ΔΗΜΟΚΡΙΤΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΡΑΚΗΣ
DEMOCRITUS UNIVERSITY OF THRACE

Στοχαστική Ανάλυση Χρονοσειρών

Περιεχόμενο:

Βασικοί ορισμοί.
Στασιμότητα.
Διαφοροποίηση χρονοσειράς.
Στατιστικά και Διαγράμματα που Χρησιμοποιούμε στην Ανάλυση Χρονοσειρών.

Επαμεινώνδας Διαμαντόπουλος
Παρασκευή 29 Μαρτίου 2024

ΔΗΜΟΚΡΙΤΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΡΑΚΗΣ
DEMOCRITUS UNIVERSITY OF THRACE

Εισαγωγή

Βασικοί Ορισμοί

Με τον όρο **χρονοσειρά** αναφερόμαστε σε μια ακολουθία τυχαίων μεταβλητών $\{X_t : t = 1, 2,...\}$ η οποία εκφράζει την εξέλιξη της κατάστασης ενός συστήματος στο χρόνο.

Η μεταβλητή X_t δίνει την κατάσταση του συστήματος (αριθμητική τιμή) τη χρονική στιγμή t.

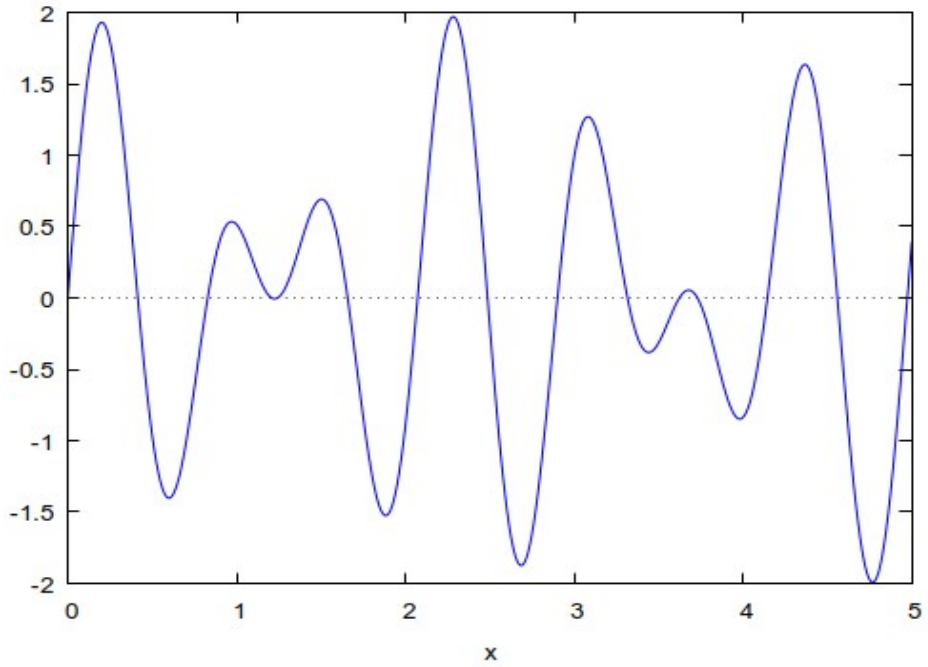
Οι χρονοσειρές διακρίνονται ως προς το είδος του χρόνου t και το είδος των δεδομένων x_t στις εξής κατηγορίες:

- Διακριτά μεγέθη x_t σε διακριτό χρόνο t.
- Συνεχή μεγέθη x_t σε διακριτό χρόνο t.
- Διακριτά μεγέθη x_t σε συνεχή χρόνο t.
- Συνεχή μεγέθη x_t σε συνεχή χρόνο t.

Οι χρονοσειρές διακρίνονται και ως προς τη διάσταση της μεταβλητής X_t σε μονομεταβλητές ή διανυσματικές χρονοσειρές.

Βασικοί Ορισμοί

Μία χρονοσειρά μπορεί να είναι **ντετερμινιστική** αν καθορίζεται από κάποιο γνωστό κανόνα, δηλαδή, αν οι τιμές της προκύπτουν ως αποτέλεσμα μίας γνωστής συνάρτησης όπως για παράδειγμα η $f(t) = \sin(2\pi t) + \sin(2\sqrt{2}\pi t)$



ΔΗΜΟΚΡΙΤΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΡΑΚΗΣ | DEMOCRITUS UNIVERSITY OF THRACE

Βασικοί Ορισμοί

Τυχαία ή **μη ντετερμινιστική** χρονοσειρά είναι αυτή που δεν μπορεί να περιγραφεί με κάποια συγκεκριμένη συναρτησιακή έκφραση. Μια χρονοσειρά μπορεί να είναι μη ντετερμινιστική επειδή:

• Δεν είναι γνωστές όλες οι πηγές της μεταβλητότητας των τιμών της, όπως π.χ. το πλήθος των ατόμων που θα επισκεφτούν ένα αρχαιολογικό χώρο μία μέρα του Ιούνη.

• Η φύση της διαδικασίας που περιγράφεται με αυτήν είναι εγγενώς τυχαία, όπως π.χ. το πλήθος των σωματιδίων που καταφτάνουν σε έναν ανιχνευτή στραμμένο προς το διάστημα.

Βασικοί Ορισμοί

Καθώς οι μη ντετερμινιστικές χρονοσειρές εξελίσσονται με τυχαίο τρόπο, η περιγραφή της εξέλιξής τους αρμόζει να γίνει με όρους της θεωρίας πιθανοτήτων.

Στο παραπάνω πλαίσιο, θεωρούμε πως πίσω από κάθε τυχαία χρονοσειρά κρύβεται κάποια στοχαστική διεργασία $\{X_t, t = 1, 2, \ldots\}$.

Το σύνολο των πιθανών εξελίξεων μιας στοχαστικής διεργασίας είναι όλος ο στατιστικός πληθυσμός.

Μία παρατηρούμενη χρονοσειρά είναι μια δειγματική μονάδα.

ΔΗΜΟΚΡΙΤΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΡΑΚΗΣ | DEMOCRITUS UNIVERSITY OF THRACE

Στόχοι της ανάλυσης μίας χρονοσειράς

Ο βασικός στόχος της ανάλυσης είναι ο προσδιορισμός ενός μοντέλου που περιγράφει το μοτίβο που ακολουθεί η χρονοσειρά. Ένα τέτοιο μοντέλο, θα μπορούσε να χρησιμοποιηθεί για:

- Να περιγράψει τα σημαντικά χαρακτηριστικά της χρονοσειράς.
- Να εξηγήσει πώς το παρελθόν επηρεάζει το μέλλον.
- Να προβλέψει τις μελλοντικές τιμές της σειράς.

ΔΗΜΟΚΡΙΤΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΡΑΚΗΣ | DEMOCRITUS UNIVERSITY OF THRACE

Σημαντικά ερωτήματα

Μερικά σημαντικά ερωτήματα που προκύπτουν κατά την παρατήρηση μιας χρονοσειράς είναι:

• Υπάρχει **τάση** στη σειρά, δηλαδή, οι μετρήσεις τείνουν κατά μέσο όρο να αυξάνονται (ή να μειώνονται) με την πάροδο του χρόνου;

• Υπάρχει **εποχικότητα**, δηλαδή υπάρχει ένα επαναλαμβανόμενο μοτίβο με υψηλές και χαμηλές τιμές που να σχετίζονται με την χρονική στιγμή;

• Υπάρχουν **ιδιάζουσες** ή **ακραίες** τιμές, δηλαδή τιμές που απέχουν ασυνήθιστα πολύ από τις υπόλοιπες παρατηρήσεις;

• Υπάρχει **μακροχρόνιος κύκλος** ή κάποια άλλη περιοδικότητα που δεν σχετίζεται με εποχιακούς παράγοντες;

• Υπάρχει **μεταβολή στη διακύμανση** των τιμών με την πάροδο του χρόνου;

ΔΗΜΟΚΡΙΤΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΡΑΚΗΣ | DEMOCRITUS UNIVERSITY OF THRACE

Σχεδιασμός ως προς την παρουσίαση της ύλης

Η σειρά με την οποία θα προσεγγίσουμε την διδασκαλία του μαθήματος είναι η εξής:

α) Κατανόηση της έννοιας της στασιμότητας μίας χρονοσειράς.

β) Αναγνώριση της αντικατάστασης των τιμών μίας χρονοσειράς με τη διαφορά τους, ως μία μέθοδο που θεραπεύει το πρόβλημα της τάσης.

γ) Εξοικείωση με τα στατιστικά και τα διαγράμματα που χρησιμοποιούνται στην ανάλυση χρονοσειρών.

δ) Παρουσίαση των κυριότερων αναλυτικών μοντέλων με τα οποία μπορούμε να μοντελοποιήσουμε μία χρονοσειρά.

ε) Κατανόηση των διαθέσιμων μεθόδων υπολογισμού των συντελεστών ενός μοντέλου.

στ) Παρουσίαση των φίλτρων Kalman.

ΔΗΜΟΚΡΙΤΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΡΑΚΗΣ | DEMOCRITUS UNIVERSITY OF THRACE

Εκπαιδευτική διαδικασία

Κατά τη διάρκεια των διαλέξεων θα υπάρχουν μικρές ασκήσεις εμπέδωσης.

Αυτές θα υλοποιούνται είτε σε λογιστικό φύλλο (MS Excel, LibreOffice Calc) είτε με τη γλώσσα R.

Καλό είναι στον υπολογιστή σας να έχετε εγκαταστήσει:

α) Τη γλώσσα R (https://ftp.cc.uoc.gr/mirrors/CRAN/)

β) Το R Studio (https://posit.co/download/rstudio-desktop/)

Κατά τη διάρκεια του εξαμήνου θα δοθούν εργασίες από τις οποίες θα διαμορφωθεί και ο τελικός βαθμός.

ΔΗΜΟΚΡΙΤΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΡΑΚΗΣ
DEMOCRITUS UNIVERSITY OF THRACE

Στασιμότητα μίας χρονοσειράς

Στάσιμες και μη στάσιμες χρονοσειρές

Μία χρονοσειρά ονομάζεται στάσιμη (stationary) όταν τόσο η διακύμανση όσο και η μέση τιμή της δεν μεταβάλλεται στο χρόνο, δηλαδή όταν οι αντίστοιχες τυχαίες μεταβλητές αποτελούν μία στάσιμη ή ασθενώς στάσιμη στοχαστική διεργασία.

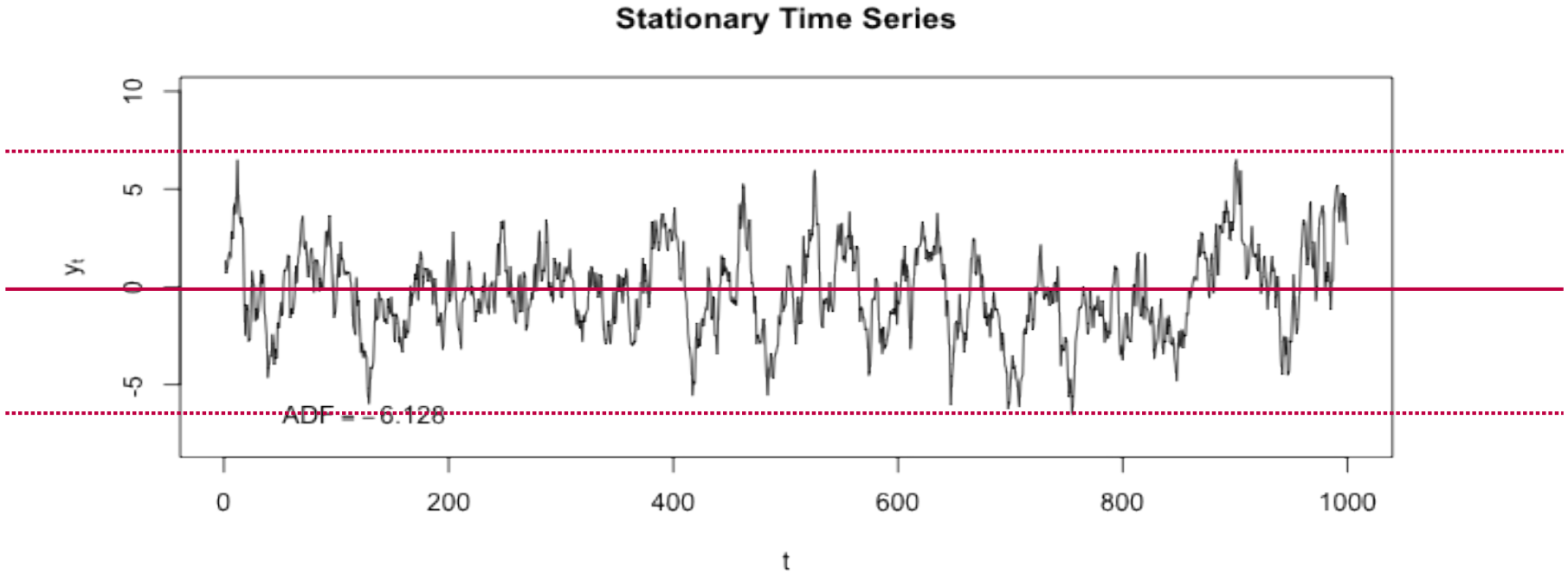
Στην πράξη, μία στάσιμη χρονοσειρά, δεν μπορεί να έχει εποχικότητα, περιοδικότητα και τάση.

Η στασιμότητα είναι μια θεμελιώδης υπόθεση για πολλά μοντέλα χρονοσειρών όπως το ARIMA (Auto Regression Integrated Moving Average) και το SARIMA (Seasonal ARIMA). Αυτά τα μοντέλα βασίζονται στη σταθερότητα των γεωμετρικών μεγεθών της χρονοσειράς όπως ο μέσος όρος και η διακύμανση με την πάροδο του χρόνου. Τα μη στάσιμα δεδομένα μπορεί να οδηγήσουν σε αναξιόπιστες προβλέψεις, ακριβώς γιατί τα μοντέλα δεν το περιμένουν.

ΔΗΜΟΚΡΙΤΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΡΑΚΗΣ
DEMOCRITUS UNIVERSITY OF THRACE

Στάσιμες και μη στάσιμες χρονοσειρές

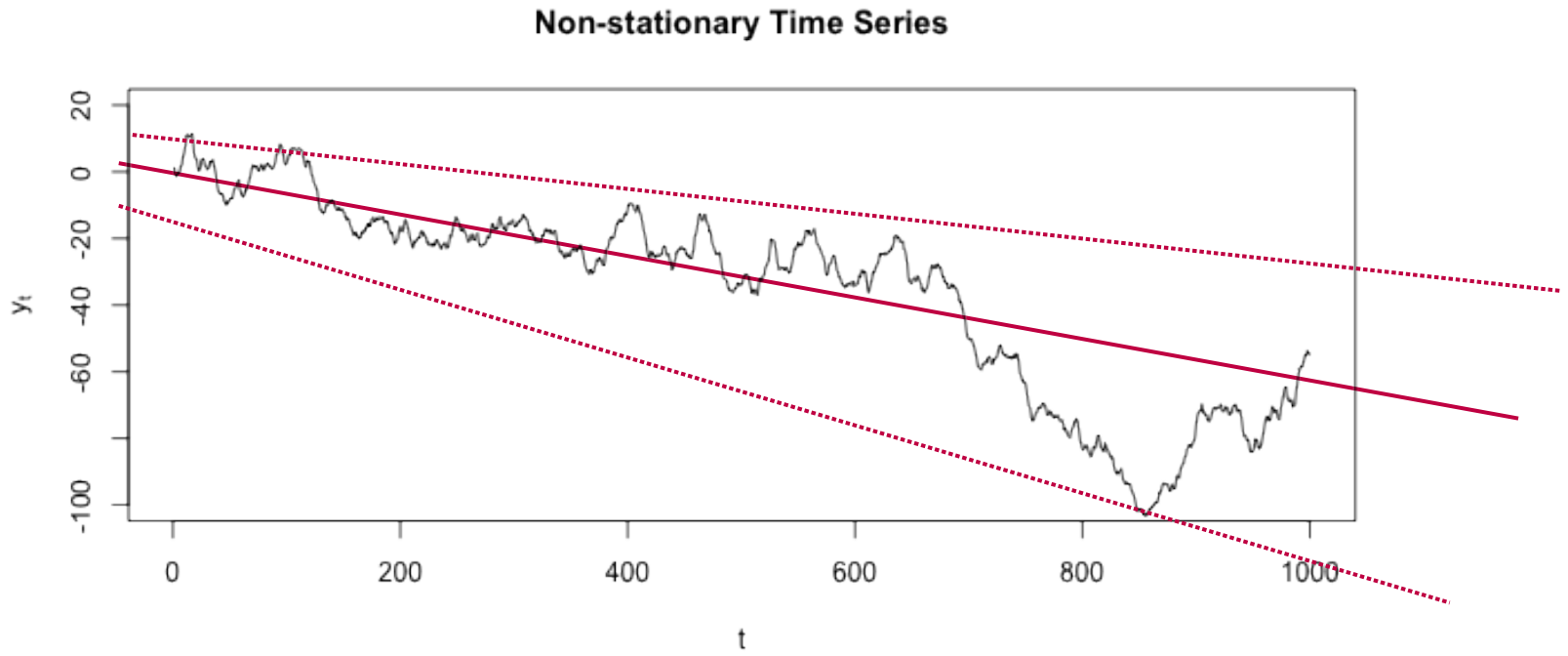
Παράδειγμα χρονοσειράς που προέρχεται από στάσιμη στοχαστική διεργασία



ΔΗΜΟΚΡΙΤΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΡΑΚΗΣ
DEMOCRITUS UNIVERSITY OF THRACE

Στάσιμες και μη στάσιμες χρονοσειρές

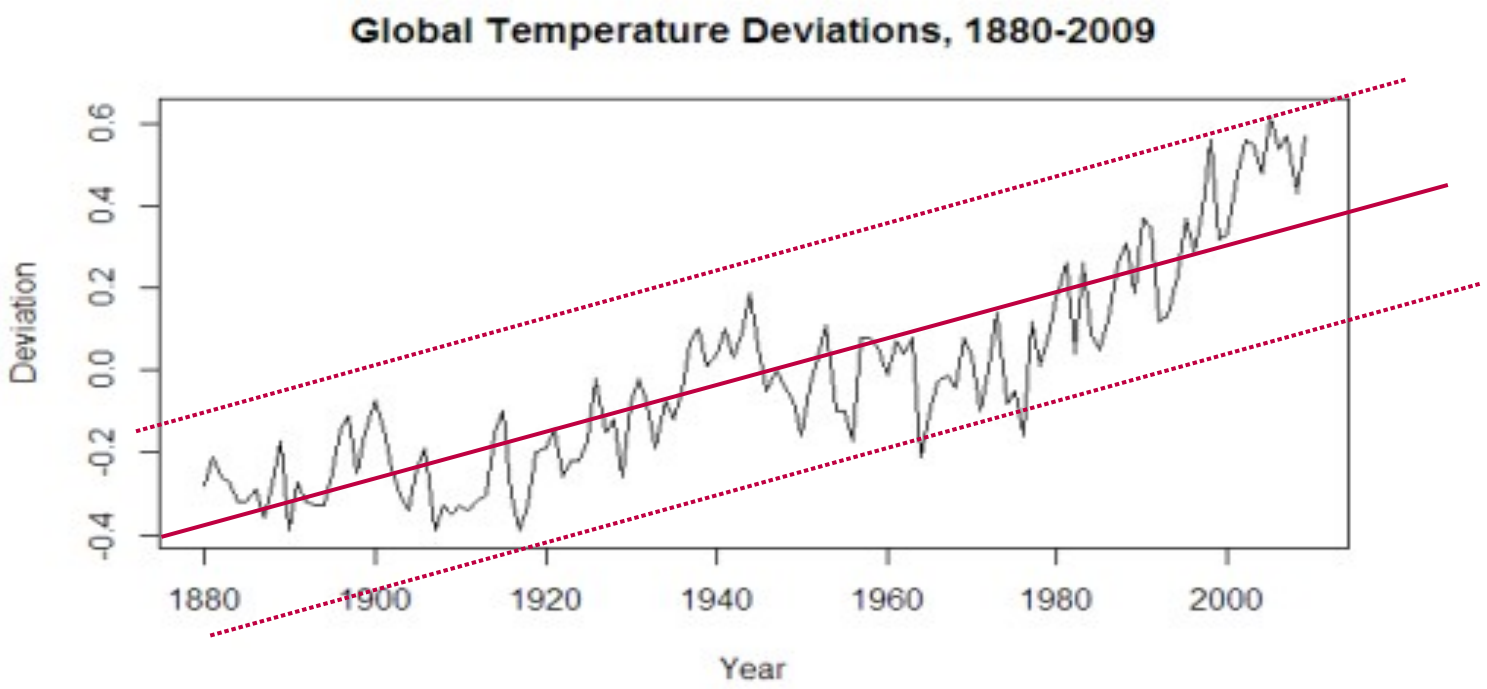
Παράδειγμα χρονοσειράς που προέρχεται από μη στάσιμη στοχαστική διεργασία



ΔΗΜΟΚΡΙΤΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΡΑΚΗΣ
DEMOCRITUS UNIVERSITY OF THRACE

Στάσιμες και μη στάσιμες χρονοσειρές

Παράδειγμα χρονοσειράς που προέρχεται από μη στάσιμη στοχαστική διεργασία



ΔΗΜΟΚΡΙΤΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΡΑΚΗΣ
DEMOCRITUS UNIVERSITY OF THRACE

Στάσιμες και μη στάσιμες χρονοσειρές

Δραστηριότητα

(a) Τιμή μετοχής της Google για 200 συνεχόμενες ημέρες

(b) Ημερήσια αλλαγή στην τιμή της μετοχής της Google για 200 συνεχόμενες ημέρες.

(c) Ετήσιος αριθμός απεργιών στις ΗΠΑ.

(d) Μηνιαίες πωλήσεις νέων μονοκατοικιών που πωλούνται στις ΗΠΑ.

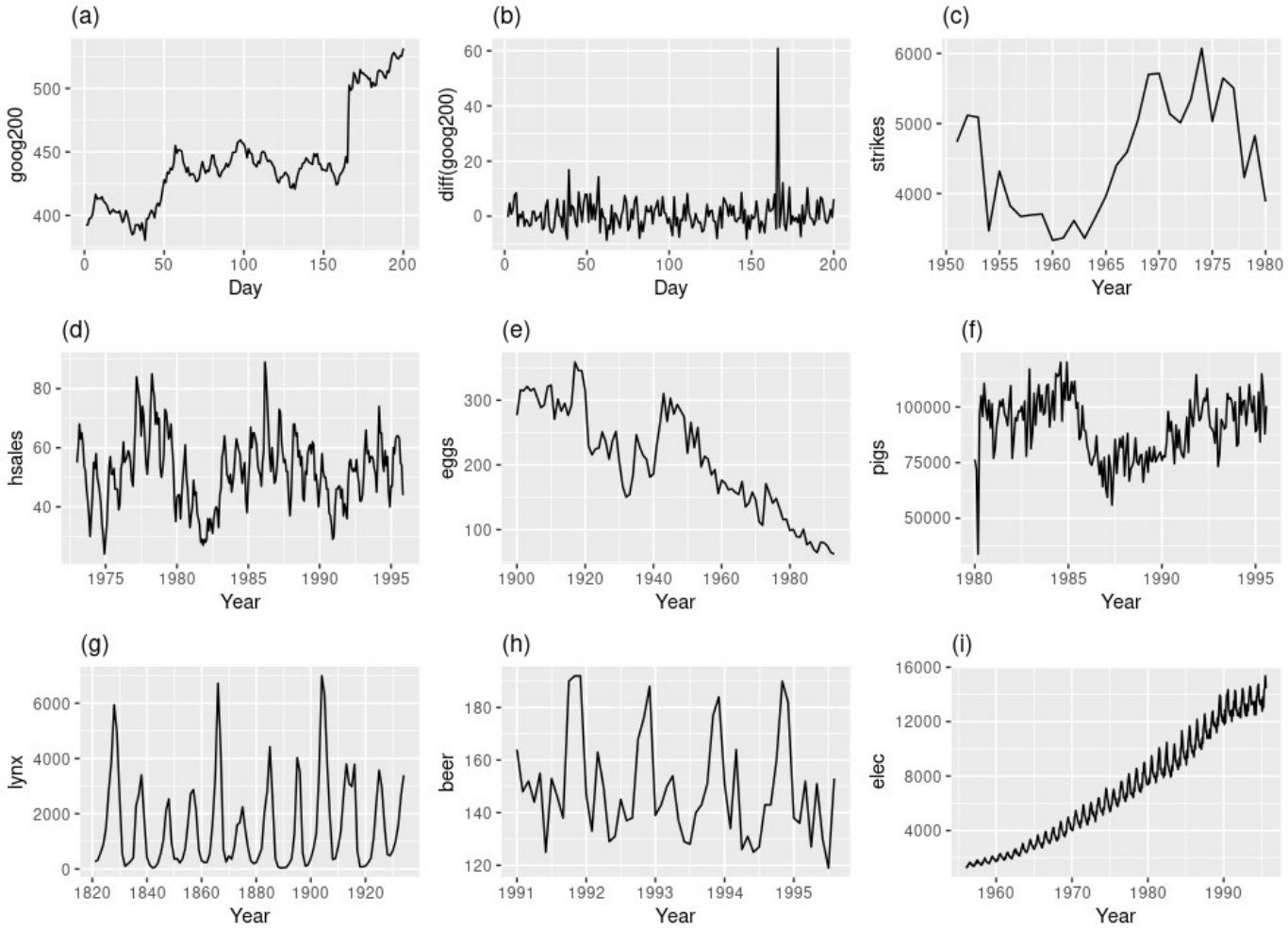
(e) Ετήσια τιμή δώδεκα αυγών στις ΗΠΑ.

(f) Μηνιαίο σύνολο χοίρων που καταναλώνονται στη Βικτώρια της Αυστραλίας.

(g) Πληθυσμός λύγκα που ζει στην περιοχή του ποταμού McKenzie του βορειοδυτικού Καναδά.

(h) Μηνιαία παραγωγή αυστραλιανής μπύρας.

(i) Μηνιαία αυστραλιανή παραγωγή ηλεκτρικής ενέργειας.



Πηγή: https://otexts.com/fpp2/stationarity.html

Στάσιμες και μη στάσιμες χρονοσειρές

Υπάρχουν στατιστικοί έλεγχοι με τις οποίες ελέγχεται η υπόθεση πως η παρατηρούμενη χρονοσειρά αποτελεί εκδοχή μίας στάσιμης διεργασίας, όπως για παράδειγμα η δοκιμασία Kwiatkowski-Phillips-Schmidt-Shin (KPSS) ([1], [2]), και η δοκιμασία Dickey–Fuller ([3]).

Επιπλέον, η δοκιμασία Ljung–Box ([4]) είναι δυνατόν να εφαρμοστεί για να ελεγχθεί η υπόθεση πως στη χρονοσειρά υπάρχει κάποιου είδους αυτοσυσχέτιση.

[1] Kwiatkowski, D., Phillips, P. C. B., Schmidt, P., & Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? Journal of Econometrics, 54(1-3), 159–178.

[2] https://en.wikipedia.org/wiki/KPSS_test

[3] Dickey, D. A.; Fuller, W. A. (1979). "Distribution of the Estimators for Autoregressive Time Series with a Unit Root". Journal of the American Statistical Association. 74 (366): 427–431. doi:10.1080/01621459.1979.10482531. JSTOR 2286348.

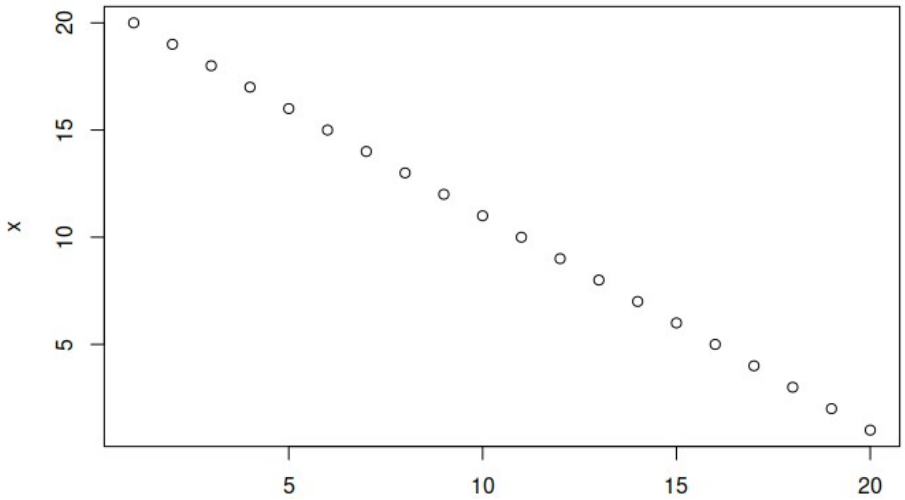
[4] G. M. Ljung; G. E. P. Box (1978). "On a Measure of a Lack of Fit in Time Series Models". Biometrika. 65 (2): 297–303. doi:10.1093/biomet/65.2.297.

ΔΗΜΟΚΡΙΤΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΡΑΚΗΣ | DEMOCRITUS UNIVERSITY OF THRACE

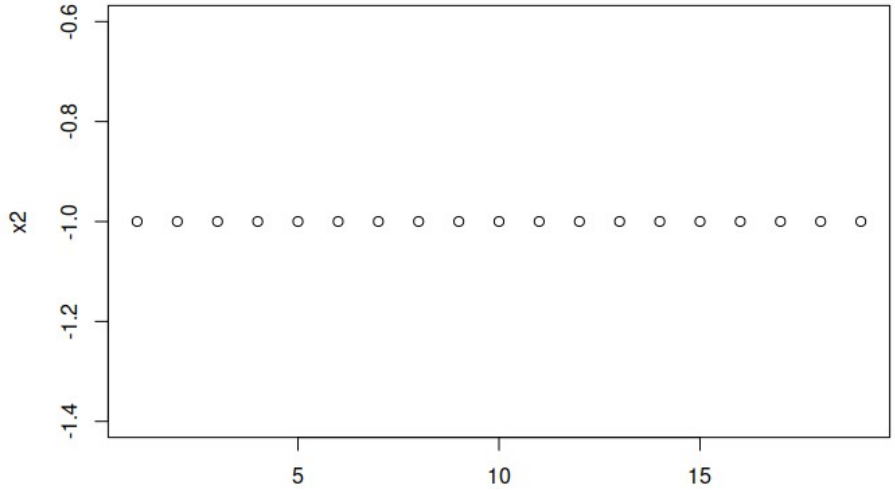
Στάσιμες και μη στάσιμες χρονοσειρές

Μετατροπή μίας μη στάσιμης σε στάσιμη χρονοσειρά.

Αν μία μη στάσιμη χρονοσειρά έχει γραμμική τάση, τότε η χρονοσειρά των διαφορών τείνει να έχει σταθερή μέση τιμή, δηλαδή να είναι στάσιμη.



x = rep(20:1)

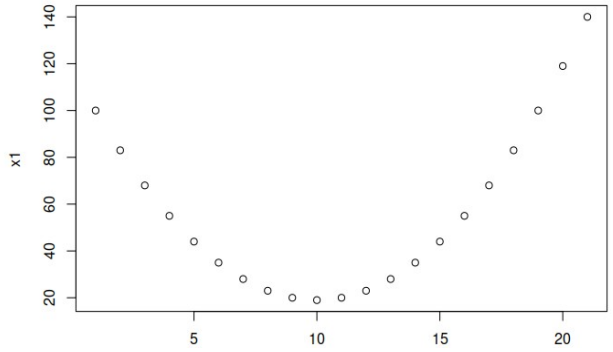


x2 <- diff(x, differences = 1)

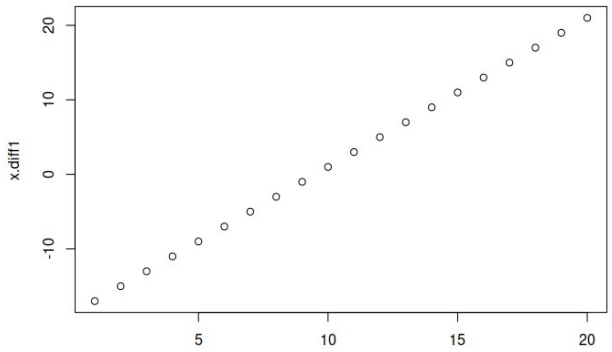
Στάσιμες και μη στάσιμες χρονοσειρές

Μετατροπή μίας μη στάσιμης σε στάσιμη χρονοσειρά.

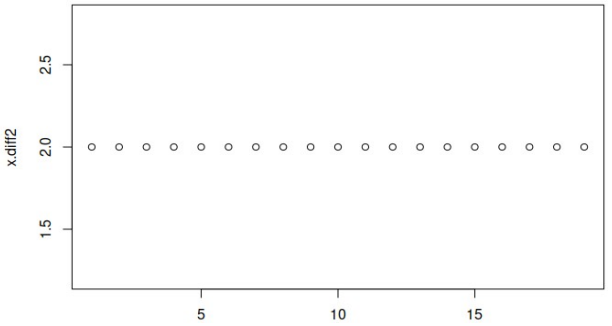
Στην περίπτωση μη γραμμικής τάσης, απαιτείται υπολογισμός διαφορών δεύτερης ή μεγαλύτερης τάξης.



f = function(x) {x^2 + 2*x + 20}
x1 = unlist(lapply(-10:10, f))



x.diff1 <- diff(x1, differences = 1)



x.diff2 <- diff(x1, differences = 2)

Στάσιμες και μη στάσιμες χρονοσειρές

Μία διαφορά 1ης τάξης στα δεδομένα, σημαίνει πως η σειρά που μελετάται είναι αυτή που προκύπτει λαμβάνοντας τις διαφορές διαδοχικών όρων.

$$\{z_1, z_2, z_3, \dots\} = \{x_2 - x_1, x_3 - x_2, x_4 - x_3, \dots\}.$$

Η διαφορά 1ης τάξης επιλέγεται για την απομάκρυνση γραμμικής τάσης στη χρονοσειρά. Στην περίπτωση αυτή, η σειρά λέγεται ότι είναι I(1). (Ολοκληρωμένη – Integrated 1ης τάξης)

Μία διαφορά 2ης τάξης στα δεδομένα, σημαίνει πως η σειρά που μελετάται είναι αυτή που προκύπτει λαμβάνοντας τις διαφορές των διαφορών διαδοχικών όρων.

$$\{w_1, w_2, w_3, \dots\} = \{z_2 - z_1, z_3 - z_2, z_4 - z_3, \dots\}$$

$$= \{(x_3 - x_2) - (x_2 - x_1), (x_4 - x_3) - (x_3 - x_2), (x_5 - x_4) - (x_4 - x_3), \dots\}$$

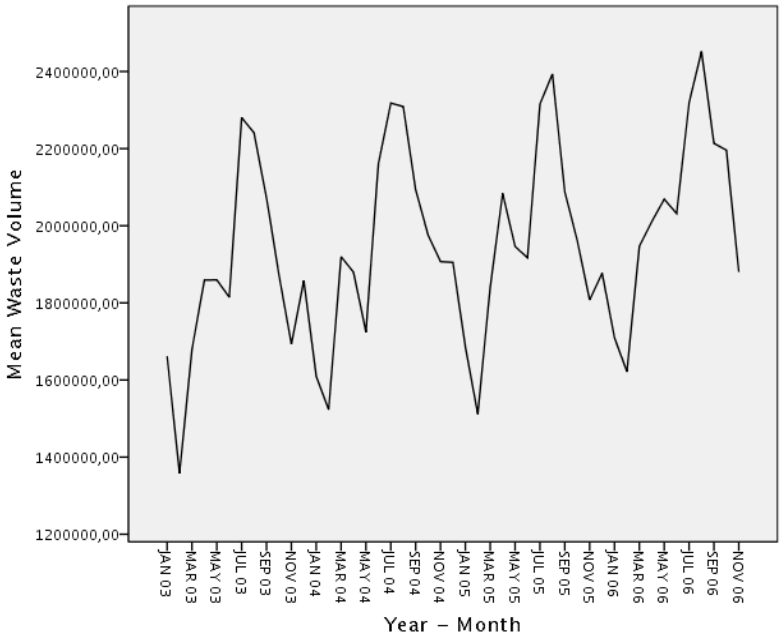
$$= \{(x_3 - 2x_2 - x_1), (x_4 - 2x_3 - x_2), (x_5 - 2x_4 - x_3), \dots\}.$$

Η διαφορά 2ης τάξης επιλέγεται για την ερμηνεία χρονοσειράς με τετραγωνική τάση.

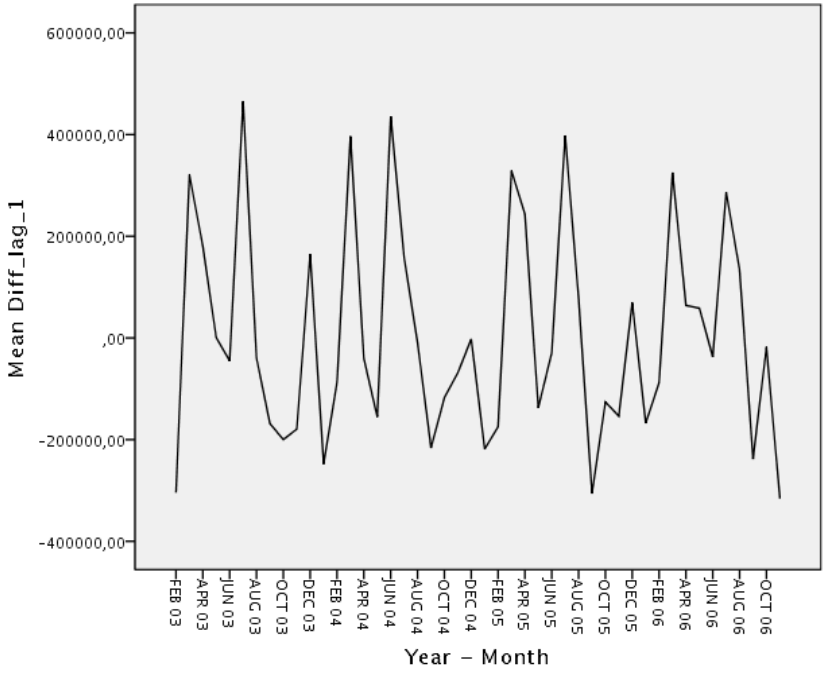
Στην περίπτωση αυτή, η σειρά λέγεται ότι είναι I(2). (Ολοκληρωμένη – Integrated 2ης τάξης)

Στάσιμες και μη στάσιμες χρονοσειρές

Παράδειγμα



$x_{n+1} - x_n$



Ανεξαρτησία (Box-Ljung test):	$x^2(1) = 19.618$, $p < 0,001$
Στασιμότητα (Dickey-Fuller test):	DF.stat = -4.1194, p = 0.013

Ανεξαρτησία (Box-Ljung test):	$x^2(1) = 0.001742$, $p = 0,967$
Στασιμότητα (Dickey-Fuller test):	DF.stat = -2.7841, p = 0.262

Δραστηριότητα Εμπέδωσης

Δίνεται η χρονοσειρά 20 σημείων 1, 3, 3, 6, 8, 9, 10, 9, 11, 13, 12, 10, 14, 17, 18, 17, 20, 19, 19, 21.

(α) Να υπολογίσετε τις διαφορές $1^{ης}$ τάξης.

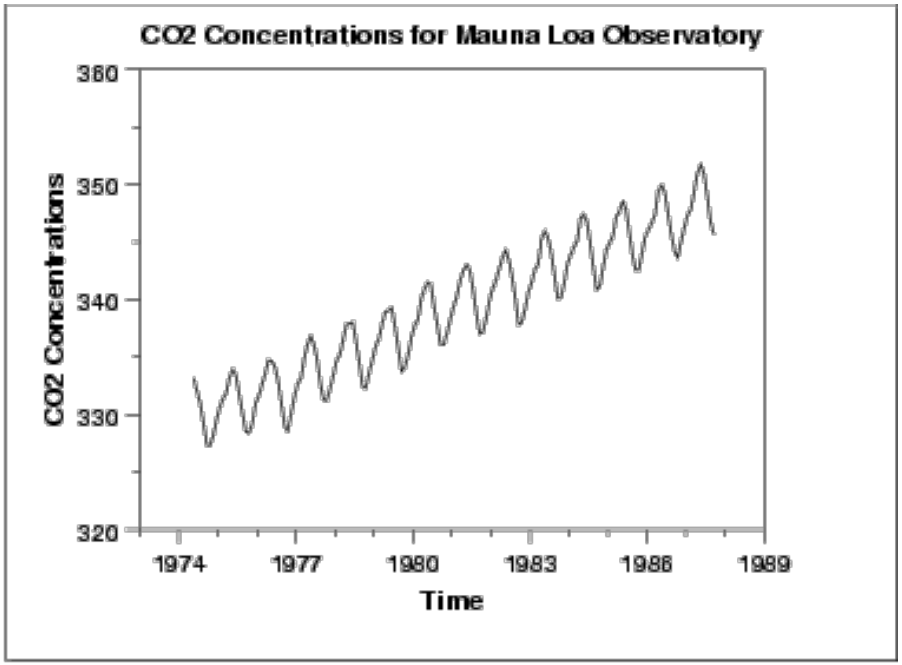
(β) Να αναπαραστήσετε γραφικά την αρχική σειρά και τις διαφορές της σε ένα διάγραμμα.

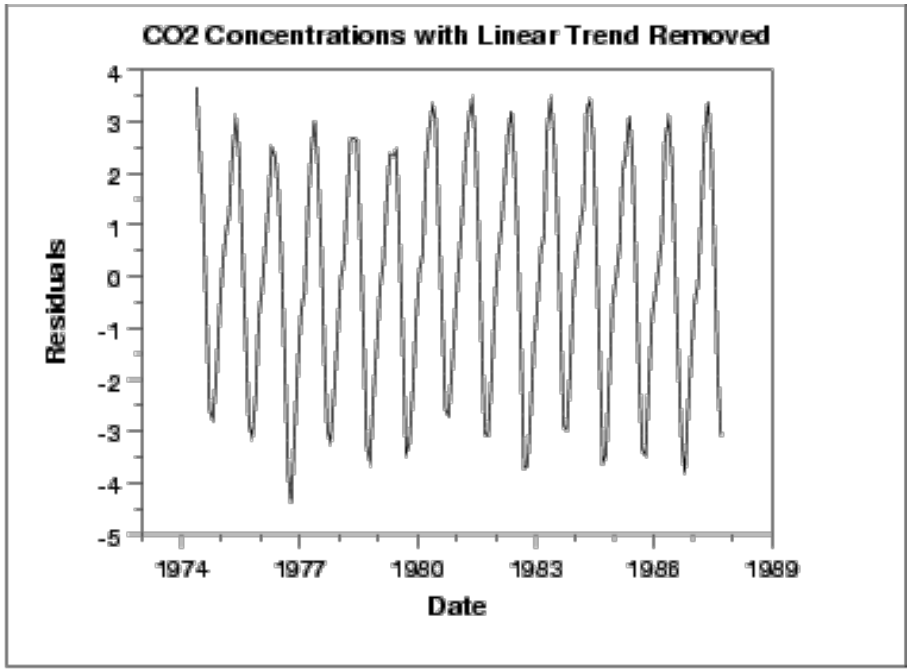
Σημείωση:
Μπορείτε να χρησιμοποιήσετε το MS Excel, το LibreOffice Calc ή τη γλώσσα R.
Στην περίπτωση της γλώσσας R, αρκεί να εκτελέσετε τον κώδικα:

```
x = c(1, 3, 3, 6, 8, 9, 10, 9, 11, 13, 12, 10, 14, 17, 18, 17, 20, 19, 19, 21)
x.diff1 <- diff(x, differences = 1)
plot(1:20, x, type = "l", col = 2, xlab = "Χρόνος", ylab = "Τιμές", ylim = c(-3, 20), lwd = 2)
lines(2:20, x.diff1, type = "l", col = 3, lwd = 2)
legend("topleft", c("Τιμές x_t", " Τιμές Δx_t"), lty = 1, col = 2:3, lwd = 2)
```

Στάσιμες και μη στάσιμες χρονοσειρές

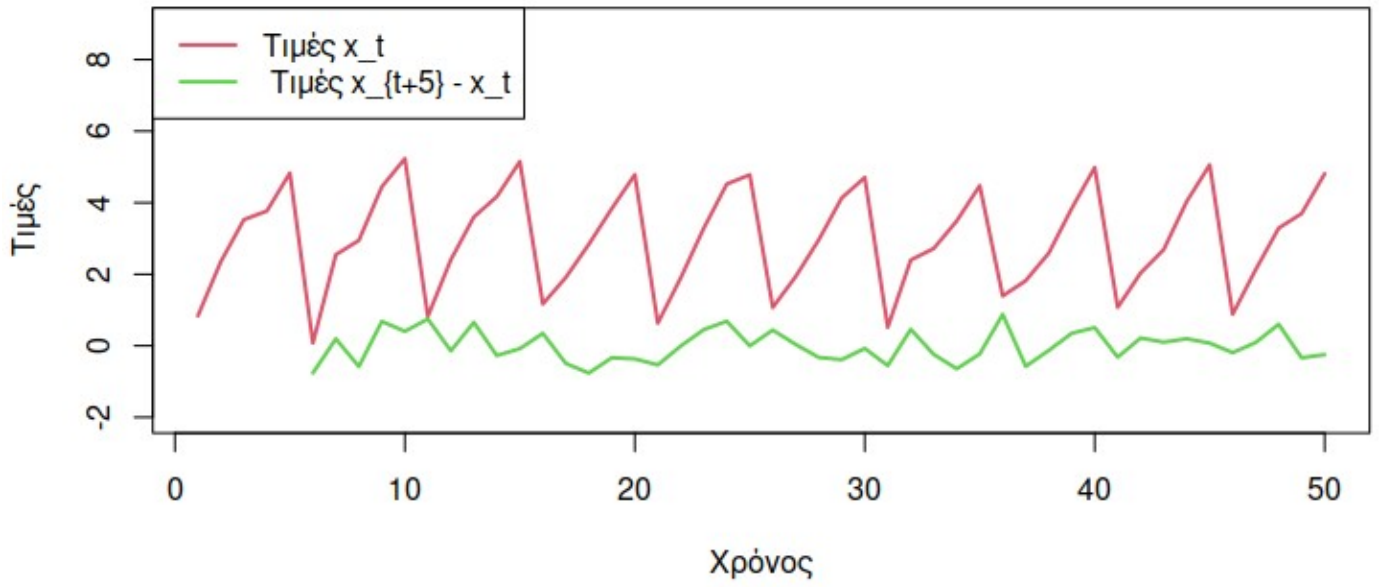
Η διαφορά, αφαιρεί την τάση αλλά όχι την εποχικότητα.





Στάσιμες και μη στάσιμες χρονοσειρές

Η αντικατάσταση της αρχικής χρονοσειράς με τις D τάξης διαφορές της μπορεί να οδηγήσει στην απομάκρυνση περιοδικότητας με περίοδο T = D.



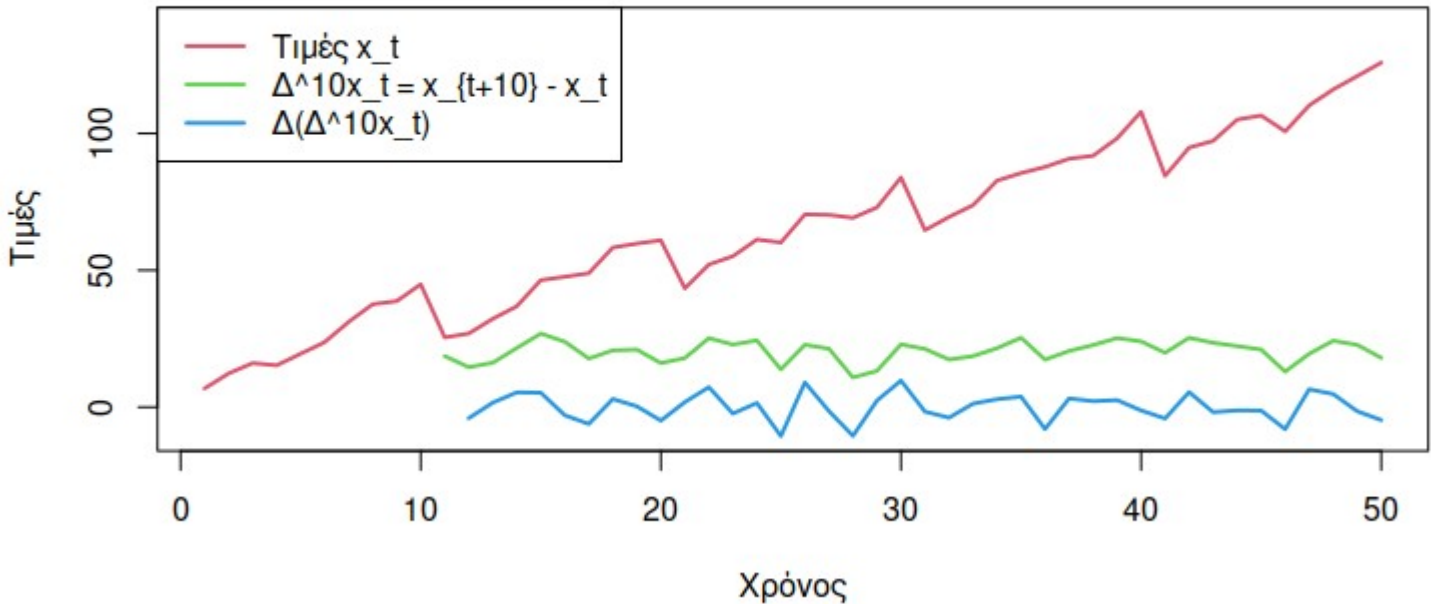
Στάσιμες και μη στάσιμες χρονοσειρές

Κώδικας R

```
x = c(rep(1:5, 10)) + rnorm(50, 0, 0.3)
x.diff1 <- diff(x, differences = 1, lag = 5)
plot(1:50, x, type = "l", col = 2, xlab = "Χρόνος", ylab = "Τιμές", ylim = c(-2, 9), lwd = 2)
lines(6:50, x.diff1, type = "l", col = 3, lwd = 2)
legend("topleft", c("Τιμές x_t", " Τιμές x_{t+5} - x_t"), lty = 1, col = 2:3, lwd = 2)
```

Στάσιμες και μη στάσιμες χρονοσειρές

Ο υπολογισμός της εποχιακής διαφοράς είναι δυνατό να αφαιρέσει παράλληλα και την κύρια τάση.




```
x_trend = 2*c(1:50) + 3 + rnorm(50, 0, 3)
x = x_trend + c(rep(seq(2, 20, 2), 5))

x.D10 <- diff(x, differences = 1, lag = 10)
x.D10.d1 <- diff(x.D10, differences = 1)
```

```
plot(1:50, x, type = "l", col = 2, xlab = "Χρόνος", ylab = "Τιμές", ylim = c(-10, 100), lwd = 2)
lines(11:50, x.D10, type = "l", col = 3, lwd = 2)
lines(12:50, x.D10.d1, type = "l", col = 4, lwd = 2)
legend("topleft", c("Τιμές x_t", "Δ^10x_t = x_{t+10} - x_t", "Δ(Δ^10x_t)"), lty = 1, col = 2:4, lwd = 2)
```

ΔΗΜΟΚΡΙΤΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΡΑΚΗΣ | DEMOCRITUS UNIVERSITY OF THRACE

Δραστηριότητα Εμπέδωσης

Δίνεται η χρονοσειρά 20 σημείων 4, 4, 6, 8, 3, 5, 6, 8, 3, 6, 5, 8, 3, 5, 4, 6, 4, 6, 6, 7.

(α) Να αναπαραστήσετε γραφικά την αρχική σειρά και να ανιχνεύσετε την περιοδικότητά της.

(β) Βρείτε την κατάλληλης τάξης διαφορά για να αφαιρέσετε την περιοδικότητα από τη σειρά.

Σημείωση:
Μπορείτε να χρησιμοποιήσετε το MS Excel, το LibreOffice Calc ή τη γλώσσα R.
Στην περίπτωση της γλώσσας R, αρκεί να εκτελέσετε τον κώδικα (αντικαταστήστε όπου D την κατάλληλη περίοδο):

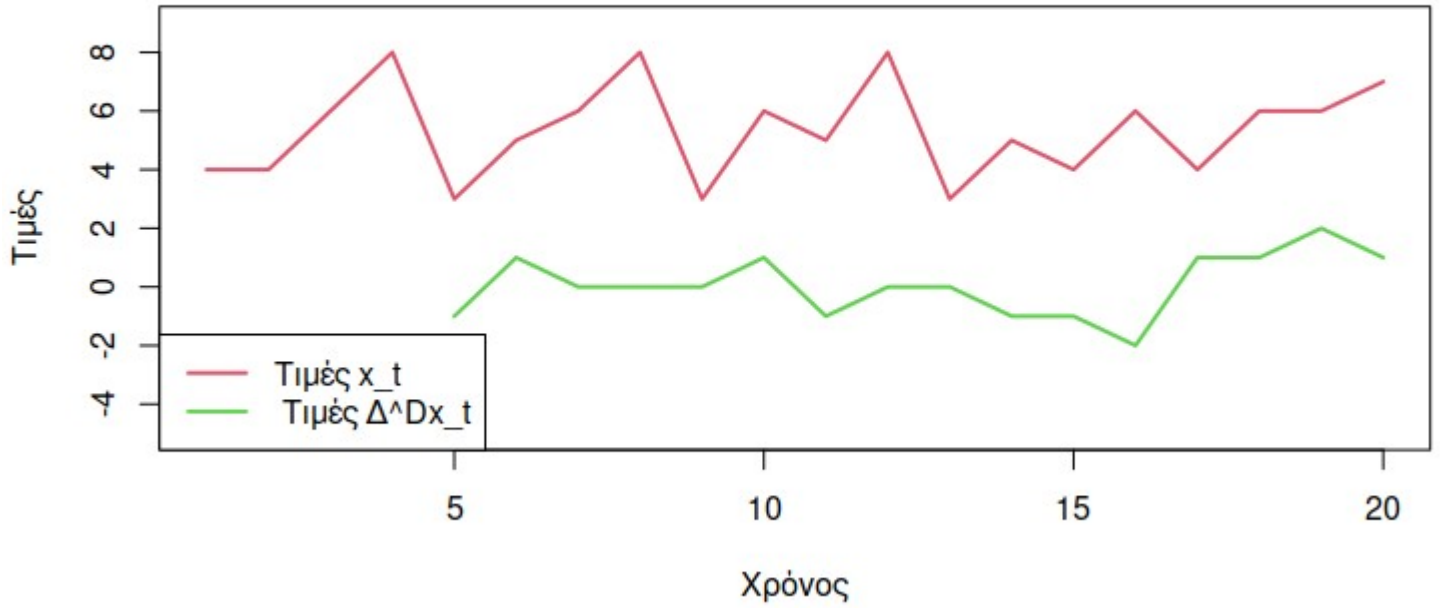
```
x = c(4, 4, 6, 8, 3, 5, 6, 8, 3, 6, 5, 8, 3, 5, 4, 6, 4, 6, 6, 7)
x.D <- diff(x, differences = 1, lag = D)
plot(1:20, x, type = "l", col = 2, xlab = "Χρόνος", ylab = "Τιμές", ylim = c(-5, 9), lwd = 2)
lines(5:20, x.D, type = "l", col = 3, lwd = 2)
legend("bottomleft", c("Τιμές x_t", " Τιμές Δ^Dx_t"), lty = 1, col = 2:3, lwd = 2)
```

Δραστηριότητα Εμπέδωσης

Δίνεται η χρονοσειρά 20 σημείων 4, 4, 6, 8, 3, 5, 6, 8, 3, 6, 5, 8, 3, 5, 4, 6, 4, 6, 6, 7.

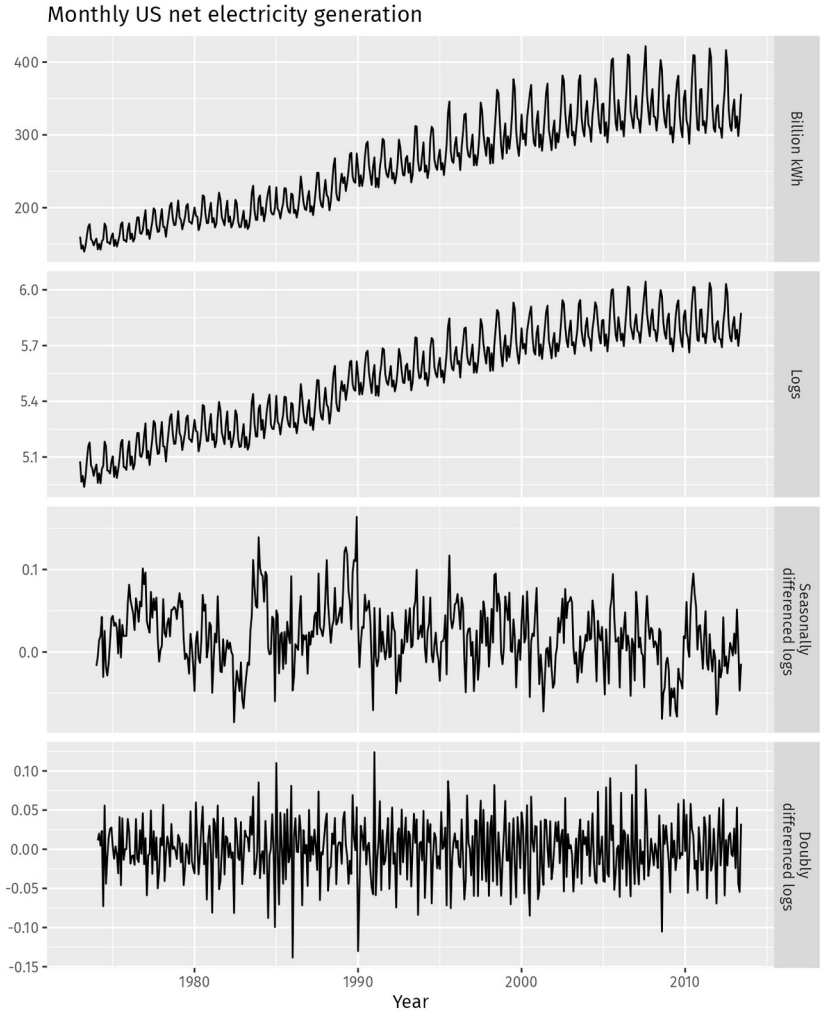
(α) Να αναπαραστήσετε γραφικά την αρχική σειρά και να ανιχνεύσετε την περιοδικότητά της.

(β) Βρείτε την κατάλληλης τάξης διαφορά για να αφαιρέσετε την περιοδικότητα από τη σειρά.



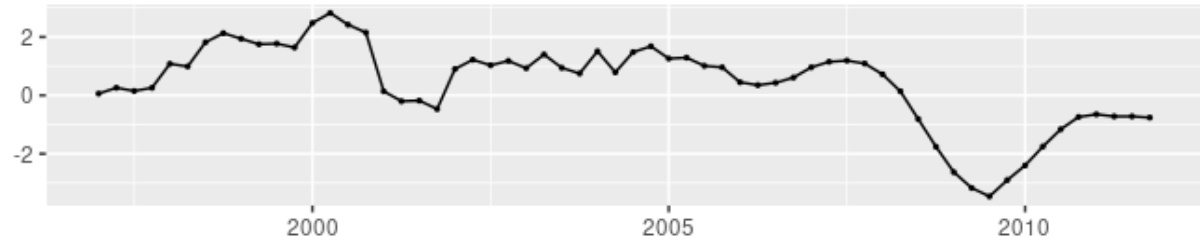
Στάσιμες και μη στάσιμες χρονοσειρές

Στην περίπτωση όπου η κύρια τάση παραμένει, είναι δυνατός ο μετασχηματισμός της σειράς, ή ο επιπλέον υπολογισμός διαφοράς 1ης ή 2ης τάξης.

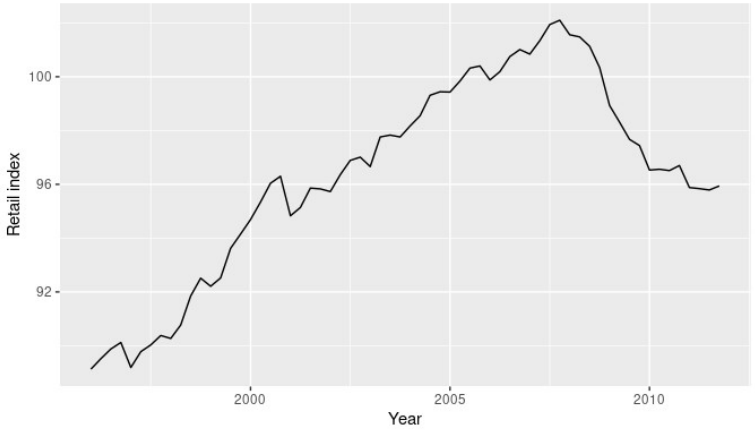


Πηγή διαγράμματος: https://otexts.com/fpp2/stationarity.html

Στάσιμες και μη στάσιμες χρονοσειρές



εποχιακή διαφορά τεσσάρων περιόδων (m = 4 τρίμηνα = 1 έτος)



τριμηνιαία στοιχεία του ευρωπαϊκού λιανικού εμπορίου από το 1996 έως το 2011.

πρώτη διαφορά, των d τιμών

Πηγή διαγράμματος: https://otexts.com/fpp2/seasonal-arima.html

Στάσιμες και μη στάσιμες χρονοσειρές

Μετατροπή μίας μη στάσιμης σε στάσιμη χρονοσειρά.

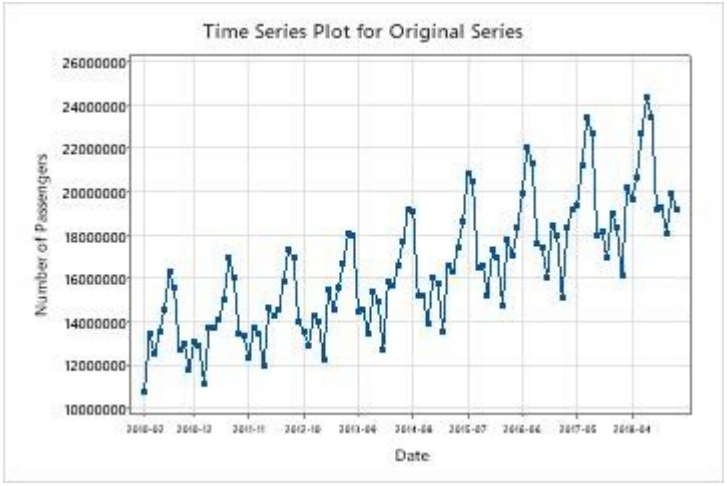
Στην περίπτωση μεταβλητής διακύμανσης, μπορεί να γίνει εφαρμογή στις τιμές με το λογάριθμο ή την τετραγωνική ρίζα ή ακόμα και να εφαρμοστεί ο μετασχηματισμός Box – Cox με τον οποίο εξομαλύνεται η διακύμανση.

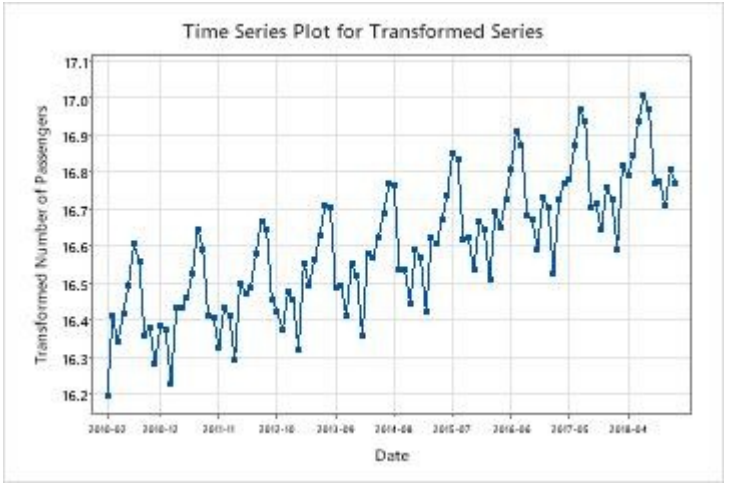
$$y_i^{(\lambda)} = \begin{cases} \dfrac{y_i^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0, \\ \ln y_i & \text{if } \lambda = 0, \end{cases}$$

Στάσιμες και μη στάσιμες χρονοσειρές

Μετατροπή μίας μη στάσιμης σε στάσιμη χρονοσειρά.

Στην περίπτωση μεταβλητής διακύμανσης, μπορεί να γίνει εφαρμογή στις τιμές με το λογάριθμο ή την τετραγωνική ρίζα ή ακόμα και να εφαρμοστεί ο μετασχηματισμός Box – Cox με τον οποίο εξομαλύνεται η διακύμανση.

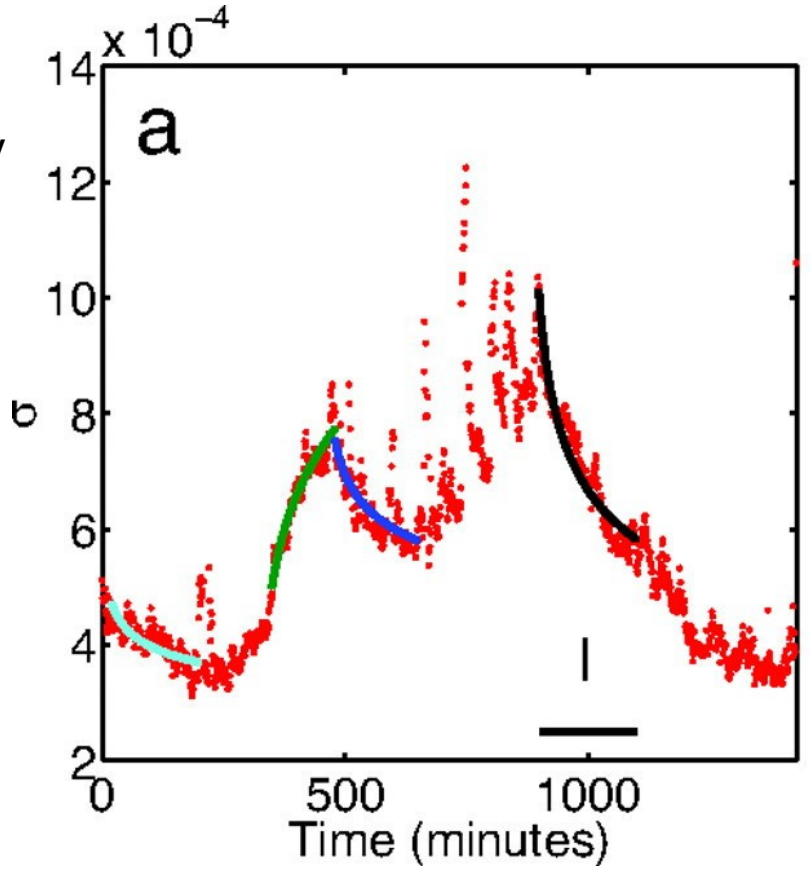




ΔΗΜΟΚΡΙΤΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΡΑΚΗΣ | DEMOCRITUS UNIVERSITY OF THRACE

Στάσιμες και μη στάσιμες χρονοσειρές

Διάγραμμα που παρουσιάζει την τυπική απόκλιση (σ) της χρονοσειράς που εκφράζει την ισοτιμία ευρώ – δολαρίου σε ημερήσια βάση (1440 λεπτά) μεταξύ 1999 και 2004([1]).



[1]. Kevin E. Bassler, Joseph L. McCauley, Gemunu H. Gunaratne (2007). Nonstationary increments, scaling distributions, and variable diffusion processes in financial markets. Proceedings of the National Academy of Sciences Oct 2007, 104 (44) 17287-17290; DOI: 10.1073/pnas.0708664104

ΔΗΜΟΚΡΙΤΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΡΑΚΗΣ | DEMOCRITUS UNIVERSITY OF THRACE

Στατιστικά και Διαγράμματα που Χρησιμοποιούμε στην Ανάλυση Χρονοσειρών

Στατιστικά και Διαγράμματα που Χρησιμοποιούμε στην Ανάλυση Χρονοσειρών

Συντελεστής Συσχέτισης Pearson

Η συνδιακύμανση Cov(X, Y) είναι μία στατιστική ποσότητα που μας επιτρέπει να ποσοτικοποιήσουμε τον τρόπο με τον οποίο συμμεταβάλλονται δύο τυχαίες μεταβλητές X, Y. Η κανονικοποιημένη εκδοχή της συνδιακύμανσης είναι ο συντελεστής συσχέτισης Pearson $\rho_{X,Y}$, ο οποίος παίρνει τιμές από -1 έως +1.

$$\rho_{X,Y} = \operatorname{corr}(X,Y) = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{\mathrm{E}[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

Δειγματικός Συντελεστής Συσχέτισης

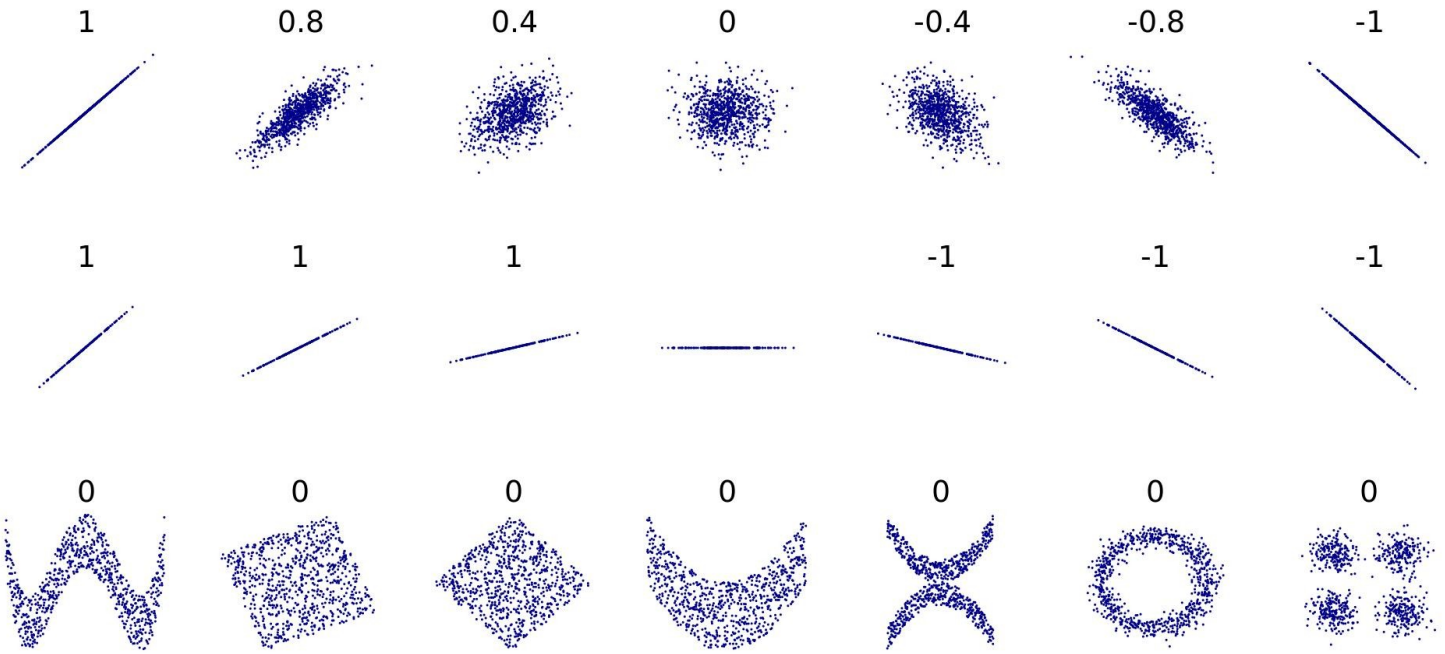
$$r_{xy} \overset{\text{def}}{=} \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{(n-1)s_x s_y} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \sum_{i=1}^{n}(y_i-\bar{y})^2}},$$

ΔΗΜΟΚΡΙΤΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΡΑΚΗΣ
DEMOCRITUS UNIVERSITY OF THRACE

Στατιστικά και Διαγράμματα που Χρησιμοποιούμε στην Ανάλυση Χρονοσειρών

Συντελεστής Συσχέτισης Pearson

$$r_{xy} \overset{\text{def}}{=} \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$



Στατιστικά και Διαγράμματα που Χρησιμοποιούμε στην Ανάλυση Χρονοσειρών

Μερικός Συντελεστής Συσχέτισης (Partial Correlation Coefficient)

Ο μερικός συντελεστής συσχέτισης (partial correlation coefficient) υπολογίζει τη συσχέτιση δύο μεταβλητών X, Y ύστερα από την απομάκρυνση της μεταβλητότητας των X, Y που εξηγείται από μία τρίτη μεταβλητή Z.

$$\rho_{X,\,Y\,|\,Z} = \frac{\mathrm{Cov}\left(X,\,Y\,|\,Z\right)}{\sigma_{X\,|\,Z}\cdot\sigma_{Y\,|\,Z}}$$

Ο μερικός συντελεστής συσχέτισης δίνει περισσότερο ακριβή εικόνα για τη σχέση δύο μεταβλητών στην περίπτωση που είναι γνωστή μία τρίτη μεταβλητή που συσχετίζεται και με τις δύο. Ο υπολογισμός του μερικού συντελεστή συσχέτισης μπορεί να γενικευθεί για περισσότερες της μίας μεταβλητής, Z_1, Z_2, …, Z_n.

$$\rho_{X,\,Y\,|\,Z_1,\,Z_2,\,\ldots,\,Z_n} = \frac{\mathrm{Cov}\left(X,\,Y\,|\,Z_1,\,Z_2,\,\ldots,\,Z_n\right)}{\sigma_{X\,|\,Z_1,\,Z_2,\,\ldots,\,Z_n}\cdot\sigma_{Y\,|\,Z_1,\,Z_2,\,\ldots,\,Z_n}}$$

Μερικός Συντελεστής Συσχέτισης (Partial Correlation Coefficient)

Παράδειγμα Ι

Σε ένα δείγμα μαθητών δημοτικού, αν

X: Ικανότητα Απομνημόνευσης, Y: Ρυθμός ομιλίας, Z: Ηλικία,

τότε ο απλός συντελεστής συσχέτισης Pearson $\rho_{X,\ Y}$, δίνει μία εκτίμηση για τη σχέση των δύο μεταβλητών. Καθώς όμως, τόσο η ικανότητα απομνημόνευσης όσο και ο ρυθμός ομιλίας εξαρτώνται από την ηλικία (Z) του παιδιού, ο μερικός συντελεστής συσχέτισης $\rho_{X,\ Y\ |\ Z}$ αποτελεί περισσότερο ακριβής εκτίμηση για την πραγματική σχέση των δύο μεταβλητών.

Μερικός Συντελεστής Συσχέτισης (Partial Correlation Coefficient)

Παράδειγμα ΙΙ

Στα πλαίσια μίας έρευνας σε 50 πόλεις των Η.Π.Α. καταγράφηκαν:

- οι επενδύσεις στο χώρο της δημόσιας υγείας (funding)
- το πλήθος των αναφερόμενων ασθενών (disease)
- το πλήθος των επισκέψεων στα νοσοκομεία και τα κέντρα υγείας (visits).

Από τα δεδομένα προέκυψε πως οι επενδύσεις στο χώρο της υγείας είναι ισχυρά θετικά συσχετισμένες με το πλήθος των ασθενειών που αναφέρονται ($r(50) = 0,737$, $p < 0,001$).

Αν η συσχέτιση ερμηνευθεί ως αιτιότητα τότε αυτό θα σήμαινε πως δεν πρέπει να δαπανούμε χρήματα για την υγεία καθώς αυτό συνδυάζεται με αύξηση των ασθενών άρα αποβαίνει εις βάρος των πολιτών, ένα συμπέρασμα προφανώς λανθασμένο.

Μερικός Συντελεστής Συσχέτισης (Partial Correlation Coefficient)

Παράδειγμα ΙΙ

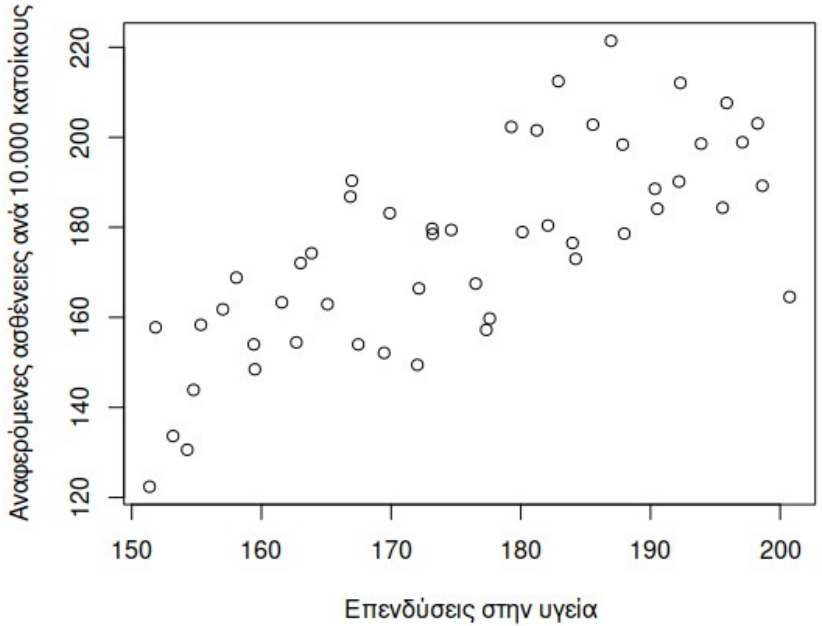
Η κατάσταση γίνεται πιο σαφής αν μελετηθεί επιπλέον

- η σχέση των επενδύσεων με το πλήθος επισκέψεων στις δομές υγείας.
- η σχέση των επισκέψεων στις δομές υγείας με το πλήθος των αναφερόμενων ασθενειών.

Υποδεικνύεται ότι η παρατηρούμενη θετική σχέση μεταξύ των δαπανών για την υγεία και των ασθενειών που δηλώνονται οφείλεται στην ενδιάμεση σχέση μεταξύ κάθε μίας από τις μεταβλητές αυτές με το ρυθμό επίσκεψης στις δομές υγείας, η οποία είναι ισχυρά θετική.

ΔΗΜΟΚΡΙΤΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΡΑΚΗΣ | DEMOCRITUS UNIVERSITY OF THRACE

Στατιστικά και Διαγράμματα που Χρησιμοποιούμε στην Ανάλυση Χρονοσειρών

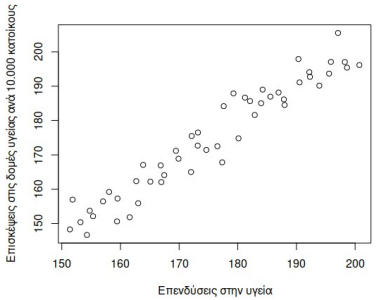


Επενδύσεις → Ασθένειες

r(50) = 0,964
(p < 0,001)

Δεν είναι πραγματικό!

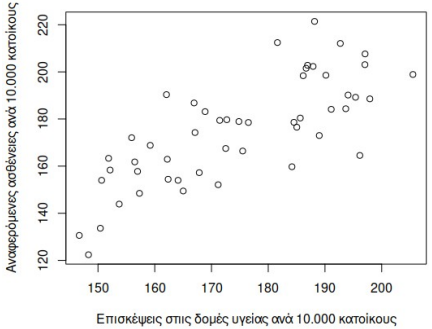
Στατιστικά και Διαγράμματα που Χρησιμοποιούμε στην Ανάλυση Χρονοσειρών



Επισκέψεις

r(50) = 0,964
(p < 0,001)

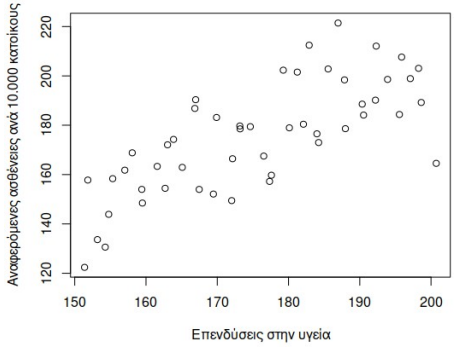
r(50) = 0,762
(p < 0,001)



Επενδύσεις

Ασθένειες

$r_p(50) = 0,013$
(p = 0,928)



ΔΗΜΟΚΡΙΤΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΡΑΚΗΣ | DEMOCRITUS UNIVERSITY OF THRACE

Στατιστικά και Διαγράμματα που Χρησιμοποιούμε στην Ανάλυση Χρονοσειρών

Αυτοδιακύμανση και Αυτοσυσχέτιση

Η αυτοδιακύμανση (autocovariance) και η αυτοσυσχέτιση (autocorrelation) είναι δύο στατιστικές ποσότητες που μας επιτρέπουν να ανιχνεύσουμε εξάρτηση των τιμών μίας χρονοσειράς $\{X_n, n \in N^*\} = \{X_1, X_2, X_3, \ldots\}$ με ένα μετρήσιμο τρόπο. Ειδικότερα, με τον τρόπο αυτό ανιχνεύουμε περιοδικότητα.

Η **αυτοδιακύμανση** είναι η συνδιακύμανση της μεταβλητής X_t των τιμών της χρονοσειράς με ένα αντίγραφό της X_{t-h} χρονικά μετατοπισμένο κατά πλήθος χρονικών στιγμών h = 1, 2, …

h = 1: $\mathrm{Cov}(X_t, X_{t-1}) = \mathrm{Cov}(\{X_1, X_2, X_3, \ldots\}, \{X_2, X_3, X_4, \ldots\})$

h = 2: $\mathrm{Cov}(X_t, X_{t-2}) = \mathrm{Cov}(\{X_1, X_2, X_3, \ldots\}, \{X_3, X_4, X_5, \ldots\})$

….

h = n: $\mathrm{Cov}(X_t, X_{t-n}) = \mathrm{Cov}(\{X_1, X_2, X_3, \ldots\}, \{X_{n+1}, X_{n+2}, X_{n+3}, \ldots\})$

Σημείωση: Φανερά, κάθε μία αυτοδιακύμανση υπολογίζεται σε πλήθος παρατηρήσεων μειωμένο κατά ένα από την προηγούμενη.

Στατιστικά και Διαγράμματα που Χρησιμοποιούμε στην Ανάλυση Χρονοσειρών

Αυτοδιακύμανση και Αυτοσυσχέτιση

Η αυτοσυσχέτιση είναι η κανονικοποιημένη ως προς το πιθανό εύρος τιμών εκδοχή της αυτοδιακύμανσης.

$$\rho_1 = \text{Corr}(X_t, X_{t-1}) = \text{Cov}(X_t, X_{t-1}) / [\text{SD}(X_t)\text{SD}(X_{t-1})]$$

$$\rho_2 = \text{Corr}(X_t, X_{t-2}) = \text{Cov}(X_t, X_{t-2}) / [\text{SD}(X_t)\text{SD}(X_{t-2})]$$

$$\dots$$

$$\rho_n = \text{Corr}(X_t, X_{t-n}) = \text{Cov}(X_t, X_{t-n}) /[\text{SD}(X_t)\text{SD}(X_{t-n})]$$

Στην περίπτωση όπου $\{x_1, x_2, x_3, \dots\}$ είναι συγκεκριμένες αριθμητικές τιμές τότε με αυτές εκφράζεται μία εκδοχή της στοχαστικής διεργασίας $\{X_n, n \in N^*\} = \{X_1, X_2, X_3, \dots\}$ που τις παράγει.

Στην περίπτωση αυτή η $\text{Cov}(x_t, x_{t-n})$ ονομάζεται **δειγματική αυτοδιακύμανση** και η $\text{Corr}(x_t, x_{t-n})$ **δειγματική αυτοσυσχέτιση**.

Στατιστικά και Διαγράμματα που Χρησιμοποιούμε στην Ανάλυση Χρονοσειρών

Δειγματική Συνάρτηση Αυτοσυσχέτισης (ACF)

Ο υπολογισμός της αυτοσυσχέτισης για όλες τις υστερήσεις Lag = 1, 2, 3, …, οδηγεί στον ορισμό της δειγματικής συνάρτησης αυτοσυσχέτισης (Autocorrelation Function – ACF). Πιο συγκεκριμένα, αν x_1, x_2, x_3, …, x_n είναι τα χρονικά δεδομένα, τότε η ACF είναι η συνάρτηση $r(h) = r_h: N \rightarrow R$, με

$$r_0 = \text{Corr}(\{x_1, x_2, x_3, \ldots, x_n\}, \{x_1, x_2, x_3, \ldots, x_n\}),$$

$$r_1 = \text{Corr}(\{x_1, x_2, x_3, \ldots, x_{n-1}\}, \{x_2, x_3, x_4, \ldots, x_n\}),$$

$$r_2 = \text{Corr}(\{x_1, x_2, x_3, \ldots, x_{n-2}\}, \{x_3, x_4, x_5, \ldots, x_n\}),$$

….

$$r_h = \text{Corr}(\{x_1, x_2, x_3, \ldots, x_{n-h}\}, \{x_{h+1}, x_{h+2}, x_{h+3}, \ldots, x_n\}),$$

ΔΗΜΟΚΡΙΤΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΡΑΚΗΣ | DEMOCRITUS UNIVERSITY OF THRACE

Στατιστικά και Διαγράμματα που Χρησιμοποιούμε στην Ανάλυση Χρονοσειρών

Τρόπος υπολογισμού r_h για τη χρονοσειρά $\{x_1, x_2, x_3, \ldots, x_n\}$

1ο βήμα: Υπολογισμός γενικού αριθμητικού μέσου

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

2ο βήμα: Υπολογισμός δειγματικής αυτοσυσχέτισης

$$r_h = \frac{\mathrm{Cov}(x_n, x_{n-h})}{\mathrm{Var}(x_n)} = \frac{\sum_{i=1+h}^{n} (x_i - \bar{x})(x_{i-h} - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}, \; h = 0, 1, , \ldots, n-1.$$

Η τιμή της r_h, μειώνεται καθώς το h αυξάνεται. Στην πράξη, υπολογίζουμε την αυτοσυσχέτιση της χρονοσειράς με τον εαυτό της σε όλες τις πιθανές μετατοπίσεις και παρατηρούμε τις τιμές που προκύπτουν. Συνήθως, αυτές απεικονίζονται σε ένα διάγραμμα, το Correlogram (διάγραμμα αυτοσυσχετίσεων ή διάγραμμα ACF).

Στατιστικά και Διαγράμματα που Χρησιμοποιούμε στην Ανάλυση Χρονοσειρών

Δειγματική Συνάρτησης Μερικής Αυτοσυσχέτισης (PACF)

Αν $x_1, x_2, x_3, \ldots, x_n$ είναι τα χρονικά δεδομένα, τότε η συνάρτηση μερικής αυτοσυσχέτισης φ_h με υστέρηση (lag) h, (Partial Autocorrelation Function – PACF) είναι η αυτοσυσχέτιση μεταξύ των $x_1, x_2, x_3, \ldots, x_{n-h}$ και $x_{h+1}, x_{h+2}, x_{h+3}, \ldots, x_n$ από την οποία αφαιρείται η γραμμική εξάρτηση όλων των ενδιάμεσων υστερήσεων 1, 2, …, h – 1.

Δηλαδή:

$$\varphi_1 = \text{Corr}(x_{t+1}, x_t) = r_1, \quad \varphi_2 = \text{Corr}\left(x_{t+2} - \hat{x}_{t+2}, x_t - \hat{x}_t\right), \quad \varphi_3 = \text{Corr}\left(x_{t+3} - \hat{x}_{t+3}, x_t - \hat{x}_t\right), \ldots$$

Όπου

α) Για το φ_2, $\hat{x}_{t+2}$ και $\hat{x}_t$ είναι εκτιμητές ελαχίστων τετραγώνων με ανεξάρτητη μεταβλητή το x_{t+1}

β) Για το φ_3, $\hat{x}_{t+3}$ και $\hat{x}_t$ είναι εκτιμητές ελαχίστων τετραγώνων με ανεξάρτητες μεταβλητές τις x_{t+1}, x_{t+2},

κλπ...

Στατιστικά και Διαγράμματα που Χρησιμοποιούμε στην Ανάλυση Χρονοσειρών

Ερμηνεία του ACF

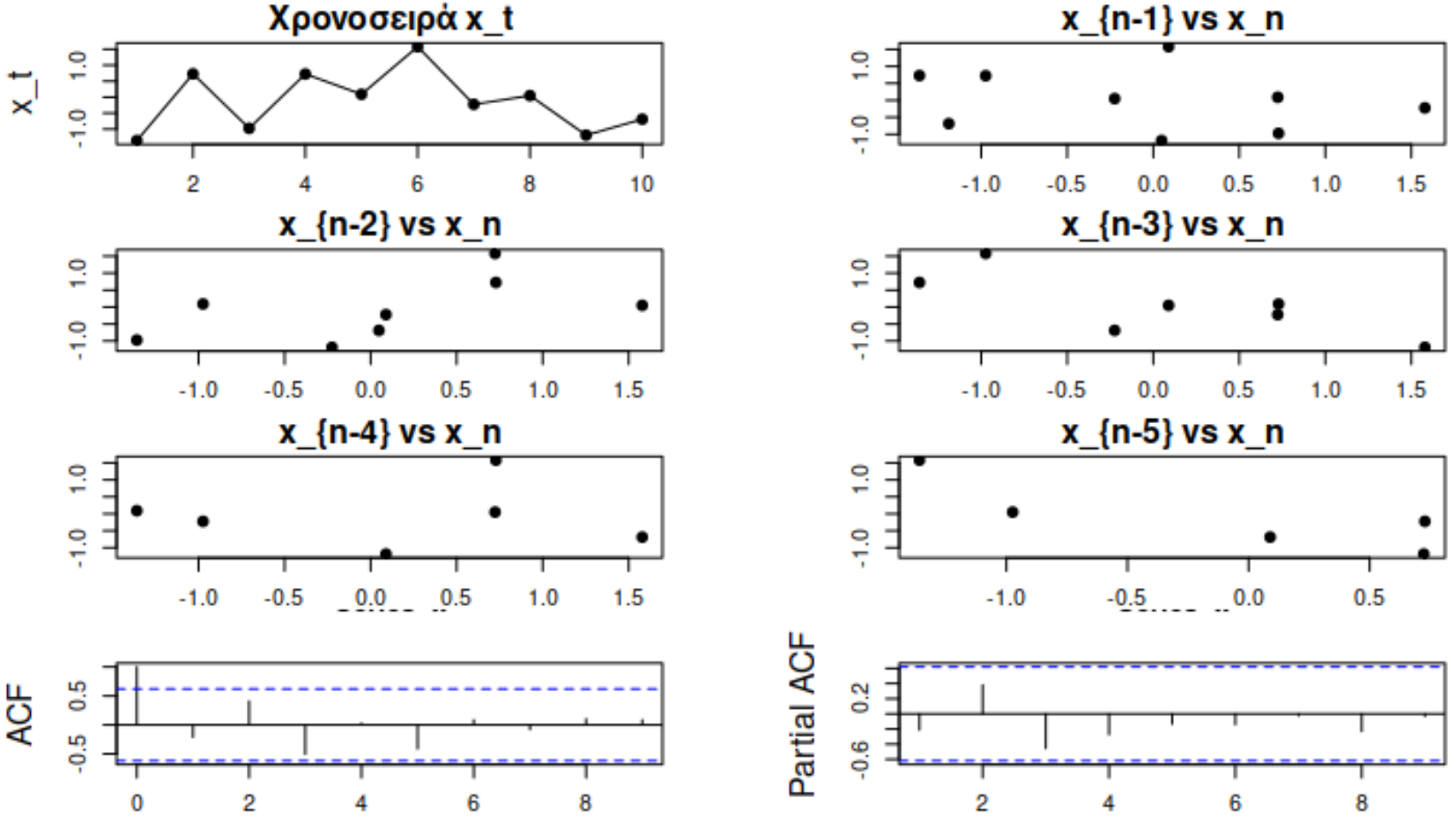
Τυχαία χρονοσειρά (Λευκός Θόρυβος)

Εάν μια χρονοσειρά είναι εντελώς τυχαία, τότε για ικανά μεγάλο δείγμα μεγέθους K, θα είναι $r_h \approx 0$, $h > 0$. Αυτό συμβαίνει γιατί σε μια τυχαία χρονοσειρά (αποδεικνύεται ότι) είναι περίπου $r_h \sim N(0, 1/K)$.

Έτσι, στο ACF αναμένουμε το 95% των τιμών (19 από τις 20 τιμές) να βρίσκονται μεταξύ $-2/K^{½}$ και $+2/K^{½}$.

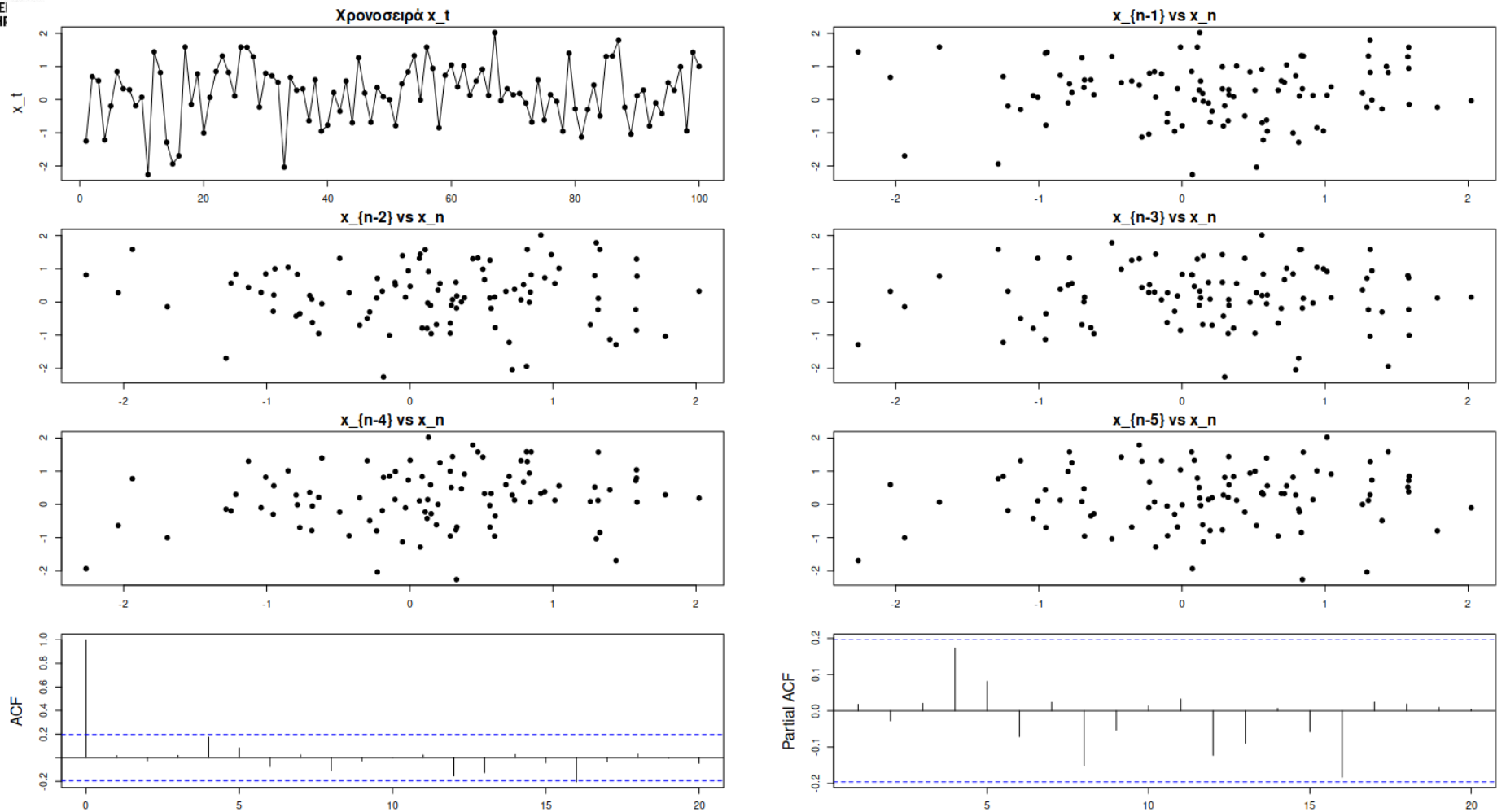
ΔΗΜΟΚΡΙΤΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΡΑΚΗΣ | DEMOCRITUS UNIVERSITY OF THRACE

x = rnorm(10, 0, 1)



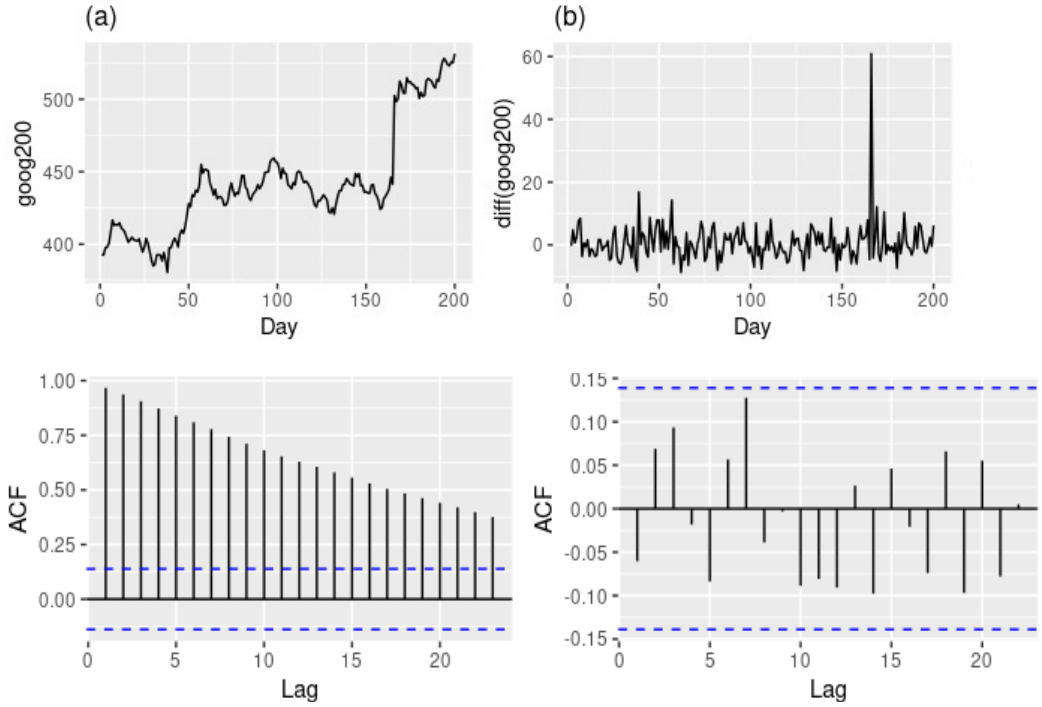
ΔΗΜΟΚΡΙΤΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΡΑΚΗΣ | DEMOCRITUS UNIVE OF THI

x = rnorm(100, 0, 1)



Στατιστικά και Διαγράμματα που Χρησιμοποιούμε στην Ανάλυση Χρονοσειρών

Το διάγραμμα των τιμών αυτοσυσχέτισης (correlogram) αντανακλά τη στασιμότητα ή μη μίας χρονοσειράς. Για μια στάσιμη χρονοσειρά, η τιμή της αυτοσυσχέτισης (ACF) θα μειωθεί σχετικά γρήγορα, ενώ η ACF μίας μη στάσιμης χρονοσειράς μειώνεται με αργό ρυθμό. Επίσης, για μία μη στάσιμη χρονοσειρά, η τιμή του r_1 είναι συχνά μεγάλη και θετική.



Στατιστικά και Διαγράμματα που Χρησιμοποιούμε στην Ανάλυση Χρονοσειρών

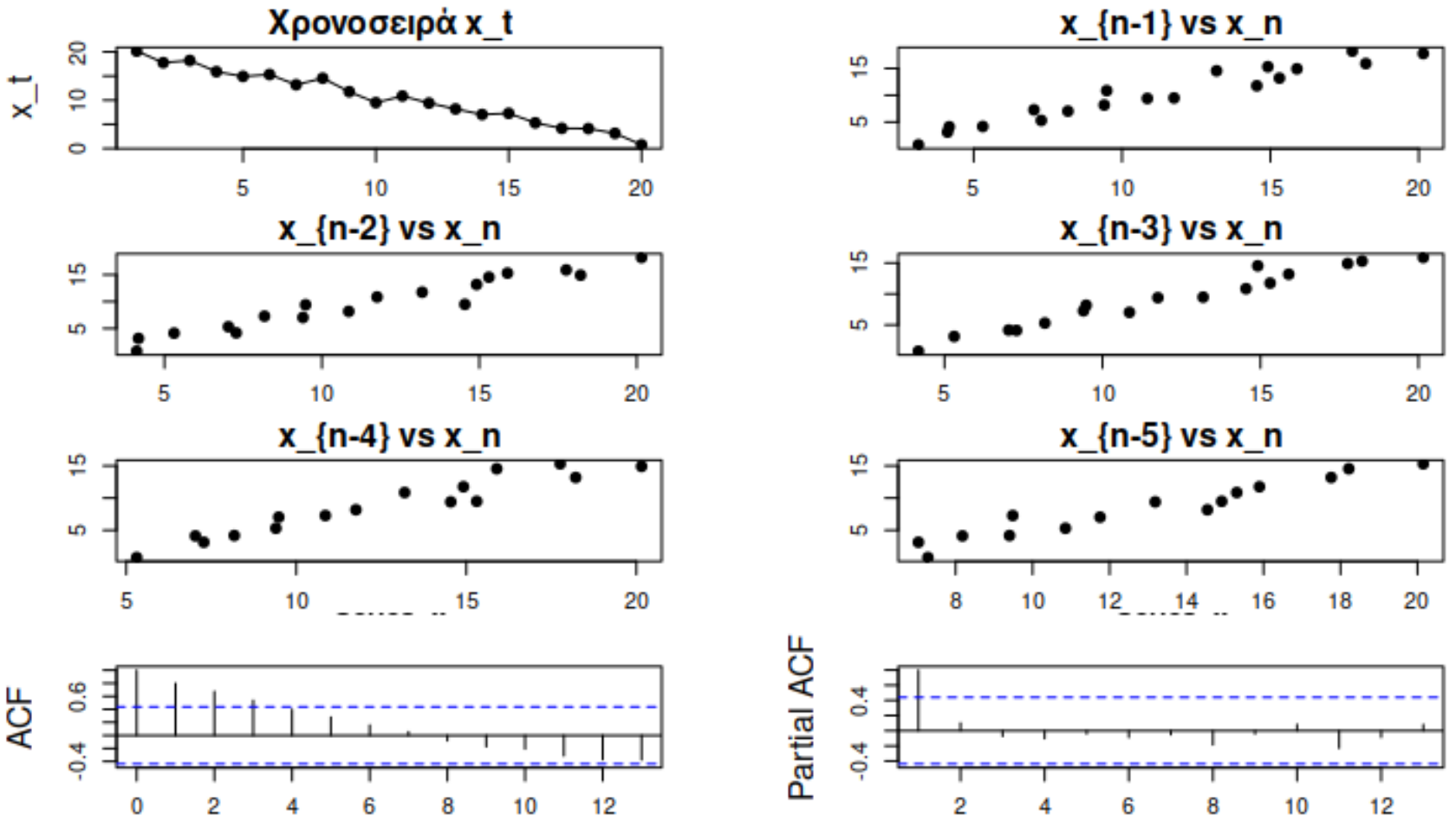
Ερμηνεία του ACF

Σειρά με τάση

Εάν υπάρχει τάση (είτε θετική είτε αρνητική) αναμένουμε θετική αυτοσυσχέτιση.

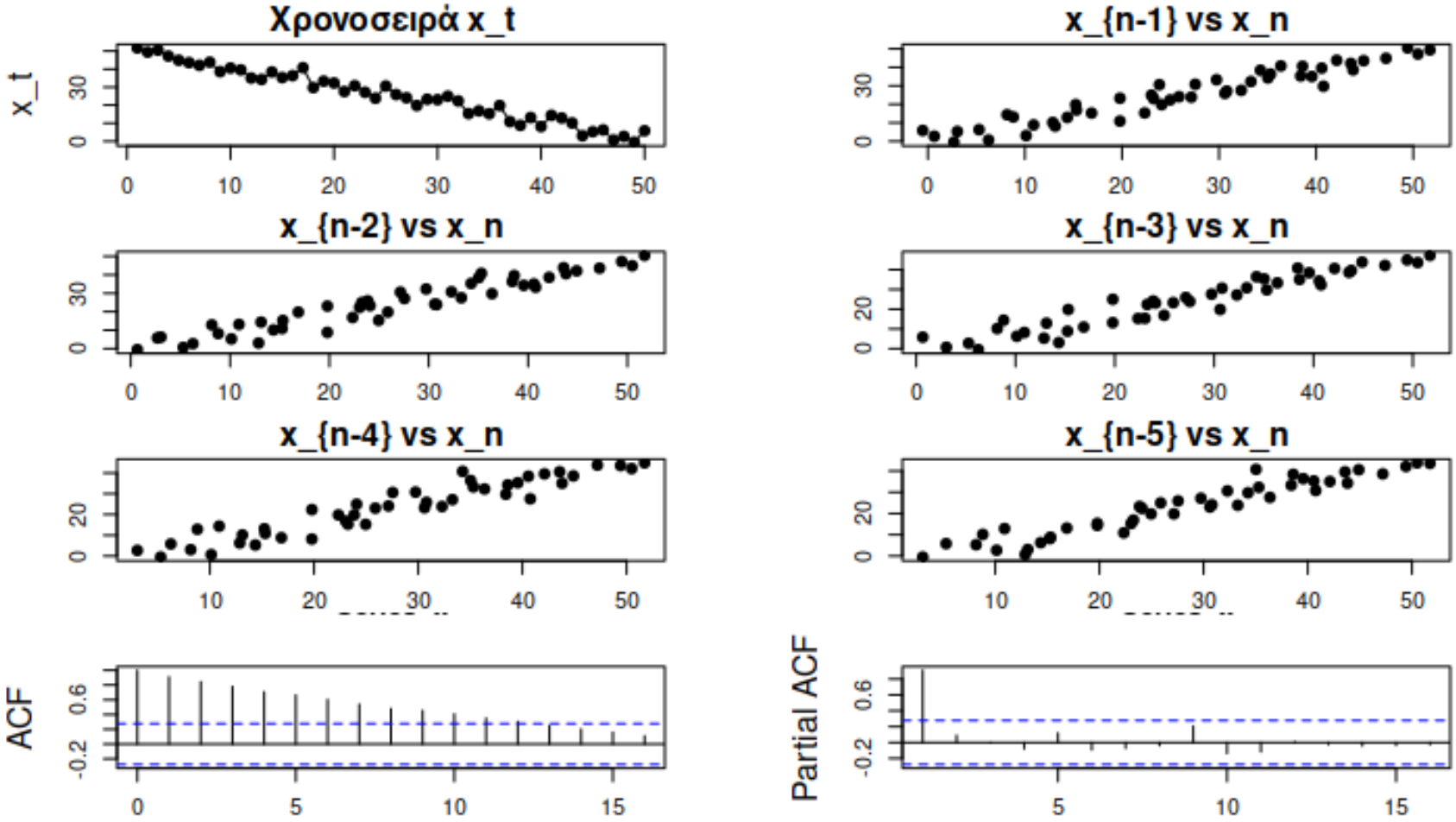
ΔΗΜΟΚΡΙΤΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΡΑΚΗΣ | DEMOCRITUS UNIVERSITY OF THRACE

x = rep(20:1) + rnorm(20)



ΔΗΜΟΚΡΙΤΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΡΑΚΗΣ | DEMOCRITUS UNIVERSITY OF THRACE

x = rep(50:1) + rnorm(50, 0, 3)



ΔΗΜΟΚΡΙΤΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΡΑΚΗΣ | DEMOCRITUS UNIVERSITY OF THRACE

Για τη δημιουργία των διαγραμμάτων χρησιμοποιήθηκε η παρακάτω συνάρτηση της R.

```
my.plot.ts = function(x){
  library(tseries)
  library(Hmisc)
  par(mfrow=c(4,2))
  par(mar = c(2, 4.5, 2, 4.5))
  plot(x, main = "Χρονοσειρά x_t", xlab = "t", ylab = "x_t", cex = 1.5, cex.lab = 1.5, cex.main = 1.5,
pch=20);
  lines(1:length(x), x, pch=20)
  for(l in 1:5){
    plot(na.omit(Lag(x, l)), x[(l+1):length(x)], main = paste0('x_{n-', l,"} vs x_n"), ylab = '', xlab =
paste0('x_{t-', l,"}"), cex = 1.5, cex.main = 1.5,cex.lab = 1.5, pch=20)
  }
  acf(x, pl=T, cex.lab = 1.5, cex.main = 1.5)
  pacf(x, pl=T, cex.lab = 1.5, cex.main = 1.5)
  par(mfrow=c(1,1))
}
```

Στατιστικά και Διαγράμματα που Χρησιμοποιούμε στην Ανάλυση Χρονοσειρών

Ερμηνεία του ACF

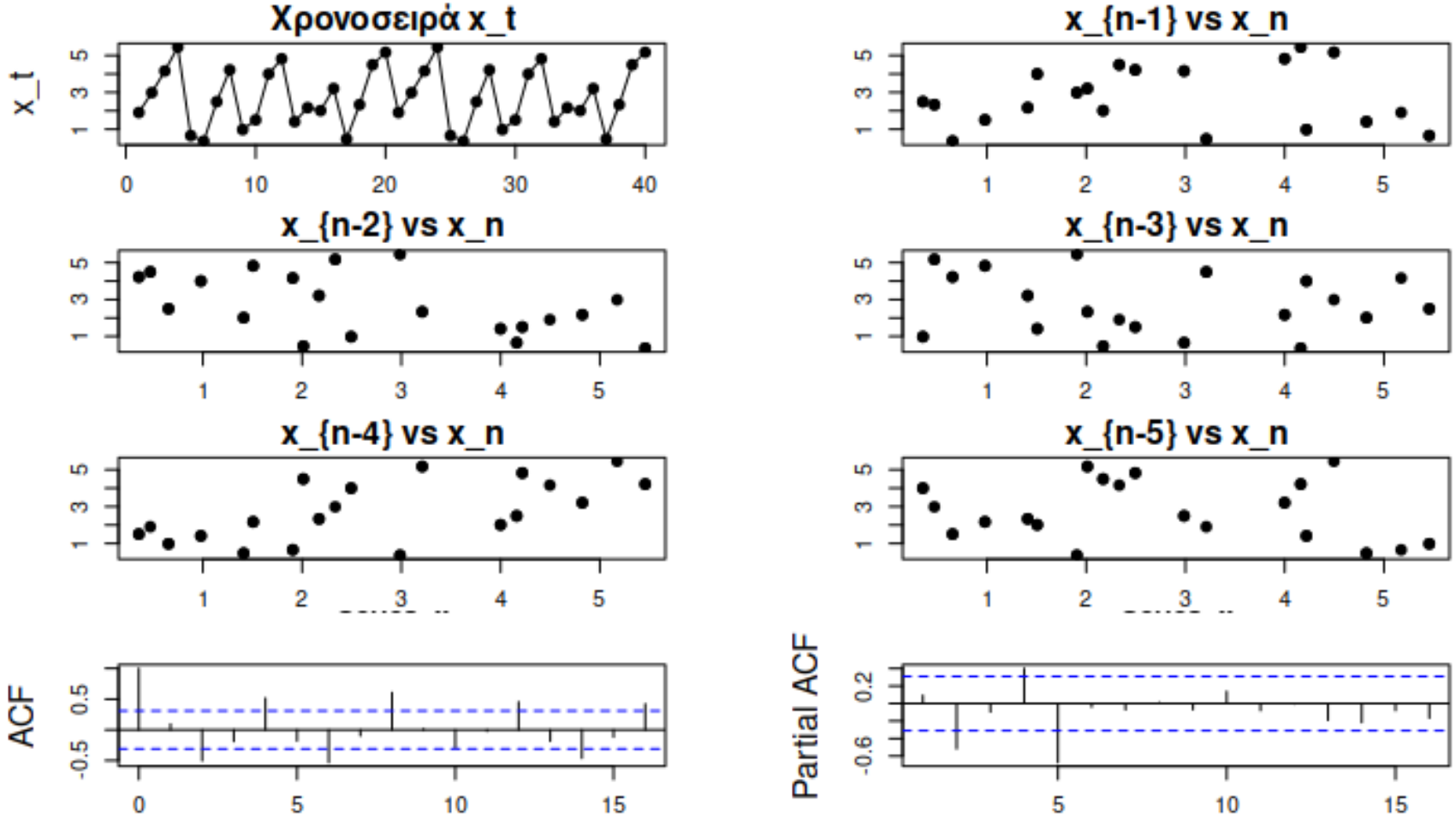
Σειρά με εποχιακή διακύμανση (περιοδικότητα)

Εάν μια χρονοσειρά περιέχει μια εποχιακή διακύμανση, τότε η τιμή της αυτοσυσχέτισης r_h, θα παρουσιάζει επίσης μια ταλάντωση στην ίδια περίοδο της διακύμανσης.

Ειδικότερα, αν η χρονοσειρά $\{x_n\}$ ακολουθεί ένα ημιτονοειδές μοτίβο, τότε το ίδιο θα συμβαίνει για την r_h.

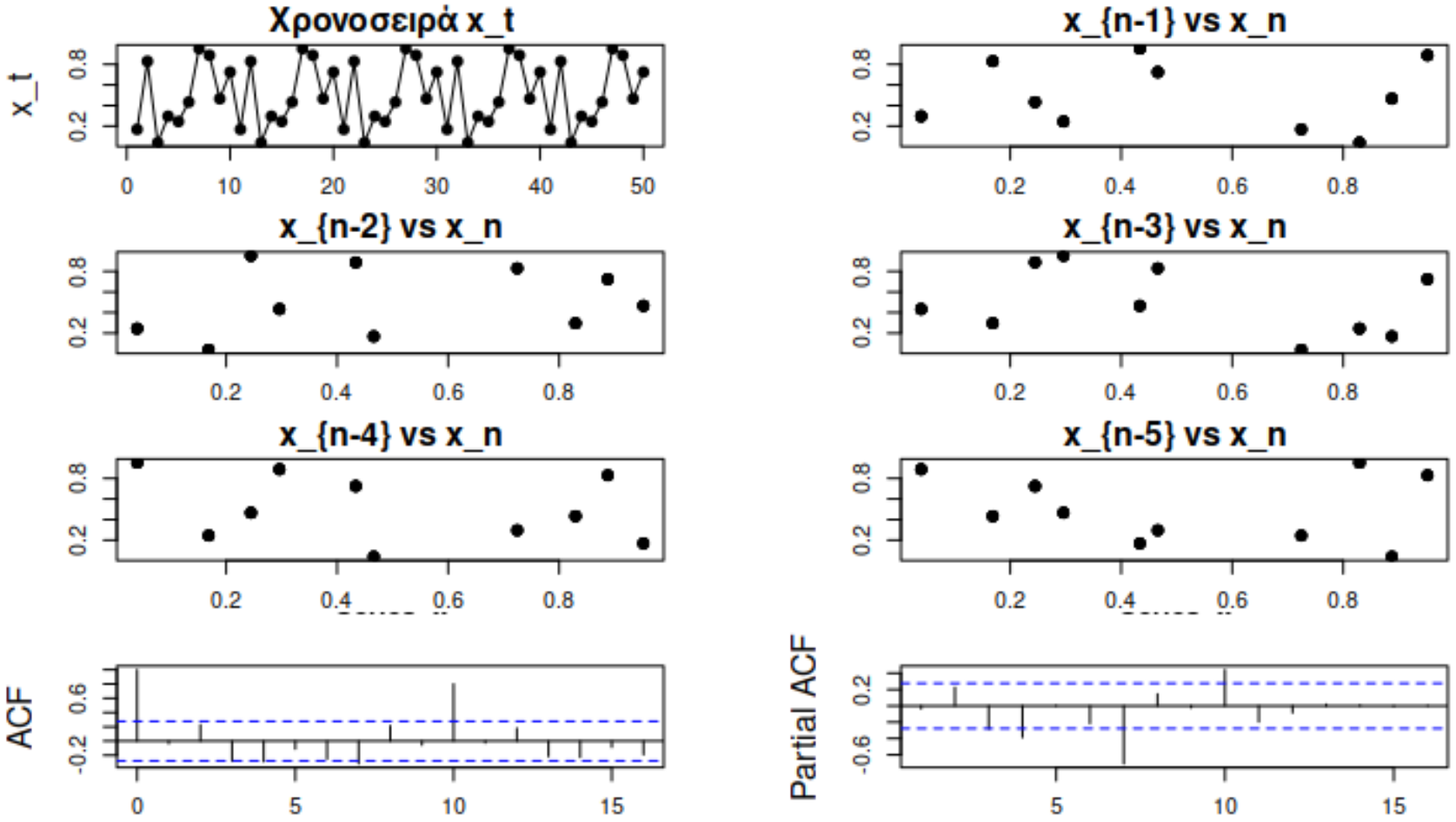
ΔΗΜΟΚΡΙΤΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΡΑΚΗΣ | DEMOCRITUS UNIVERSITY OF THRACE

x = c(rep(1:4, 10)) + rnorm(20)



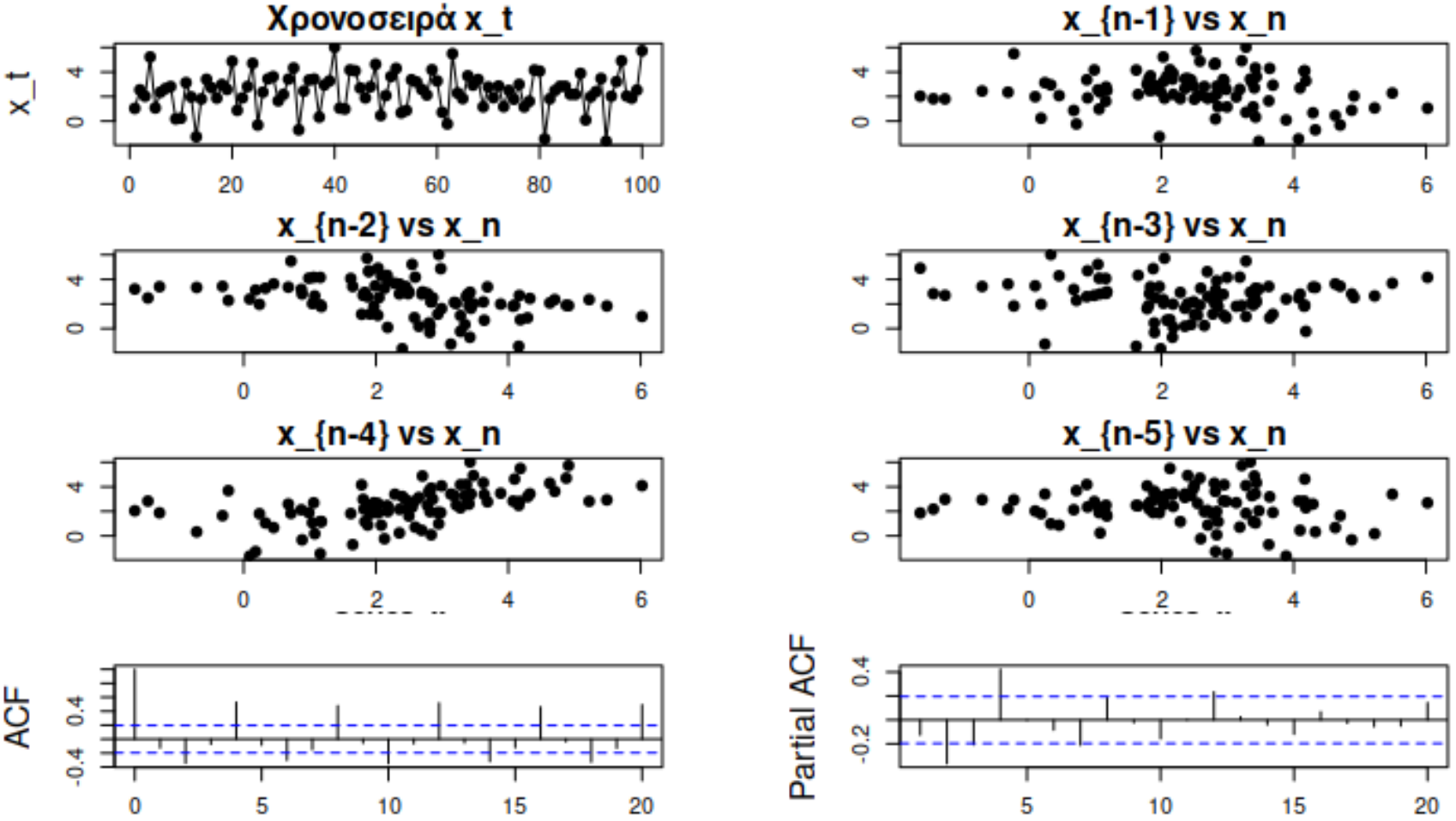
ΔΗΜΟΚΡΙΤΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΡΑΚΗΣ | DEMOCRITUS UNIVERSITY OF THRACE

x = rep(runif(10, 0, 1), 5)



ΔΗΜΟΚΡΙΤΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΡΑΚΗΣ | DEMOCRITUS UNIVERSITY OF THRACE

x = c(rep(1:4, 25)) + rnorm(100)



ΔΗΜΟΚΡΙΤΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΡΑΚΗΣ | DEMOCRITUS UNIVERSITY OF THRACE

Στατιστικά και Διαγράμματα που Χρησιμοποιούμε στην Ανάλυση Χρονοσειρών

Δραστηριότητα

Τι μπορούμε να καταλάβουμε για μία χρονοσειρά με τα εξής στοιχεία;

lag	acf	pacf
1	0.9	0.9
2	0.85	0.4
3	0.7	0.1

Στατιστικά και Διαγράμματα που Χρησιμοποιούμε στην Ανάλυση Χρονοσειρών

Δραστηριότητα

Τι μπορούμε να καταλάβουμε για μία χρονοσειρά με τα εξής στοιχεία;

lag	acf	pacf
1	0.9	0.9
2	0.85	0.4
3	0.7	0.1

Απάντηση

Από τις τιμές του ACF υποδεικνύεται πως κάθε όρος x_t, έχει σημαντική γραμμική σχέση με τους τρεις προηγούμενούς του x_{t-1}, x_{t-2}, x_{t-3}. Όμως, από τις τιμές του PACF, προκύπτει πως η γραμμική σχέση μεταξύ των x_t, x_{t-3}, παύει να είναι σημαντική ύστερα από την αφαίρεση της επιρροής των ενδιάμεσων δύο x_{t-1}, x_{t-2}. Συμπεραίνουμε, πως ένα μοντέλο που θα μπορούσε να προσαρμοστεί στη χρονοσειρά αυτή είναι το εξής: $x_t = \mu + \alpha \cdot x_{t-1} + \beta \cdot x_{t-2}$.