητα είναι σταθερή στο χρόνο, ίση με  $\exp(\alpha)$ :

$$(1) \quad \log(\mu(t)) = \alpha + \log(N(t))$$

Στο βασικό αυτό μοντέλο μπορούν να προστεθούν διάφορες συμμεταβλητές (covariates)  $f(X)$ :

$$\log(\mu(t)) = \alpha + \beta * f(X) + \log(N(t))$$

Με τον τρόπο αυτό η αναμενόμενη αναλογική νοσηρότητα μπορεί να εκφραστεί ως συνάρτηση τόσο του χρόνου  $t$ , όσο και διαφόρων άλλων παραγόντων (π.χ. εποχή του χρόνου, ημέρα της εβδομάδας, θερμοκρασία, περίοδοι νηστείας όπως το Ραμαζάνι, κλπ).

Για την παρακολούθηση των φυσιολογικών μακροχρόνιων τάσεων της νοσηρότητας προτείνεται η ενσωμάτωση στο μοντέλο του χρόνου  $t$  μέσα από μια natural cubic spline συνάρτηση  $S(t, k)$ . Αν ο χρόνος  $t$  εκφράζεται σε ημέρες, οι βαθμοί ελευθερίας  $k$  προτείνεται να είναι ίσοι με τον αριθμό των συμπληρωμένων μηνών της χρονοσειράς. Πρακτικά αυτό σημαίνει πως η χρονική τάση (time trend) της αναμενόμενης αναλογικής νοσηρότητας δύναται να αλλάζει περίπου κάθε μήνα, και αποφεύγονται έτσι οι πολύ βραχυχρόνιες διακυμάνσεις οι οποίες θα συνιστούσαν υπερπροσαρμογή (overfitting).

$$(2) \quad \log(\mu(t)) = \alpha + \beta * S(t, k) + \log(N(t))$$

Για πολύ σπάνια σύνδρομα είναι λογικότερο η αναμενόμενη νοσηρότητα να θεωρηθεί σταθερή στο χρόνο, εκτός αν η χρονοσειρά είναι μεγάλη (πολλών ετών). Ομοίως δεν έχει νόημα να ενσωματωθεί ο χρόνος στο μοντέλο αν η χρονοσειρά είναι μικρότερη από ένα μήνα.

Συνεπώς χρησιμοποιείται το μοντέλο (2), εκτός αν:

- η χρονοσειρά είναι διάρκειας μικρότερης του ενός μήνα, ή/και
- ο αριθμός περιστατικών  $Y(t)$  είναι μηδέν σε τουλάχιστον 75% των παρατηρήσεων

Στην περίπτωση αυτή προτιμάται το μοντέλο (1).

## Βήμα 2: Υπολογισμός τυπικής απόκλισης και ορίου επιφυλακής

Για τον υπολογισμό της τυπικής απόκλισης (standard deviation) του παρατηρούμενου αριθμού περιστατικών  $Y(t)$  χρησιμοποιείται ο μετασχηματισμός στην  $2/3$  δύναμη<sup>1</sup>, ο οποίος διορθώνει τη λοξότητα (skewness) που χαρακτηρίζει την κατανομή Poisson. Τελικά το επίπεδο επιφυλακής  $Y_a(t)$  για τον παρατηρούμενο αριθμό περιστατικών ορίζεται σε συνάρτηση του αριθμού των τυπικών αποκλίσεων  $Z$ , ως εξής:

$$(3) \quad Y_a(t) = \mu(t) * \left( 1 + \frac{2}{3} Z \sqrt{\frac{\varphi + \text{Var}(\mu(t))/\mu(t)}{\mu(t)}} \right)^{3/2}$$

όπου  $\varphi$  ο συντελεστής υπερδιασποράς (overdispersion factor) όπως εκτιμάται από το μοντέλο παλινδρόμησης (1) ή (2). Με παρόμοιο τρόπο μπορεί να υπολογιστεί και το z-score του παρατηρούμενου αριθμού περιστατικών, δηλαδή ο αριθμός τυπικών αποκλίσεων σε σχέση με τον αναμενόμενο αριθμό.

Για  $Z=1.64$  ο τύπος (3) ορίζει ένα μονόπλευρο 95% διάστημα πρόγνωσης (one-sided 95% prediction interval) για τον παρατηρούμενο αριθμό περιστατικών  $Y(t)$ . Για  $Z=2$  τυπικές αποκλίσεις, το διάστημα πρόγνωσης είναι 97.5%, δηλαδή μόνο το 2.5% των παρατηρήσεων αναμένεται τυχαία να βρεθούν υψηλότερα της τιμής  $Y_a(t)$ . **Ως όριο επιφυλακής προτείνονται οι 2 τυπικές αποκλίσεις ( $Z=2$ ).** Εξυπακούεται ότι διαιρώντας  $Y_a(t)/N(t)$  προκύπτει το όριο επιφυλακής για την αναλογική νοσηρότητα.

## Βήμα 3: Απόρριψη ασυνήθιστων τιμών και επανυπολογισμός

Το μοντέλο παλινδρόμησης του βήματος 1 εκτιμά τον αναμενόμενο αριθμό περιστατικών που αντιστοιχεί στη συνήθως παρατηρούμενη νοσηρότητα. Τυχόν ασυνήθιστες επιδημικές εξάρσεις θα αυξήσουν την τυπική απόκλιση του παρατηρούμενου αριθμού περιστατικών, και άρα θα ανεβάσουν το όριο επιφυλακής. Επομένως οι παρατηρούμενες τιμές που πιστοποιημένα αντιστοιχούν σε επιδημικά επίπεδα, είναι ορθό να αποκλείονται από τον υπολογισμό της αναμενόμενης νοσηρότητας και του ορίου επιφυλακής. Από την άλλη μεριά, αποκλεισμός τιμών που μπορεί τυχαία να είναι υψηλές χωρίς όμως να αντιστοιχούν σε επιδημία, οδηγεί σε υποεκτίμηση της τυπικής απόκλισης και ψευδώς χαμηλότερο όριο επιφυλακής.

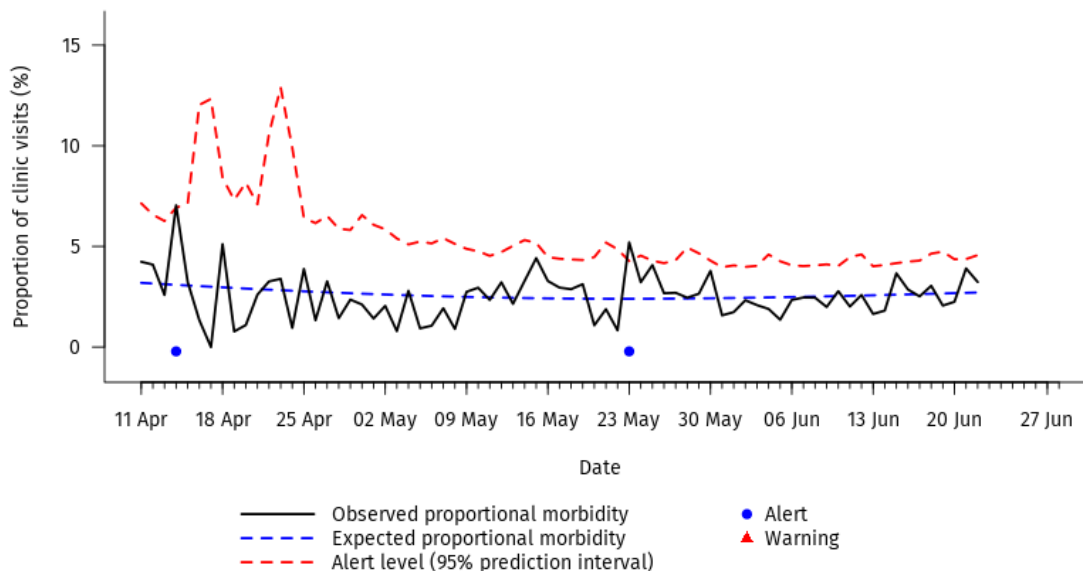
Για το σκοπό αυτό, προτείνεται η εξής διαδικασία: μετά τα βήματα 1 και 2, ελέγχεται αν υπάρχουν σημεία στη χρονοσειρά με z-score μεγαλύτερο του 3, που είναι σχεδόν απόλυτα βέβαιο ότι αντιστοιχούν σε επιδημία. Αν υπάρχουν, τότε ξαναπροσαρμόζεται το μοντέλο παλινδρόμησης (βήμα 1) αποκλείοντας τα σημεία αυτά, και επανυπολογίζεται το όριο επιφυλακής (βήμα 2). Ο κύκλος επαναλαμβάνεται αν χρειάζεται, μέχρι τα σημεία που χρησιμοποιούνται στον υπολογισμό του μοντέλου να έχουν όλα z-score μικρότερο από 3.

Συμπληρωματικά υπάρχει και η δυνατότητα χειροκίνητης εξαίρεσης τιμών από τον αλγόριθμο, εφόσον το z-score είναι μικρότερο από 3 αλλά θεωρούνται μέρος επιβεβαιωμένης επιδημικής δραστηριότητας.

## Παράδειγμα

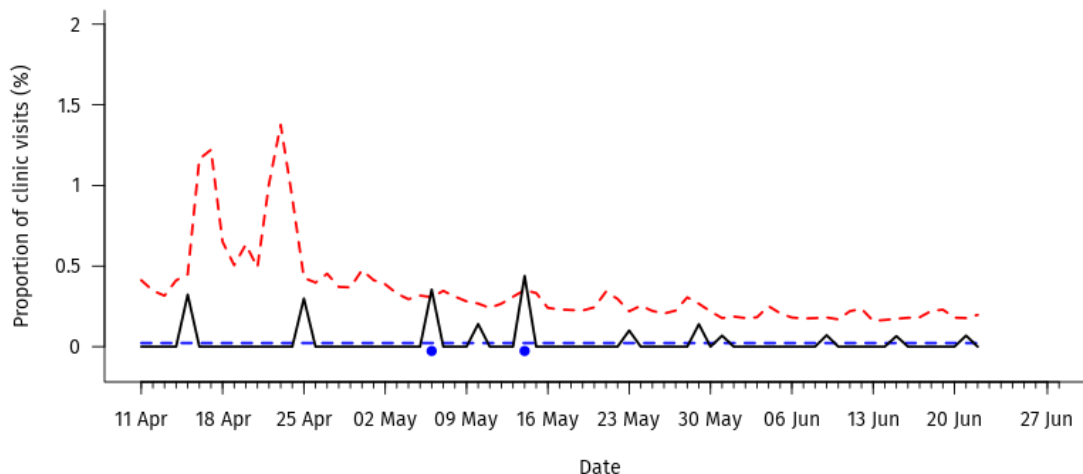
Στο πρώτο διάγραμμα, το μοντέλο περιλαμβάνει natural cubic spline με 2 βαθμούς ελευθερίας. Τα σκαμπανεβάσματα στο όριο επιφυλακής αφορούν το συνολικό αριθμό επισκέψεων (δηλαδή τον παρονομαστή)· αν είναι χαμηλότερος, το όριο επιφυλακής είναι υψηλότερο.

Gastroenteritis without blood in the stool



Στο δεύτερο διάγραμμα, το επιτηρούμενο σύνδρομο είναι σπάνιο και το μοντέλο παλινδρόμησης είναι intercept-only. Η αναμενόμενη νοσηρότητα είναι σταθερή στο χρόνο και πάρα πολύ μικρή. Οι διακυμάνσεις της παρατηρούμενης νοσηρότητας οφείλονται σε μεμονωμένα περιστατικά, εκτός από δύο φορές που παρατηρήθηκαν δύο περιστατικά την ίδια μέρα· και τις δύο φορές, ξεπεράστηκε ελαφρά το όριο επιφυλακής.

Bloody diarrhoea



## Βιβλιογραφία

1. Farrington CP, Andrews NJ, Beale AD, Catchpole MA. A Statistical Algorithm for the Early Detection of Outbreaks of Infectious Disease. *Journal of the Royal Statistical Society Series A (Statistics in Society)*. 1996;159(3):547.
2. Hulth A, Andrews N, Ethelberg S, Dreesman J, Faensen D, van Pelt W, et al. Practical usage of computer-supported outbreak detection in five European countries. *Euro Surveill*. 2010 Sep 9;15(36).