SPSS STATISTICS

PART 2

ΔΕΣΠΟΙΝΑ ΚΟΠΑΝΑΚΗ

ΠΑΛΙΝΔΡΟΜΗΣΗ

- Σε μεγάλο αριθμό προβλημάτων έχουμε πειραματικά δεδομένα της μορφής (x, y).
- Σκοπός: να προσδιορίσουμε την εξίσωση που τα περιγράφει.
- Η διαδικασία εύρεσης της εξίσωσης ονομάζεται **παλινδρόμηση (regression)**.
- Αποτέλεσμα: ένας πίνακας δεδομένων αντικαθίσταται από μια απλή εξίσωση.
- Κριτήριο που ορίζει τον καλύτερο τρόπο περιγραφής δεδομένων: **κριτήριο ελαχίστων τετραγώνων**.
  - Ορίζει ως καλύτερη καμπύλη εκείνη που περνά μέσα από τα σημεία (x_i, y_i) και για την οποία το άθροισμα των τετραγώνων των *υπολοίπων* είναι ελάχιστο.
  - *Υπόλοιπο* ορίζουμε την διαφορά μεταξύ πραγματικής και θεωρητικής τιμής y σε μια ορισμένη τιμή x.

ΠΑΛΙΝΔΡΟΜΗΣΗ - ΠΑΡΑΔΕΙΓΜΑ

- Μεταβολή του μήκους του βραχίονα νηπίων σε mm με το χρόνο σε εβδομάδες
- Να γίνει η γραφική παράσταση και να εκτιμηθεί η ηλικία δυο νηπίων με μήκος βραχίονα 50 και 55 mm

- **Ανεξάρτητη** (ενεργητική, πειραματική, ερέθισμα, επεξηγηματική): Αυτή που ελέγχεται, μεταβάλλεται κατά βούληση του ερευνητή, το αίτιο.
  - Μήκος βραχίονα
- **Εξαρτημένη** (παθητική, αντίδραση, κριτήριο): Αυτή που δέχεται την επίδραση της ανεξάρτητης, το αποτέλεσμα.
  - Ηλικία

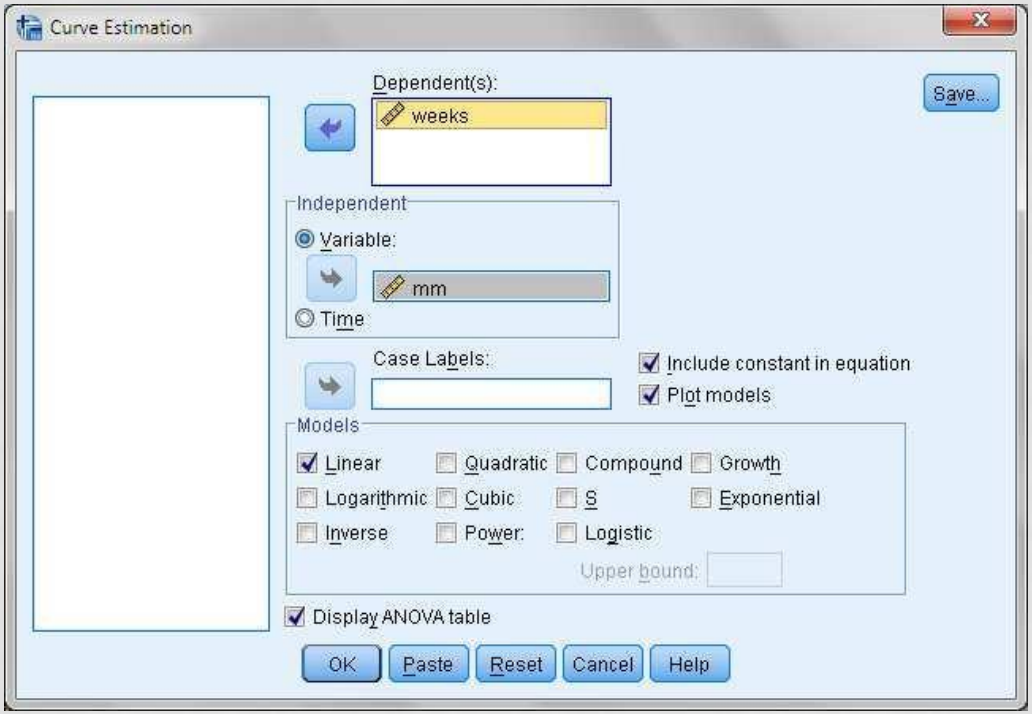
ΠΑΛΙΝΔΡΟΜΗΣΗ - ΠΑΡΑΔΕΙΓΜΑ

- Μεταφορά δεδομένων στο SPSS

mm	weeks	mm	weeks
42	28	65	37
45	27	65	38
58	32	68	4
59	34	7	4
59	35	7	4
61	35	72	41
64	36	75	45

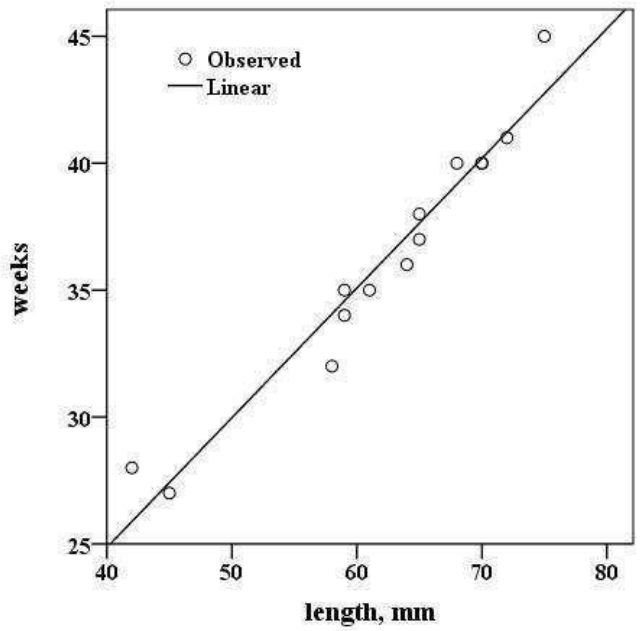
ΠΑΛΙΝΔΡΟΜΗΣΗ - ΠΑΡΑΔΕΙΓΜΑ

- Analyze > Regression > Curve Estimation



- Include constant in equation: το επιλέγουμε πάντα εκτός κι αν έχουμε στοιχεία ότι όταν x=0 τότε και y=0.

ΠΑΛΙΝΔΡΟΜΗΣΗ - ΠΑΡΑΔΕΙΓΜΑ



Coefficients

	Unstandardized Coefficients		Standardized Coefficients		
	B	Std. Error	Beta	t	Sig.
mm	,511	,035	,973	14,541	,000
(Constant)	4,420	2,215		1,996	,069

- y = a + b x
- y = 4,42 + 0,511x
- a=4,42 ± 2,215
- b=0,511± 0,035

- Η σταθερά α είναι στατιστικά μη σημαντική **(Sig. > 0,05)** άρα μπορεί να παραληφθεί από τη μελέτη

- 4.42 + 0,511*50=29,97
  - (περίπου 30 εβδομάδες)
- 4.42 + 0,511*55=32,525
  - (περίπου 32,5 εβδομάδες)

ΠΑΛΙΝΔΡΟΜΗΣΗ – 2ο ΠΑΡΑΔΕΙΓΜΑ

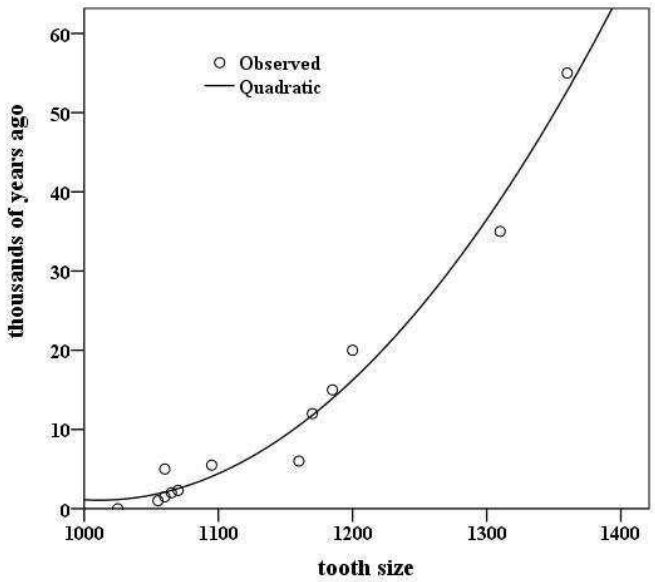
- Μεταβολή διαστάσεων των δοντιών με το πέρασμα των χιλιετιών
- Να γίνει η γραφική παράσταση των τιμών του πίνακα και να εκτιμηθεί η χρονολογία των δειγμάτων 1150 και 1250mm^2

Thousands years ago	Tooth-size (mm^2)	Thousands years ago	Tooth-size (mm^2)
0	1025	6	1160
1	1055	12	1170
1.5	1060	15	1185
2	1065	20	1200
2.3	1070	35	1310
5	1060	55	1360
5.5	1095		

- Ανεξάρτητη μεταβλητή: tooth-size
- Εξαρτημένη μεταβλητή: years

ΠΑΛΙΝΔΡΟΜΗΣΗ – 2ο ΠΑΡΑΔΕΙΓΜΑ

- Analyze > Regression > Curve Estimation
- Models > Quadratic



Coefficients

	Unstandardized Coefficients		Standardized Coefficients		
	B	Std. Error	Beta	t	Sig.
tooth size	-,859	,184	-5,520	-4,677	,001
tooth size ** 2	,000	,000	6,479	5,490	,000
(Constant)	435,137	108,157		4,023	,002

- $y = a + bx + cx^2$
- a=435,137 ± 108,157
  - b=-0,859± 0,184
    - c=0,0004246

- $435{,}137 - 0{,}858616 * 1150 + 0{,}0004246 * 1150^2$
  - 9,26 εβδομάδες
- $435{,}137 - 0{,}858616 * 1250 + 0{,}0004246 * 1250^2$
  - 25,3 εβδομάδες

ΠΑΛΙΝΔΡΟΜΗΣΗ – 3ο ΠΑΡΑΔΕΙΓΜΑ

- Η ποσότητα y του νερού που εξατμίζεται από το έδαφος εξαρτάται από
  - τη μέγιστη (T_1) θερμοκρασία του εδάφους
  - την ελάχιστη (T_2) θερμοκρασία του εδάφους
  - τη μέγιστη (T_3) θερμοκρασία του αέρα
  - την ελάχιστη (T_4) θερμοκρασία του αέρα
- Να προσδιοριστεί το γραμμικό μοντέλο, δηλαδή η συνάρτηση
- $y = a_0 + a_1T_1 + a_2T_2 + a_3T_3 + a_4T_4$

y	T_1	T_2	T_3	T_4
30	28	18	29	15
34	28	18	30	16
33	26	18	28	17
26	27	19	28	18
41	28	20	31	20
10	23	18	25	19
12	22	18	25	20
20	23	19	28	20
31	28	20	31	21
38	30	22	32	24
43	31	22	32	24
47	32	23	34	24
45	31	22	34	23
45	31	22	33	21
22	27	20	30	20
5	15	20	28	20
30	28	15	30	18
29	28	21	30	20
23	25	21	31	21

ΠΑΛΙΝΔΡΟΜΗΣΗ – 3ο ΠΑΡΑΔΕΙΓΜΑ

- Analyze > Regression > Linear
- Ανεξάρτητες μεταβλητές: T_1, T_2, T_3, T_4
- Εξαρτημένη μεταβλητή: y
- Method: επιλέγουμε τη μέθοδο που θα χρησιμοποιηθεί για τον υπολογισμό των σταθερών της συνάρτησης
  - Enter το πρόγραμμα υπολογίζει όλες τις σταθερές
  - Backward: το πρόγραμμα αρχικά υπολογίζει όλες τις σταθερές και μετά αρχίζει να αφαιρεί μία-μία τις στατιστικά μη σημαντικές
  - *Forward:* εισάγει αρχικά τον σταθερό όρο και μετά τη σταθερά που αντιστοιχεί στη μεταβλητή που έχει τη μεγαλύτερη συσχέτιση με την εξαρτημένη μεταβλητή. Εξετάζεται αν είναι στατιστικά σημαντική και μετά εισάγει την επόμενη μεταβλητή με την καλύτερη συσχέτιση με την εξαρτημένη μεταβλητή κ.ο.κ
  - *Stepwise:* συνδυασμός *Backward* και *Forward*

ΠΑΛΙΝΔΡΟΜΗΣΗ – 3ο ΠΑΡΑΔΕΙΓΜΑ

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-75.412	12.218		-6.172	.000
	T1	1.882	.386	.621	4.876	.000
	T2	.212	.959	.035	.221	.829
	T3	1.990	.774	.418	2.572	.022
	T4	-.465	.659	-.098	-.705	.492

a. Dependent Variable: x

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-75.412	12.218		-6.172	.000
	T1	1.882	.386	.621	4.876	.000
	T2	.212	.959	.035	.221	.829
	T3	1.990	.774	.418	2.572	.022
	T4	-.465	.659	-.098	-.705	.492
2	(Constant)	-75.037	11.710		-6.408	.000
	T1	1.870	.370	.617	5.058	.000
	T3	2.063	.677	.433	3.047	.008
	T4	-.367	.473	-.077	-.776	.450
3	(Constant)	-75.494	11.549		-6.537	.000
	T1	1.933	.356	.638	5.425	.000
	T3	1.776	.560	.373	3.172	.006

a. Dependent Variable: x

ΣΥΣΧΕΤΙΣΗ ΜΕΤΑΒΛΗΤΩΝ

- Σχετίζεται έμμεσα με την παλινδρόμηση και τα ελάχιστα τετράγωνα είναι το πρόβλημα της συσχέτισης (correlation) δύο μεταβλητών
- Χρήσιμο να γνωρίζουμε αν δύο τυχαίες μεταβλητές σχετίζονται ή όχι, αν δηλαδή η μεταβολή της μιας μεταβάλει και την άλλη

- **Συντελεστή Pearson r [-1, 1]**
- Αρνητικές τιμές του r σημαίνουν ότι όταν η μεταβλητή x αυξάνει, η y ελαττώνεται και το αντίστροφο
- r = 0 σημαίνει παντελή έλλειψη **συσχέτισης**
- r θετικό σημαίνει ότι όταν η μια μεταβλητή αυξάνει, αυξάνει και η άλλη
- ***Χρησιμοποιείται μόνο όταν τα δεδομένα ακολουθούν την κανονική κατανομή***

ΣΥΣΧΕΤΙΣΗ ΜΕΤΑΒΛΗΤΩΝ

- Αν δεν ακολουθούν την κανονική κατανομή, υπολογίζουμε τον **συντελεστή Spearman, ρ [-1, 1]**
- Μη παραμετρική μέθοδος
- Ομοίως, αρνητικές τιμές του r σημαίνουν ότι όταν η μεταβλητή x αυξάνει, η y ελαττώνεται και το αντίστροφο

ΣΥΣΧΕΤΙΣΗ ΜΕΤΑΒΛΗΤΩΝ-ΠΑΡΑΔΕΙΓΜΑ

- Έστω ότι έχουμε τη διάθεσή μας τις μεταβλητές height και body mass.
- Να εξετασθεί αν υπάρχει συσχέτιση μεταξύ των μεταβλητών
- 1ο βήμα: έλεγχος κάθε μεταβλητής για να διαπιστώσουμε αν ακολουθούν την κανονική μεταβλητή
  - *Analyze > Descriptive Statistics > Explore*
- Έστω ότι και οι δυο ακολουθούν κανονική κατανομή
- Μπορούμε να χρησιμοποιήσουμε τον συντελεστή Pearson και προφανώς τον συντελεστή Spearman που δεν υπόκειται σε περιορισμούς

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
height	,098	48	,200*	,963	48	,134
body mass	,116	48	,111	,965	48	,167

a. Lilliefors Significance Correction

*. This is a lower bound of the true significance.

ΣΥΣΧΕΤΙΣΗ ΜΕΤΑΒΛΗΤΩΝ-ΠΑΡΑΔΕΙΓΜΑ

- *Analyze* > *Correlate* > *Bivariate*
- Μεταφέρουμε τις μεταβλητές height και body mass στο πλαίσιο *Variables*
- επιλέγουμε στο *Correlation Coefficients* τα κριτήρια *Pearson* και Spearman

Correlations

		height	body mass
height	Pearson Correlation	1	,863**
	Sig. (2-tailed)		,000
	N	50	48
body mass	Pearson Correlation	,863**	1
	Sig. (2-tailed)	,000	
	N	48	48

**. Correlation is significant at the 0.01 level (2-tailed).

- Υψηλή συσχέτιση των μεταβλητών
  - (r = 0,863 και ρ = 0,878)

Correlations

			height	body mass
Spearman's rho	height	Correlation Coefficient	1,000	,878**
		Sig. (2-tailed)	.	,000
		N	50	48
	body mass	Correlation Coefficient	,878**	1,000
		Sig. (2-tailed)	,000	.
		N	48	48

**. Correlation is significant at the 0.01 level (2-tailed).

ΑΝΑΛΥΣΗ ΠΟΛΛΩΝ ΜΕΤΑΒΛΗΤΩΝ (MULTIVARIATE ANALYSIS)

- Συσσωρεύουμε πληθώρα δεδομένων και θέλουμε να ερευνήσουμε αν υπάρχουν ομάδες δειγμάτων με παρόμοιες ιδιότητες, και ποιες είναι αυτές
  - Υπάρχει στατιστικά σημαντική διαφορά ανάμεσα στα κεραμικά αγγεία από τις θέσεις Παλιάμπελα και Μακρύγιαλος?
    - χρησιμοποιώντας ως μεταβλητές συγχρόνως το ύψος των αγγείων, το πλάτος, τη διάμετρο του στομίου, τη διάμετρο της βάσης και άλλες δια
- Ανάλυση Κυρίων Συνιστωσών (Principal Component Analysis - PCA),
- Ανάλυση σε Ομάδες (Cluster Analysis-CA)

ΑΝΑΛΥΣΗ ΚΥΡΙΩΝ ΣΥΝΙΣΤΩΣΩΝ (PRINCIPAL COMPONENT ANALYSIS)

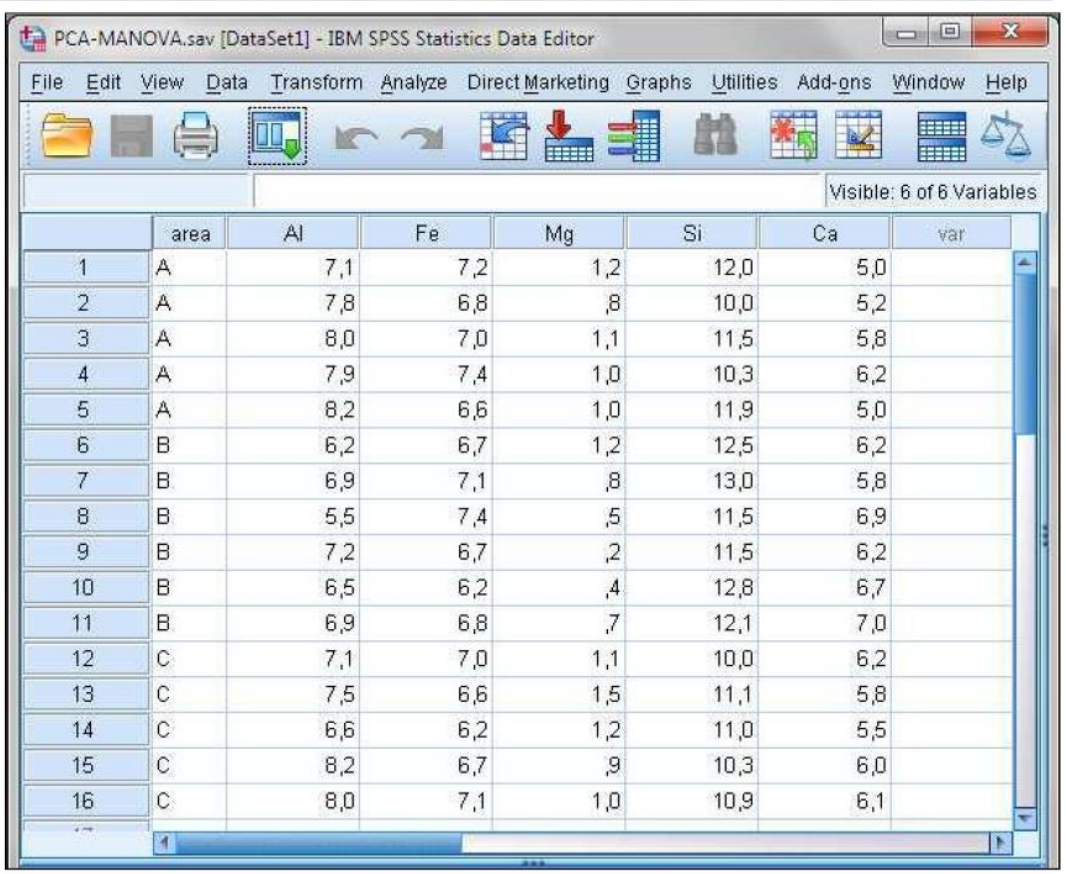
- Η μέθοδος «εξαγωγής» των παραγόντων
  - Αρχικά «εξάγεται» ο πρώτος παράγοντας ή συνιστώσα, ο οποίος ερμηνεύει το μεγαλύτερο δυνατό ποσοστό της διακύμανσης ανάμεσα στα στοιχεία (items) και τον παράγοντα (συσχέτιση)
  - Στη συνέχεια «εξάγεται» ο επόμενος παράγοντας ή συνιστώσα, ο οποίος ερμηνεύει το μεγαλύτερο δυνατό ποσοστό της διακύμανσης που έχει απομείνει από την ερμηνεία του πρώτου παράγοντα.
  - Στη συνέχεια «εξάγεται» ο επόμενος παράγοντας ή συνιστώσα μέχρι να μην μείνει ποσοστό διακύμανσης που δεν ερμηνεύεται από τα στοιχεία που μελετάμε

ΑΝΑΛΥΣΗ ΚΥΡΙΩΝ ΣΥΝΙΣΤΩΣΩΝ (PRINCIPAL COMPONENT ANALYSIS)

- Ελαττώνουμε τις διαστάσεις του πίνακα για να εξετάσουμε αν σ' έναν πίνακα δεδομένων υπάρχουν ομάδες ομοειδών δεδομένων
- Φέρνουμε έναν άξονα (μια γραμμή) μέσα από τα σημεία της γραφικής παράστασης και κατά μήκος της μεγαλύτερης διασποράς των σημείων και προβάλουμε τα σημεία αυτά πάνω στον άξονα.
- Ο άξονας ονομάζεται PC1 ή πρώτη κύρια συνιστώσα.
- Ακολούθως φέρνουμε ένα δεύτερο άξονα, τον PC2, που είναι κάθετος στον PC1 και τον περιστρέφουμε, πάντα κάθετα στον PC1, έτσι ώστε και αυτός να είναι κατά μήκος της μεγαλύτερης διασποράς των σημείων ως προς τη διεύθυνσή του.
- Οι δύο αυτοί άξονες ορίζουν ένα επίπεδο.
- Στο επίπεδο αυτό προβάλουμε όλα τα σημεία.
- Συνεχίζουμε με τον ίδιο τρόπο μέχρι να καταλήξουμε με αρκετά PCs ώστε να εξηγηθεί όλη η διασπορά του δείγματος

PCA - ΠΑΡΑΔΕΙΓΜΑ

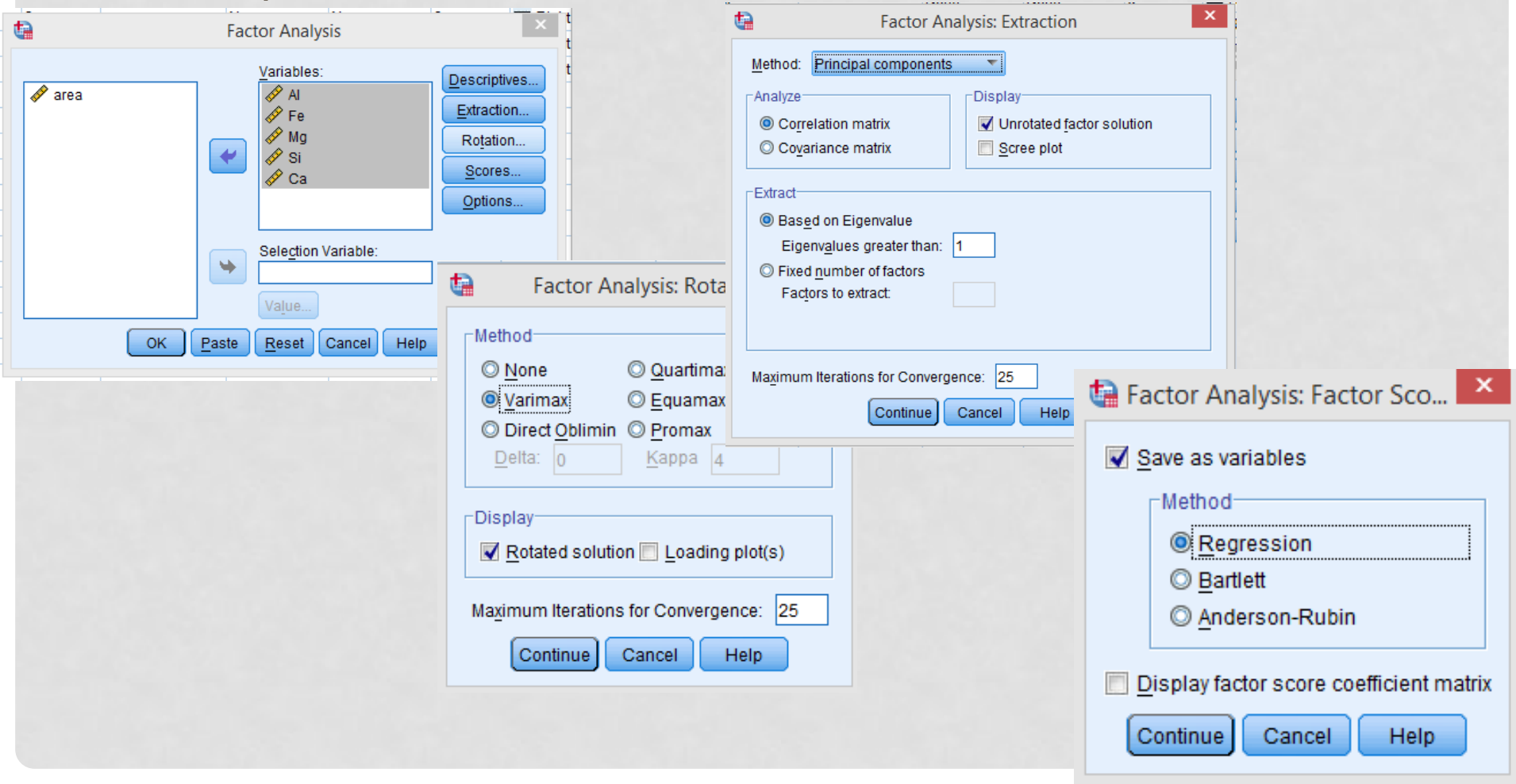
- Αποτελέσματα της χημικής ανάλυσης ειδωλίων ίδιας χρονολογίας που βρέθηκαν σε τρεις διαφορετικές περιοχές A, B, C
- Να εξαχθούν συμπεράσματα σχετικά με την προέλευση των ειδωλίων



	area	Al	Fe	Mg	Si	Ca	var
1	A	7,1	7,2	1,2	12,0	5,0	
2	A	7,8	6,8	,8	10,0	5,2	
3	A	8,0	7,0	1,1	11,5	5,8	
4	A	7,9	7,4	1,0	10,3	6,2	
5	A	8,2	6,6	1,0	11,9	5,0	
6	B	6,2	6,7	1,2	12,5	6,2	
7	B	6,9	7,1	,8	13,0	5,8	
8	B	5,5	7,4	,5	11,5	6,9	
9	B	7,2	6,7	,2	11,5	6,2	
10	B	6,5	6,2	,4	12,8	6,7	
11	B	6,9	6,8	,7	12,1	7,0	
12	C	7,1	7,0	1,1	10,0	6,2	
13	C	7,5	6,6	1,5	11,1	5,8	
14	C	6,6	6,2	1,2	11,0	5,5	
15	C	8,2	6,7	,9	10,3	6,0	
16	C	8,0	7,1	1,0	10,9	6,1	

PCA - ΠΑΡΑΔΕΙΓΜΑ

- *Analyze > Dimension Reduction > Factor*

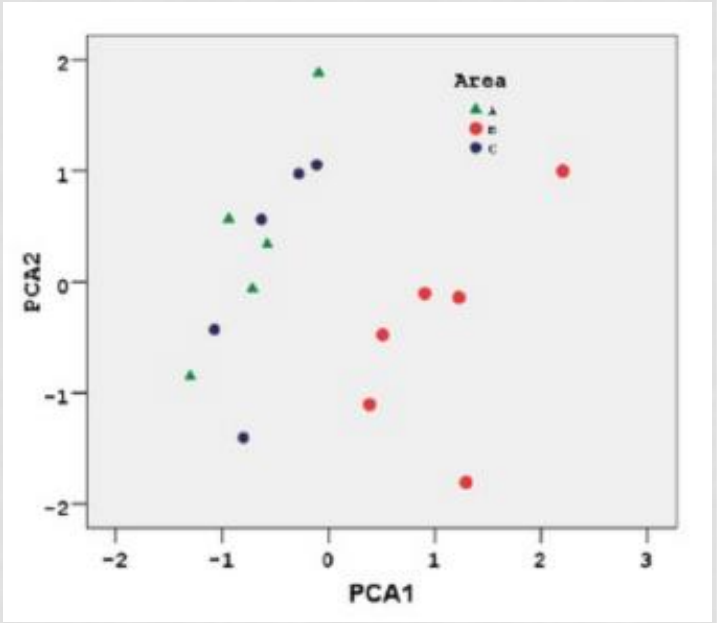


PCA - ΠΑΡΑΔΕΙΓΜΑ

- Με αυτή την επιλογή οι τιμές των PC1, PC2 αποθηκεύονται στο φύλλο εργασίας με τίτλους FAC1_1, FAC2_1.
- Με κλικ στο ***OK*** δημιουργούνται αυτόματα στον SPSS Data Editor οι στήλες FAC1_1, FAC2_1 οι οποίες περιλαμβάνουν τις τιμές των αξόνων PC1 και PC2

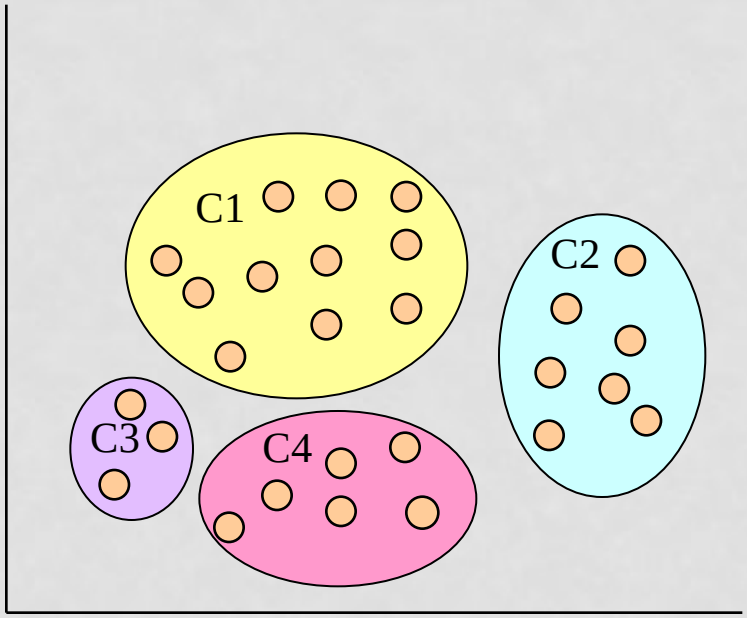
PCA - ΠΑΡΑΔΕΙΓΜΑ

- *Graphs > Legacy Dialogs > Scatter/Dot > Simple Scatter*
- *X Axis:* τη μεταβλητή REGR factor score 1
- *Y Axis:* τη μεταβλητή REGR factor score 2
- *Set Markers by:* τη μεταβλητή Area
- Η κάθε περιοχή, A, B, C, θα έχει διαφορετικό σύμβολο
- τα σημεία της περιοχής B σχηματίζουν μια ξεχωριστή ομάδα
- τα σημεία των περιοχών A και C μαζί μια άλλη ομάδα
- Αν στο διάγραμμα αποτελεσμάτων δεν ξεχωρίσουν ομάδες δοκιμάζουμε διαφορετικές μεθόδους περιστροφής



ΑΝΑΛΥΣΗ ΣΕ ΟΜΑΔΕΣ – CLUSTER ANALYSIS (CA)

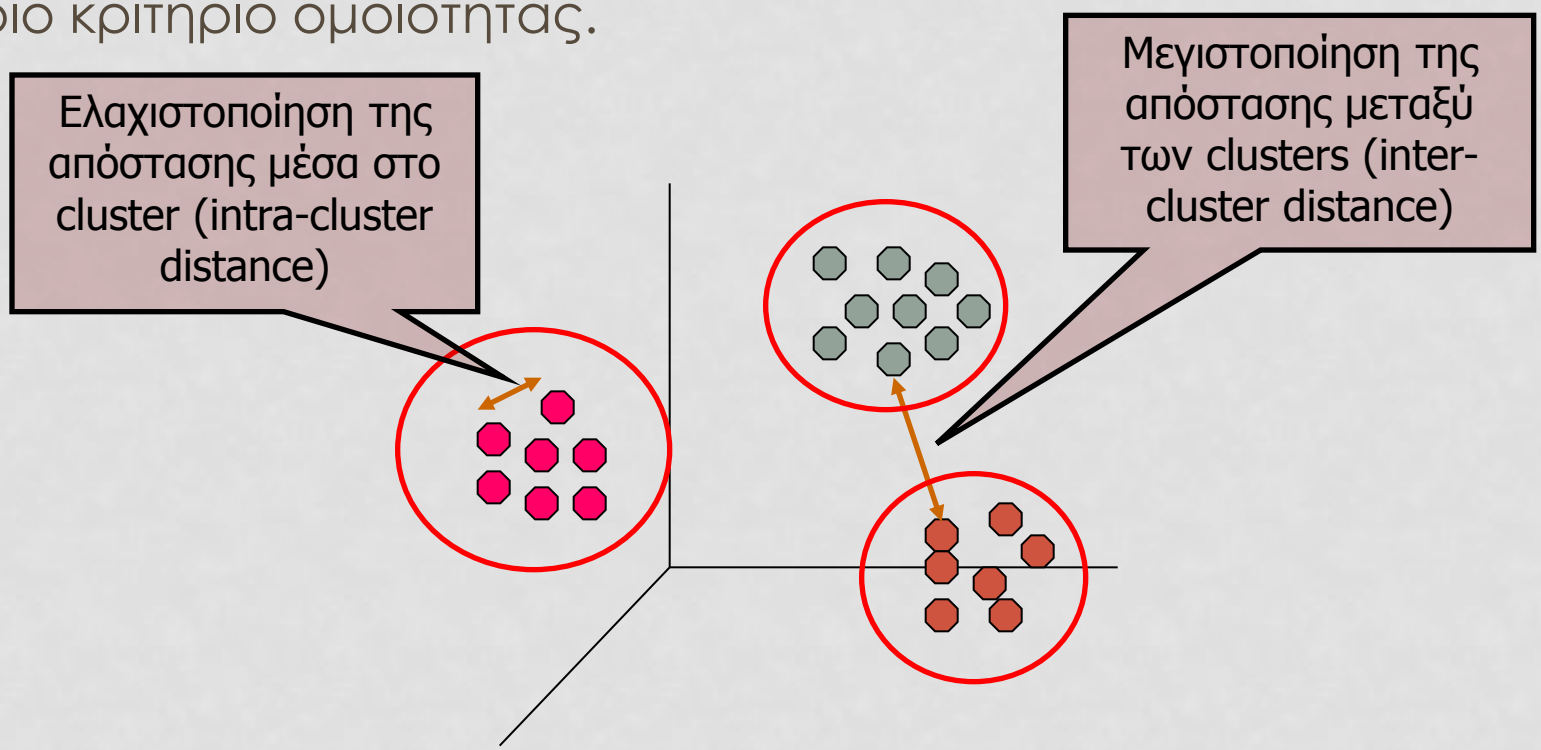
- ή αλλιώς, Ομαδοποίηση
- Εύρεση μιας "φυσικής" ομαδοποίησης των δεδομένων, χωρίς προκαθορισμό των ομάδων



- Πόσες / ποιες συστάδες;

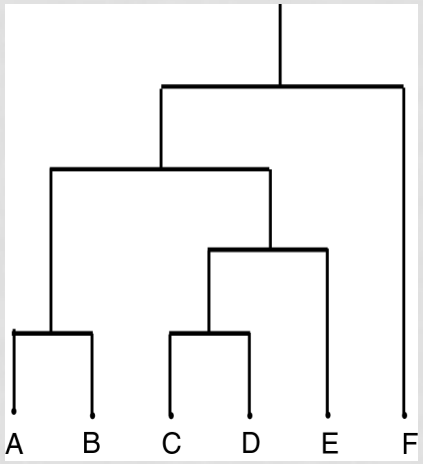
ΣΥΣΤΑΔΟΠΟΙΗΣΗ (CLUSTERING)

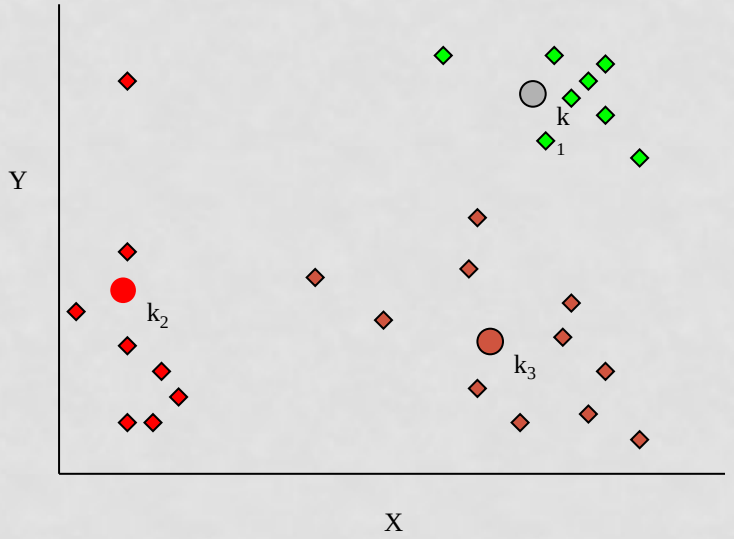
- **Διαμέριση** μίας ΒΔ πελατών σε ομάδες (συστάδες) με βάση κάποιο κριτήριο ομοιότητας.



ΤΥΠΟΙ ΣΥΣΤΑΔΟΠΟΙΗΣΗΣ

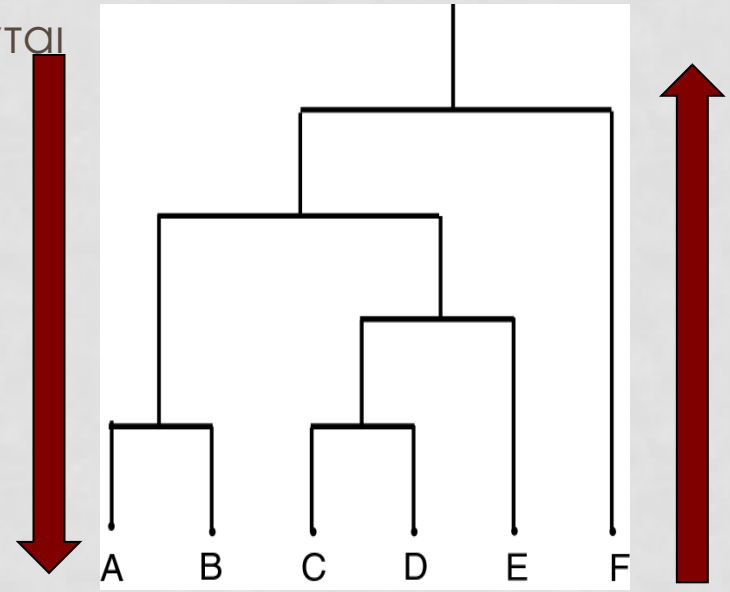
- **Ιεραρχική** vs. **Διαμέρισης**
  - δημιουργούνται εμφωλιασμένα σύνολα συστάδων
    - από 1 σε 2, 3, … N συστάδες (διαιρετικοί αλγόριθμοι) ή
    - από N σε N-1, ..., 2, 1 συστάδες (συσσωρευτικοί αλγόριθμοι)
  - ή δημιουργείται απευθείας ένα σύνολο k συστάδων
    - ο αριθμός k μπορεί να είναι είσοδος στον αλγόριθμο (ή όχι)
- Για **μικρές** (που χωράνε στην κύρια μνήμη) ή **μεγάλες ΒΔ**
- Για ειδικούς τύπους δεδομένων (π.χ. **κατηγορικά**)





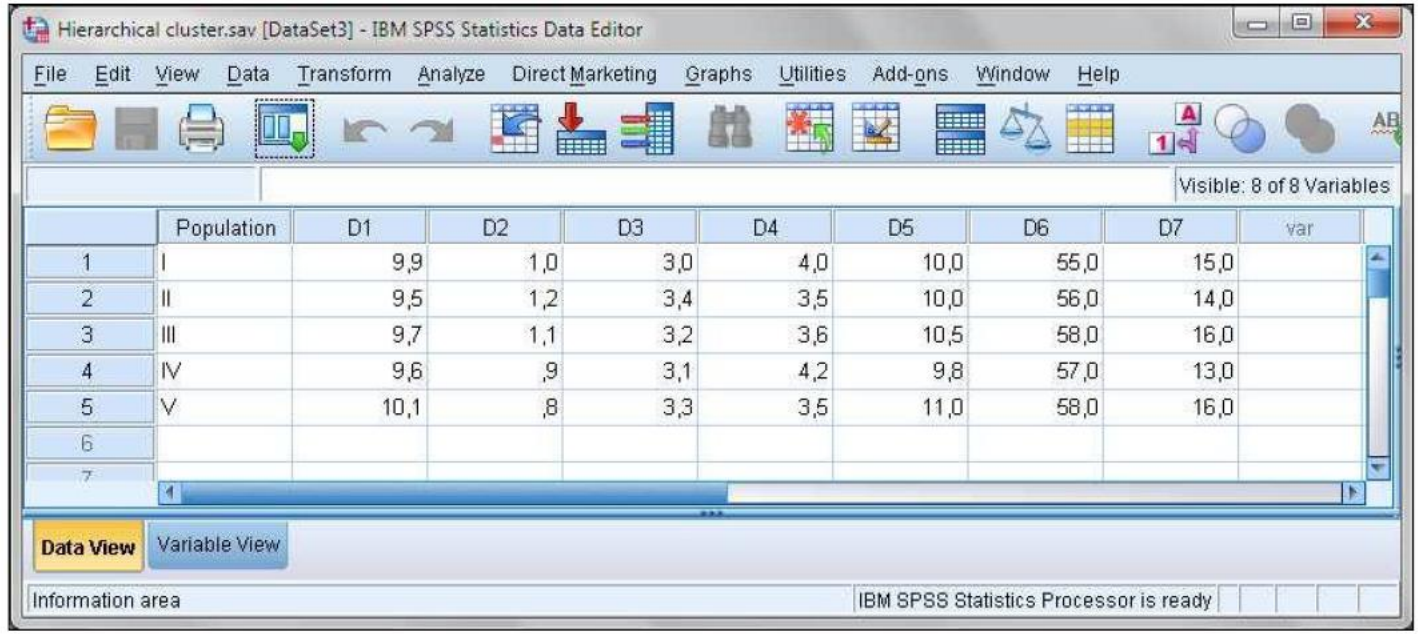
ΙΕΡΑΡΧΙΚΗ ΣΥΣΤΑΔΟΠΟΙΗΣΗ

- Οι συστάδες δημιουργούνται σε επίπεδα
  - Κάθε επίπεδο αντιπροσωπεύει ένα σύνολο από συστάδες
  - Το αποτέλεσμα απεικονίζεται σε ένα δενδρόγραμμα
- **Συσσωρευτικοί αλγόριθμοι** (agglomerative)
  - Αρχικά κάθε στοιχείο είναι μία συστάδα
  - Επαναληπτικά οι συστάδες συγχωνεύονται
  - Προσέγγιση bottom-up
- **Διαιρετικοί αλγόριθμοι** (divisive)
  - Αρχικά όλα τα στοιχεία σε μία συστάδα
  - Επαναληπτικά οι συστάδες διαιρούνται
  - Προσέγγιση top-down



CLUSTER ANALYSIS (CA)-ΠΑΡΑΔΕΙΓΜΑ

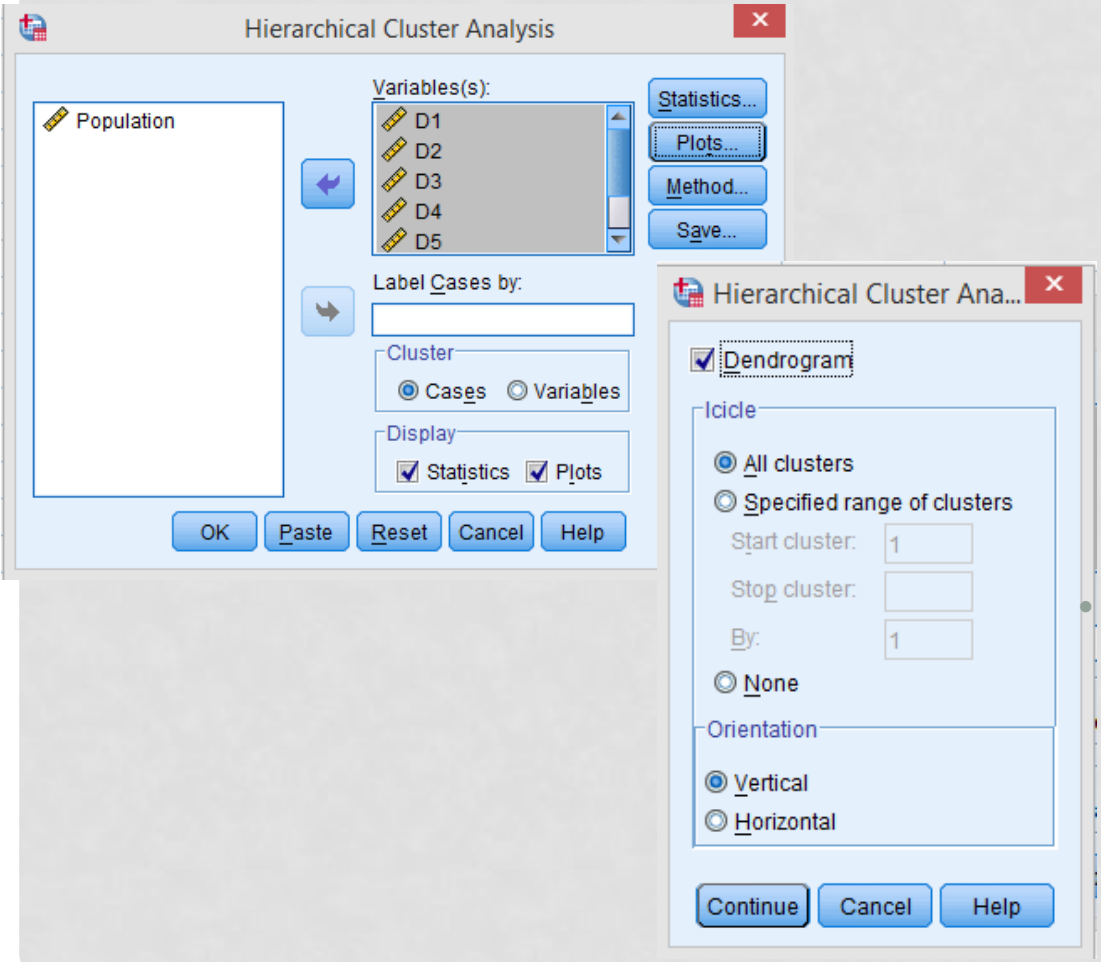
- Τιμές 7 δεικτών που χαρακτηρίζουν το σχήμα των κρανίων πέντε προϊστορικών πληθυσμών
- Να ελεγχθούν πιθανές συγγένειες μεταξύ των πληθυσμών

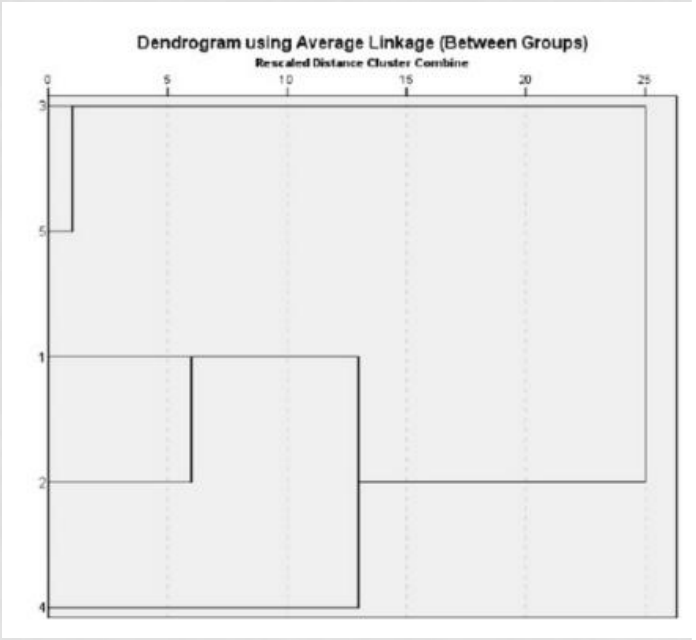


	Population	D1	D2	D3	D4	D5	D6	D7	var
1	I	9,9	1,0	3,0	4,0	10,0	55,0	15,0	
2	II	9,5	1,2	3,4	3,5	10,0	56,0	14,0	
3	III	9,7	1,1	3,2	3,6	10,5	58,0	16,0	
4	IV	9,6	,9	3,1	4,2	9,8	57,0	13,0	
5	V	10,1	,8	3,3	3,5	11,0	58,0	16,0	
6									

CLUSTER ANALYSIS (CA)-ΠΑΡΑΔΕΙΓΜΑ

- *Analyze > Classify > Hierarchical Cluster*





οι πληθυσμοί με βάση τα κρανιακά δεδομένα μπορούν να χωριστούν σε δύο ομάδες:

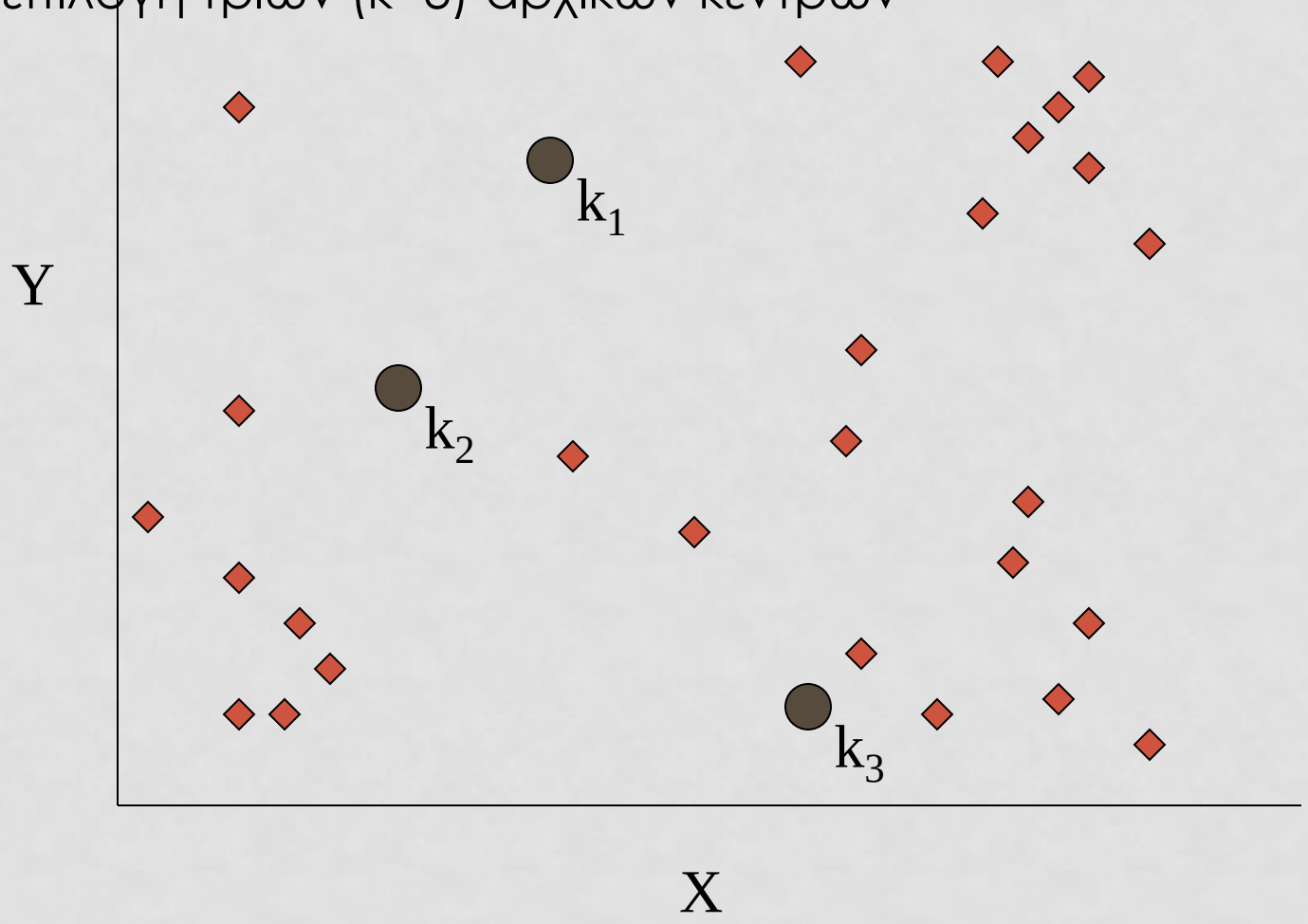
- Οι πληθυσμοί III και V έχουν στενή συγγένεια,
- ενώ οι I, II και IV σχηματίζουν μια δεύτερη ομάδα.
- Στην ομάδα αυτή οι I με τους II φαίνεται να σχηματίζουν μια υποομάδα

ΣΥΣΤΑΔΟΠΟΙΗΣΗ ΜΕ ΔΙΑΜΕΡΙΣΗ

- Μη ιεραρχική
- Δημιουργεί τις συστάδες σε ένα βήμα μόνο.
- (στις περισσότερες τεχνικές) ο χρήστης απαιτείται να εισάγει τον επιθυμητό αριθμό των συστάδων, k.
- Βασικός αλγόριθμος K-Means Είναι ένα εργαλείο σχεδιασμένο για να εκχωρεί τις περιπτώσεις σε ένα σταθερό αριθμό ομάδων (clusters), των οποίων τα χαρακτηριστικά δεν είναι ακόμη γνωστά, αλλά βασίζονται σε ένα σύνολο συγκεκριμένων μεταβλητών

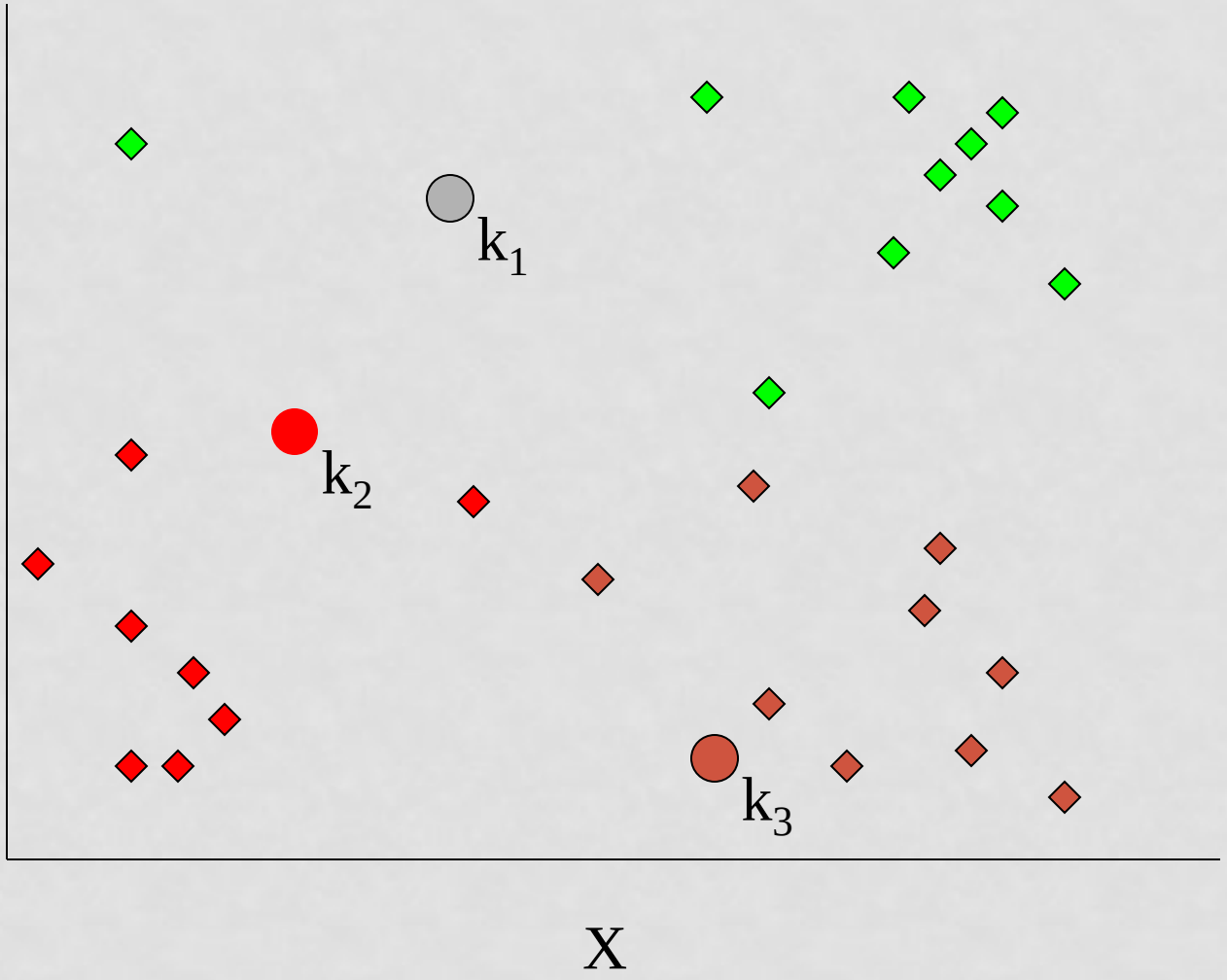
ΠΑΡΑΔΕΙΓΜΑ K-MEANS (ΣΕ 2 ΔΙΑΣΤΑΣΕΙΣ)

- Τυχαία επιλογή τριών (k=3) αρχικών κέντρων



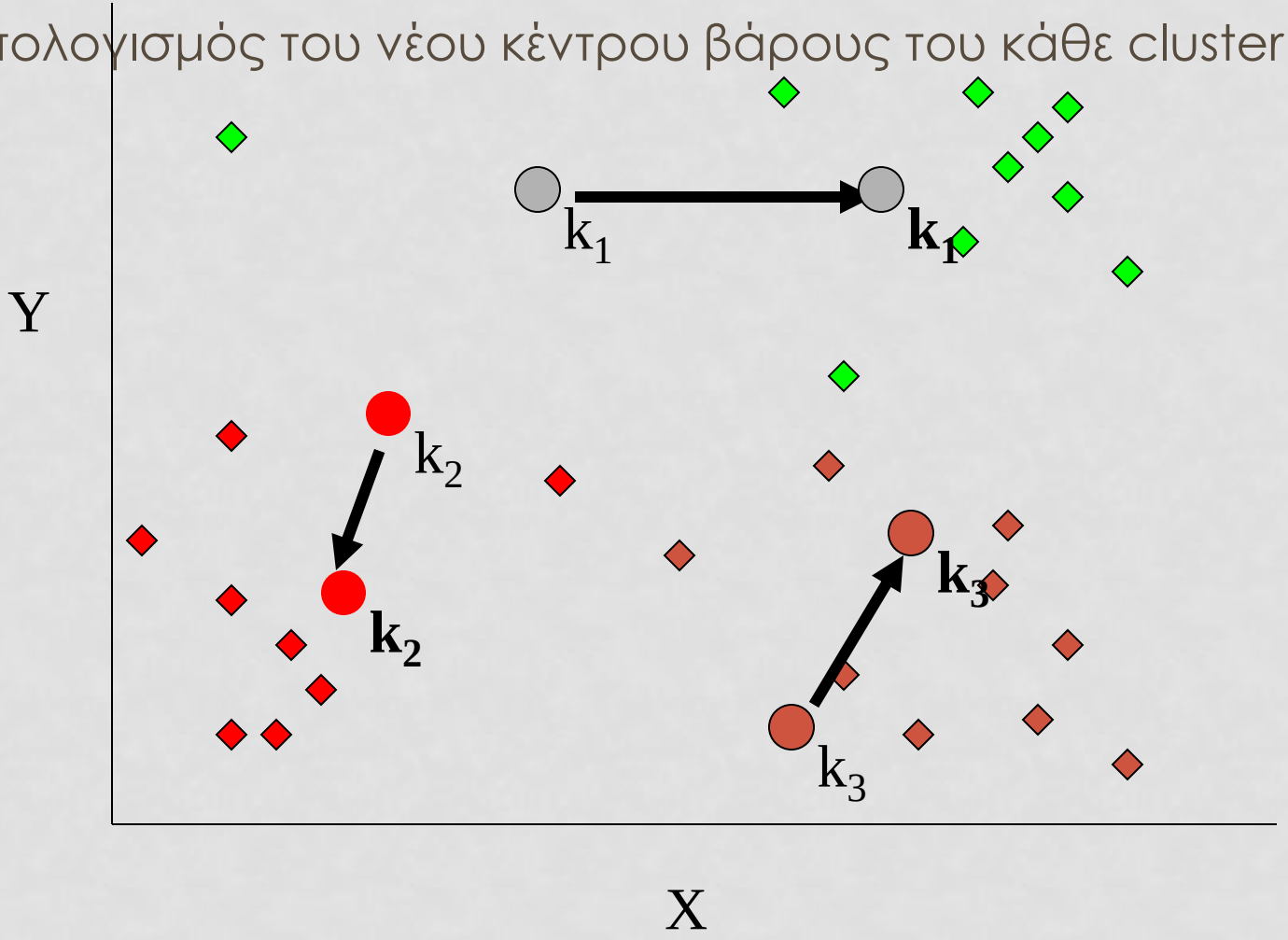
ΠΑΡΑΔΕΙΓΜΑ K-MEANS, 1Η ΕΠΑΝΑΛΗΨΗ

- Εκχώρηση κάθε στοιχείου στο πλησιέστερό του cluster (με βάση την απόσταση από το κέντρο του cluster)



ΠΑΡΆΔΕΙΓΜΑ K-MEANS, 1Η ΕΠΑΝΆΛΗΨΗ

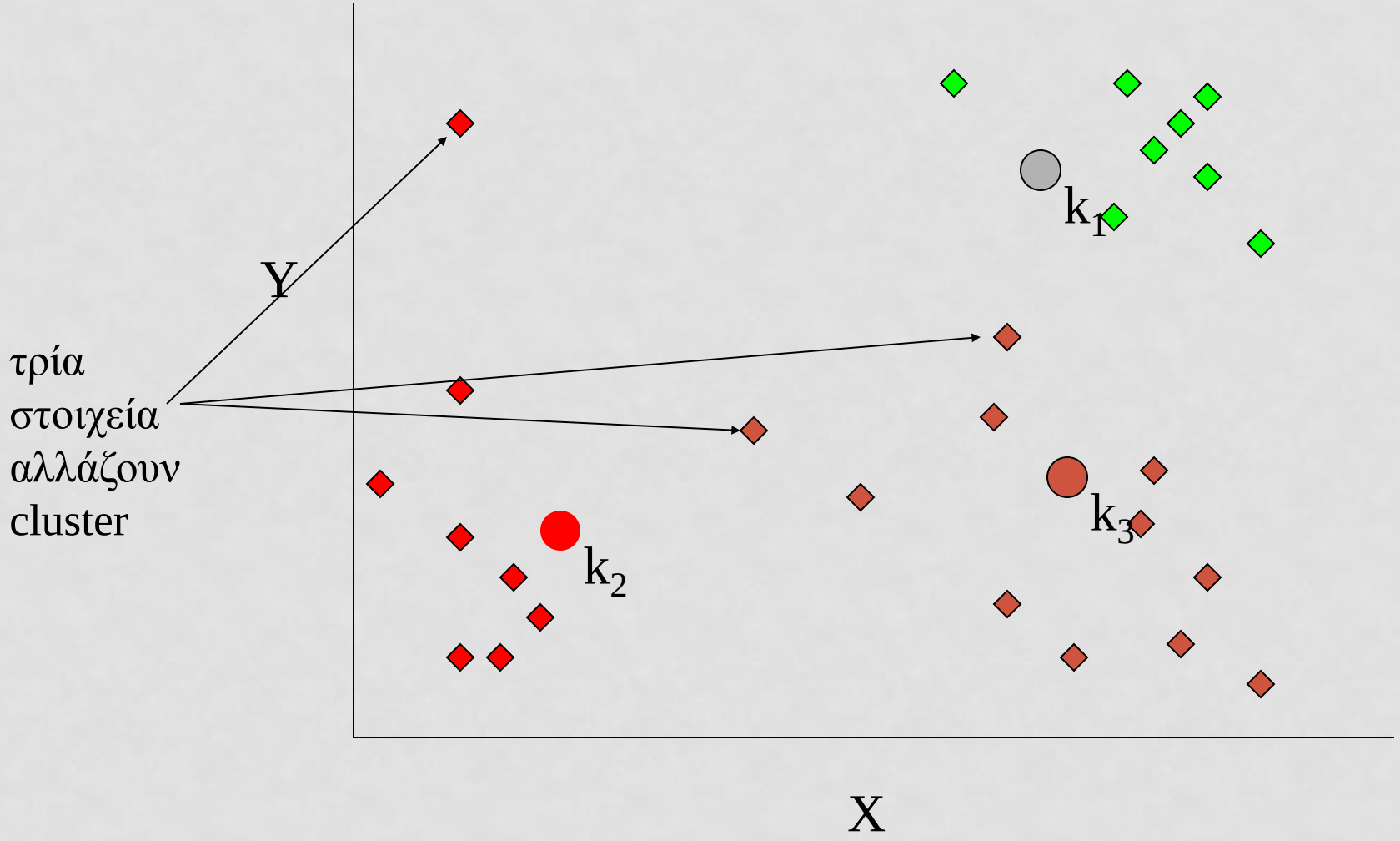
- Επανυπολογισμός του νέου κέντρου βάρους του κάθε cluster



ΠΑΡΑΔΕΙΓΜΑ K-MEANS, 2Η ΕΠΑΝΑΛΗΨΗ

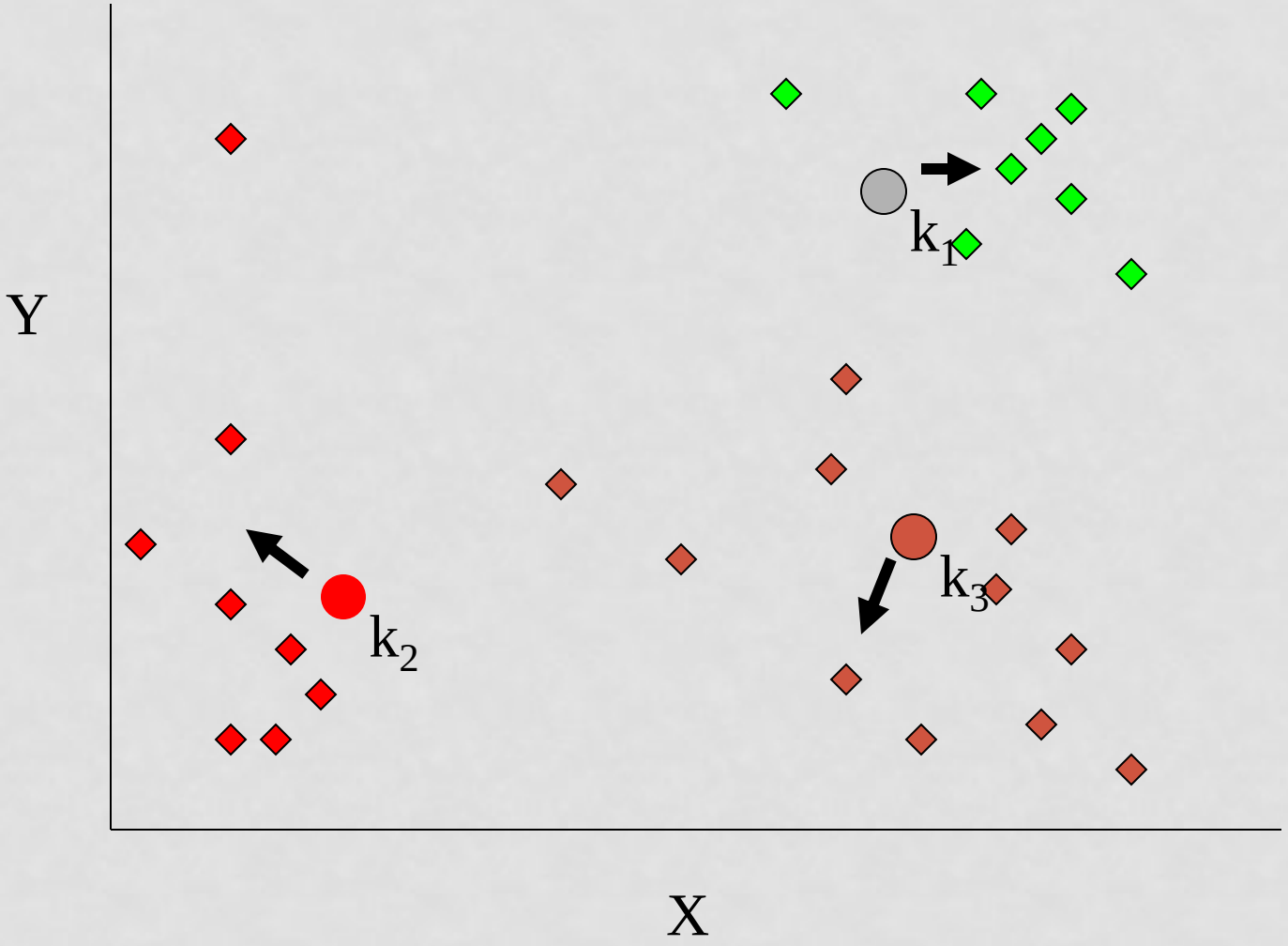
- Εκχώρηση κάθε στοιχείου στο πλησιέστερό του cluster (με βάση την απόσταση από το κέντρο του cluster)

ΠΑΡΑΔΕΙΓΜΑ K-MEANS, 2Η ΕΠΑΝΑΛΗΨΗ



ΠΑΡΑΔΕΙΓΜΑ K-MEANS, 2Η ΕΠΑΝΑΛΗΨΗ

- Επανυπολογισμός του νέου κέντρου βάρους του κάθε cluster



ΠΑΡΑΔΕΙΓΜΑ K-MEANS, 3Η ΕΠΑΝΑΛΗΨΗ

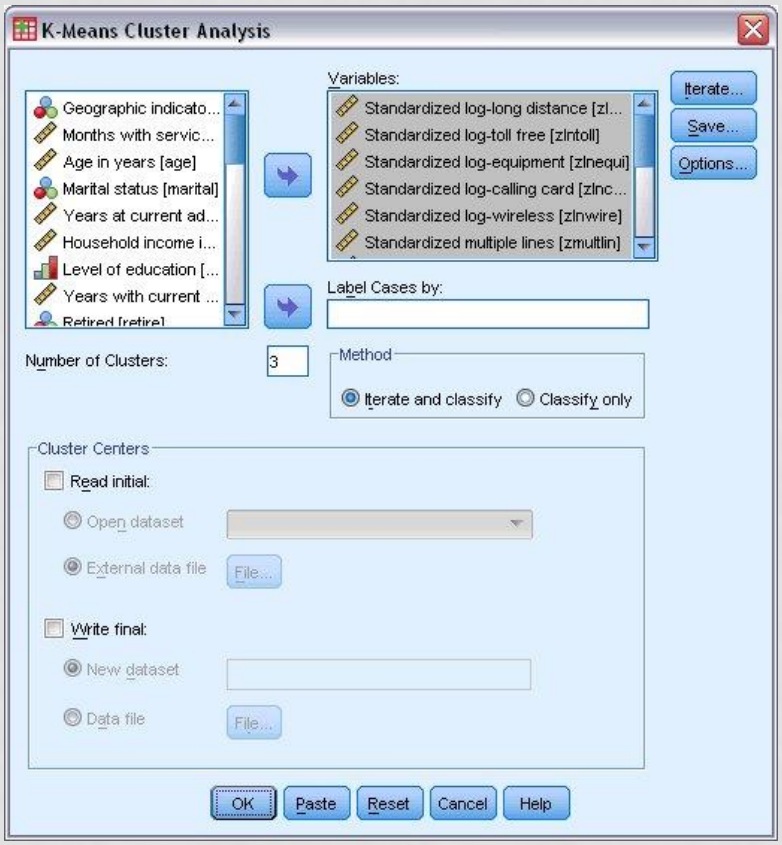
- Εκχώρηση κάθε στοιχείου στο πλησιέστερό του cluster (με βάση την απόσταση από το κέντρο του cluster)

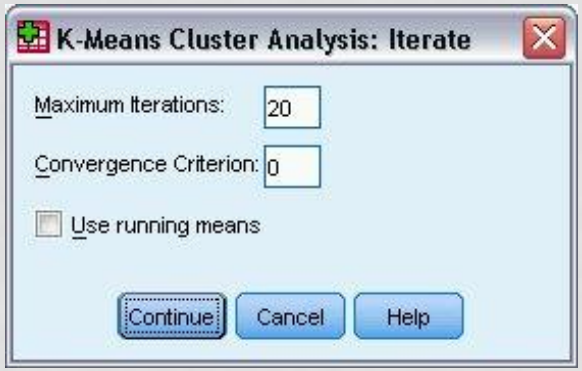
Δεν αλλάζει τίποτα
Άρα, τέλος !

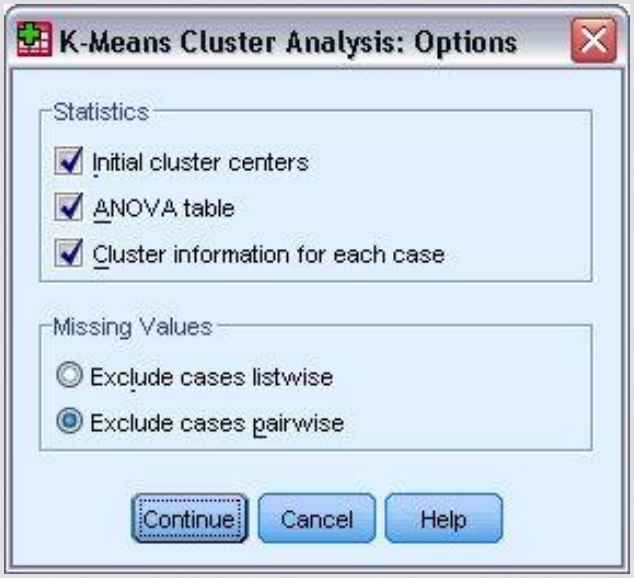
CLUSTER ANALYSIS (CA)-2ΟΠΑΡΑΔΕΙΓΜΑ

- Ένας φορέας παροχής τηλεπικοινωνιών θέλει να διαχωρίσει την πελατειακή του βάση με τα πρότυπα χρήσης της υπηρεσίας. Εάν οι πελάτες μπορούν να ταξινομηθούν με βάση τη χρήση, η εταιρεία μπορεί να προσφέρει περισσότερο ελκυστικά πακέτα στους πελάτες της. *(telco_extra.sav)*
- **Analyze** > **Classify** > **K-Means Cluster...**
- **Display Variable Labels** > **Sort by File Order**
- *Variables: Επιλέξτε Standardized log-long distance* έως *Standardized log-wireless* και *Standardized multiple lines* έως *Standardized electronic billing*
- Number of clusters=3
- **Iterate >** maximum iterations = 20
- **Options** > **ANOVA table**, **Cluster information for each group**, **Exclude cases pairwise**

CLUSTER ANALYSIS (CA)- 2οΠΑΡΑΔΕΙΓΜΑ







CLUSTER ANALYSIS (CA)- 2οΠΑΡΑΔΕΙΓΜΑ

- *Initial cluster centers for three-cluster solution*

	Cluster		
	1	2	3
Standardized log-long distance	2.48	-1.70	.12
Standardized log-toll free	2.34	-.20	-.39
Standardized log-equipment	1.34	-.65	.59
Standardized log-calling card	2.49	-.86	-1.28
Standardized log-wireless	1.14	-1.75	1.42
Standardized multiple lines	1.05	-.95	1.05
Standardized voice mail	1.51	1.51	1.51
Standardized paging	1.68	1.68	1.68
Standardized internet	1.31	-.76	1.31
Standardized caller id	1.04	1.04	-.96
Standardized call waiting	1.03	-.97	1.03
Standardized call forwarding	1.01	1.01	-.99
Standardized 3-way calling	1.00	1.00	-1.00
Standardized electronic billing	-.77	-.77	1.30

- *Iteration history for three-cluster solution*

	Change in Cluster Centers		
Iteration	1	2	3
1	3.298	3.590	3.491
2	1.016	.427	.931
3	.577	.320	.420
4	.240	.180	.195
5	.119	.125	.108
6	.093	.083	.027
7	.069	.094	.032
8	.059	.051	.018
9	.035	.085	.063
10	.025	.359	.333
11	.068	.439	.287
12	.079	.368	.177
13	.125	.139	.078
14	.077	.096	.020
15	.041	.047	.015
16	.014	.027	.000
17	.019	.038	.000
18	.000	.000	.000

CLUSTER ANALYSIS (CA)- 2οΠΑΡΑΔΕΙΓΜΑ

- ANOVA: ποιες μεταβλητές συμβάλλουν περισσότερο στην ομαδοποίηση. Μεταβλητές με μεγάλες τιμές F παρέχουν το μεγαλύτερο διαχωρισμό μεταξύ των συστάδων

	Cluster		Error			
	Mean Square	df	Mean Square	df	F	Sig.
Standardized log-long distance	13.063	2	.976	997	13.387	.000
Standardized log-toll free	43.418	2	.820	472	52.932	.000
Standardized log-equipment	99.056	2	.488	383	202.999	.000
Standardized log-calling card	6.301	2	.984	675	6.402	.002
Standardized log-wireless	52.879	2	.646	293	81.873	.000
Standardized multiple lines	38.032	2	.926	997	41.084	.000
Standardized voice mail	236.301	2	.528	997	447.554	.000
Standardized paging	298.992	2	.402	997	743.348	.000
Standardized internet	123.447	2	.754	997	163.642	.000
Standardized caller id	308.104	2	.384	997	802.474	.000
Standardized call waiting	294.674	2	.411	997	717.172	.000
Standardized call forwarding	288.343	2	.424	997	680.718	.000
Standardized 3-way calling	262.397	2	.476	997	551.678	.000
Standardized electronic billing	112.782	2	.776	997	145.381	.000

CLUSTER ANALYSIS (CA)- 2οΠΑΡΑΔΕΙΓΜΑ

- *Final cluster centers*
  - Πελάτες της ομάδας 1 τείνουν να είναι περισσότερο σπάταλοι και αγοράζουν πολλές υπηρεσίες.
  - Πελάτες της ομάδας 2 τείνουν να είναι μέτρια σπάταλοι που αγοράζουν τις υπηρεσίες "calling'
  - Πελάτες της ομάδας 3 τείνουν να ξοδεύουν πολύ λίγο και δεν αγοράζουν πολλές υπηρεσίες.

	Cluster		
	1	2	3
Standardized log-long distance	.05	.22	-.16
Standardized log-toll free	.24	.12	-1.05
Standardized log-equipment	.81	-.19	-.69
Standardized log-calling card	.17	.02	-.17
Standardized log-wireless	.42	-.75	-1.00
Standardized multiple lines	.48	-.29	-.05
Standardized voice mail	1.26	-.24	-.44
Standardized paging	1.43	-.38	-.44
Standardized internet	.81	-.59	-.02
Standardized caller id	.82	.71	-.81
Standardized call waiting	.76	.72	-.80
Standardized call forwarding	.78	.69	-.79
Standardized 3-way calling	.74	.67	-.75
Standardized electronic billing	.70	-.63	.05

CLUSTER ANALYSIS (CA)- 2οΠΑΡΑΔΕΙΓΜΑ

- Αποστάσεις μεταξύ των τελικών κέντρων των clusters
- Clusters 1 και 3 είναι πολύ διαφορετικά μεταξύ τους.
- Cluster 2 είναι εξίσου παρόμοια με τα clusters 1 και 3

Cluster	1	2	3
1		3.500	4.863
2	3.500		3.396
3	4.863	3.396	

Cluster	1	226.000
	2	292.000
	3	482.000
Valid		1000.000
Missing		.000

- Τρέχουμε ξανά τον αλγόριθμο για 4 clusters.

ΑΝΑΛΥΣΗ ΧΡΟΝΟΛΟΓΙΚΩΝ ΣΕΙΡΩΝ & ΠΡΟΒΛΕΨΗ

- Το SPSS είναι ιδιαίτερα καλό στην χρησιμοποίηση των υποδειγμάτων ARIMA για την διεξαγωγή προβλέψεων σε χρονολογικές σειρές
- Τα υποδείγματα ARIMA (autoregressive integrated moving average) είναι υποδείγματα που χρησιμοποιούνται εκτεταμένα για τις προβλέψεις χρονολογικών σειρών
- Ο σκοπός των υποδειγμάτων είναι να πραγματοποιήσουν προβλέψεις για μια χρονολογική σειρά Y_t με βάση μόνον τις παρελθούσες τιμές της σειράς και χωρίς άλλη πληροφόρηση διαρθρωτικής μορφής
  - Πχ δεν χρειάζεται να έχουμε πληροφόρηση σχετικά με το ποιες ερμηνευτικές μεταβλητές επηρεάζουν την Y_t

ΑΝΑΛΥΣΗ ΧΡΟΝΟΛΟΓΙΚΩΝ ΣΕΙΡΩΝ & ΠΡΟΒΛΕΨΗ

- Η βασική αρχή των υποδειγμάτων είναι ότι η σειρά Y_t πρέπει αν είναι στάσιμη (stationary) δηλαδή να μην υπάρχει προφανής τάση στα επίπεδα ή στην διακύμανση της μεταβλητής.
- Αν η σειρά δεν είναι στάσιμη στα επίπεδα θα πρέπει να πάρουμε πρώτες διαφορές, δηλαδή να σχηματίσουμε την σειρά

$$DY_t = Y_t - Y_{t-1}$$

- και να εφαρμόσουμε τα υποδείγματα ARIMA στην νέα σειρά DY_t.
- Αν ούτε και η DY_t είναι στάσιμη παίρνουμε δεύτερες διαφορές

$$D^2Y_t = DY_t - DY_{t-1} = Y_t - 2Y_{t-1} + Y_{t-2}$$

ΑΝΑΛΥΣΗ ΧΡΟΝΟΛΟΓΙΚΩΝ ΣΕΙΡΩΝ & ΠΡΟΒΛΕΨΗ

- κλπ μέχρι να επιτύχουμε στασιμότητα την οποία κρίνουμε διαγραμματικά.
- Γενικά την τάξη διαφορών την οποία πρέπει να επιβάλλουμε για να έχουμε στασιμότητα την συμβολίζουμε με d.
- Το υπόδειγμα ARIMA(p,d,q) έχει την μορφή

$$z_t = \beta_0 + \beta_1 z_{t-1} + \beta_2 z_{t-2} + ... + \beta_p z_{t-p} + e_t + \alpha_1 e_{t-1} + \alpha_2 e_{t-2} + ... + \alpha_q e_{t-q}$$

$$z_t = D^d Y_t$$

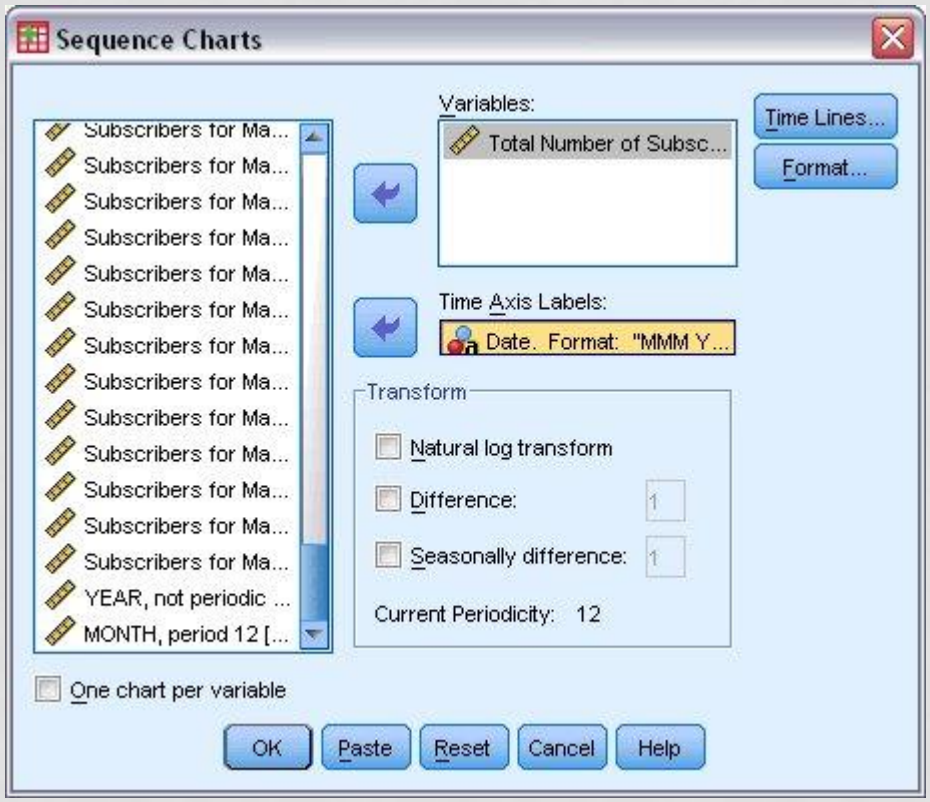
- όπου β_i είναι γνωστοί σαν αυτοπαλίνδρομοι (autoregressive) συντελεστές και
- α_i είναι γνωστοί σαν συντελεστές κινητού μέσου (moving average)

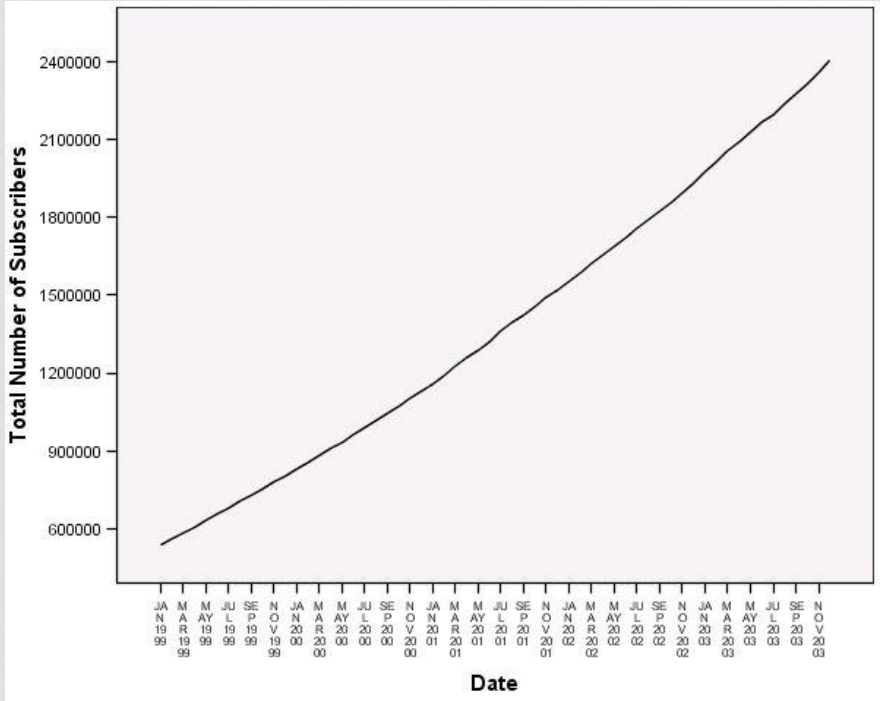
ΠΡΟΒΛΕΨΗ

- Ένας αναλυτής ενός εθνικού παρόχου ευρυζωνικών είναι υποχρεωμένος να δημιουργήσει προβλέψεις των συνδρομών των χρηστών, προκειμένου να προβλέψει τη χρήση του εύρους ζώνης.
- Οι προβλέψεις χρειάζονται για κάθε μία από τις 85 τοπικές αγορές για συνθέσουν την εθνική βάση συνδρομητών.
- Μηνιαία ιστορικά δεδομένα συλλέγονται στο αρχείο broadband_1.sav
- Θέλουμε να κάνουμε πρόβλεψη για τους επόμενους τρεις μήνες για κάθε μία από τις 85 τοπικές αγορές

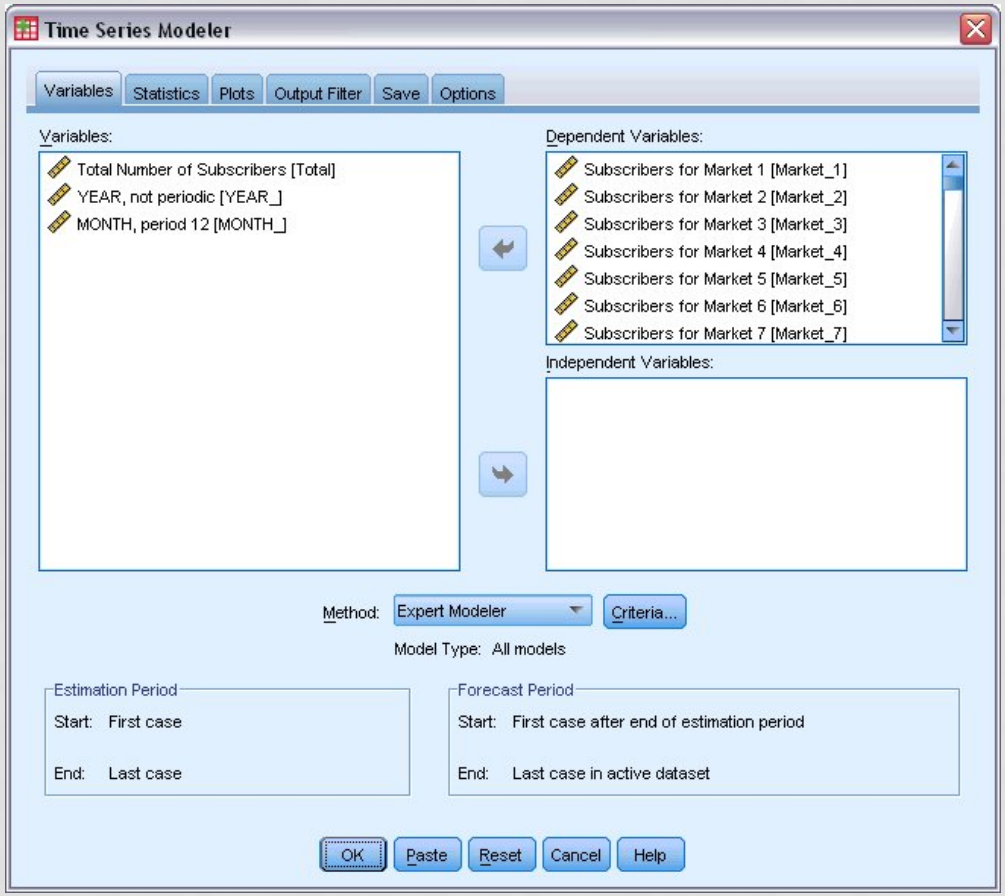
ΠΡΟΒΛΕΨΗ

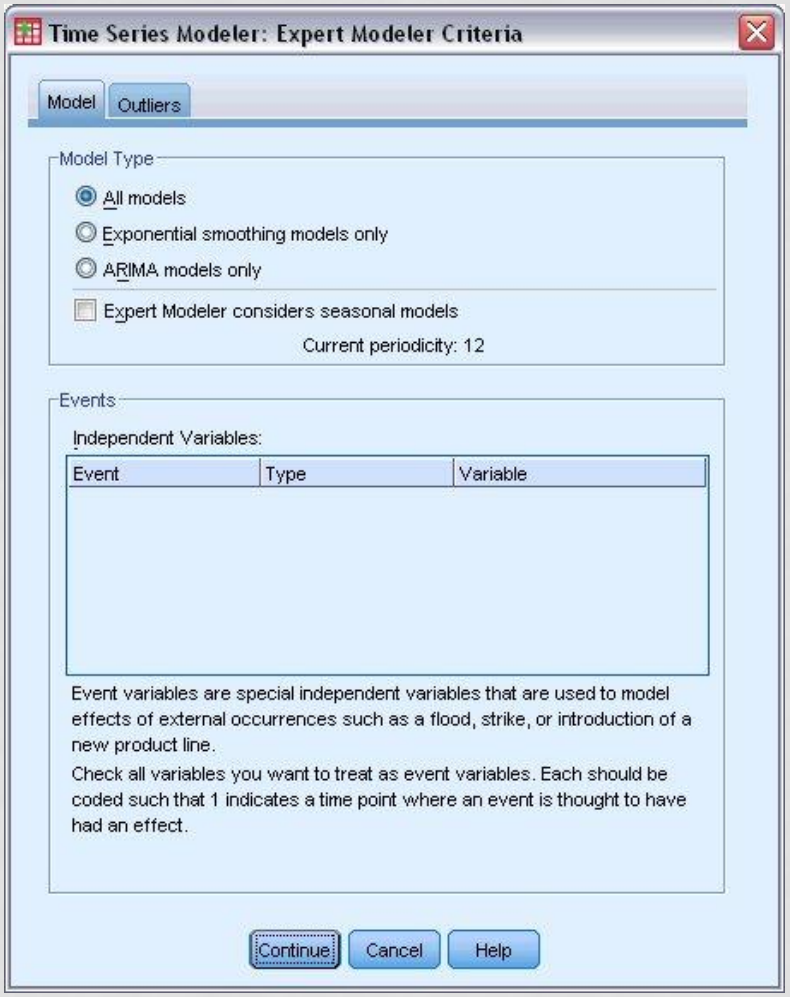
- Έλεγχος εποχικότητας με γράφημα
- Analyze > Forecasting > Sequence Charts...





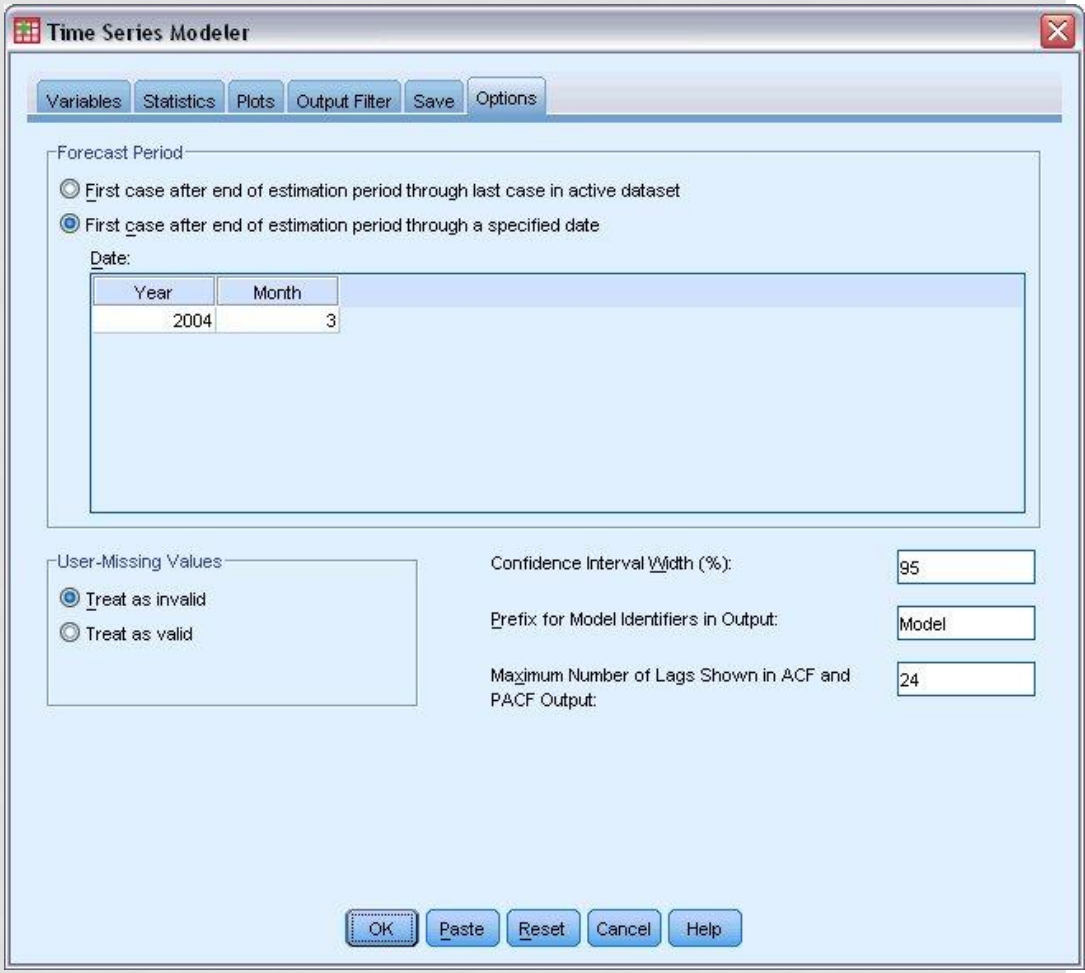
ΠΡΟΒΛΕΨΗ


Time Series Modeler
Variables
Statistics
Plots
Output Filter
Save
Options
Variables:
Total Number of Subscribers [Total]
YEAR, not periodic [YEAR_]
MONTH, period 12 [MONTH_]
Dependent Variables:
Subscribers for Market 1 [Market_1]
Subscribers for Market 2 [Market_2]
Subscribers for Market 3 [Market_3]
Subscribers for Market 4 [Market_4]
Subscribers for Market 5 [Market_5]
Subscribers for Market 6 [Market_6]
Subscribers for Market 7 [Market_7]
Independent Variables:
Method: Expert Modeler
Criteria...
Model Type: All models
Estimation Period
Start: First case
End: Last case
Forecast Period
Start: First case after end of estimation period
End: Last case in active dataset
OK
Paste
Reset
Cancel
Help


Time Series Modeler: Expert Modeler Criteria
Model
Outliers
Model Type
All models
Exponential smoothing models only
ARIMA models only
Expert Modeler considers seasonal models
Current periodicity: 12
Events
Independent Variables:
Event
Type
Variable
Event variables are special independent variables that are used to model effects of external occurrences such as a flood, strike, or introduction of a new product line.
Check all variables you want to treat as event variables. Each should be coded such that 1 indicates a time point where an event is thought to have had an effect.
Continue
Cancel
Help

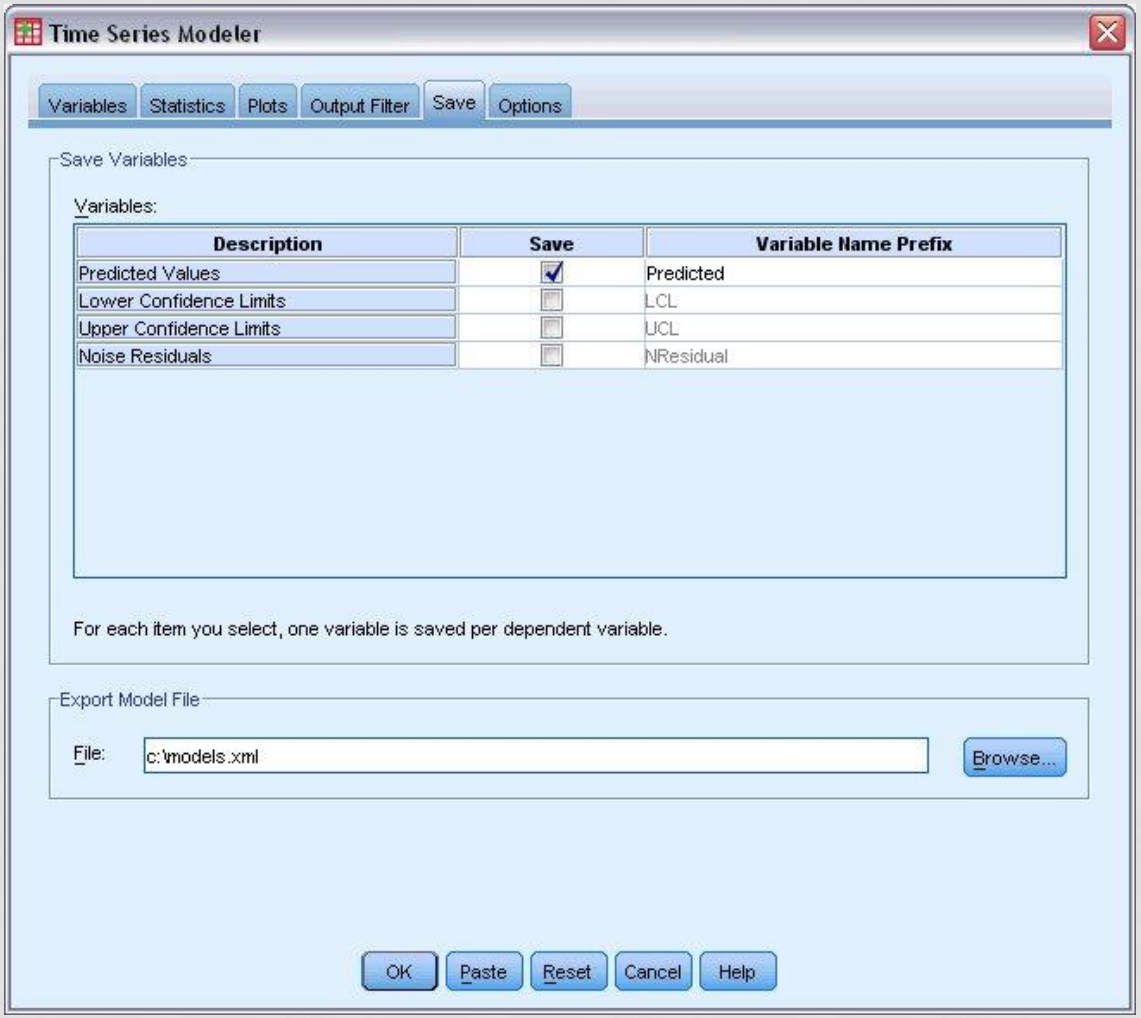
ΠΡΟΒΛΕΨΗ

- Το σύνολο δεδομένων περιλαμβάνει δεδομένα από τον Ιανουάριο 1999 έως το Δεκέμβριο του 2003. Με τις τρέχουσες ρυθμίσεις, η διάρκεια της περιόδου πρόβλεψης θα είναι από τον Ιανουάριο του 2004 έως τον Μάρτιο του 2004.

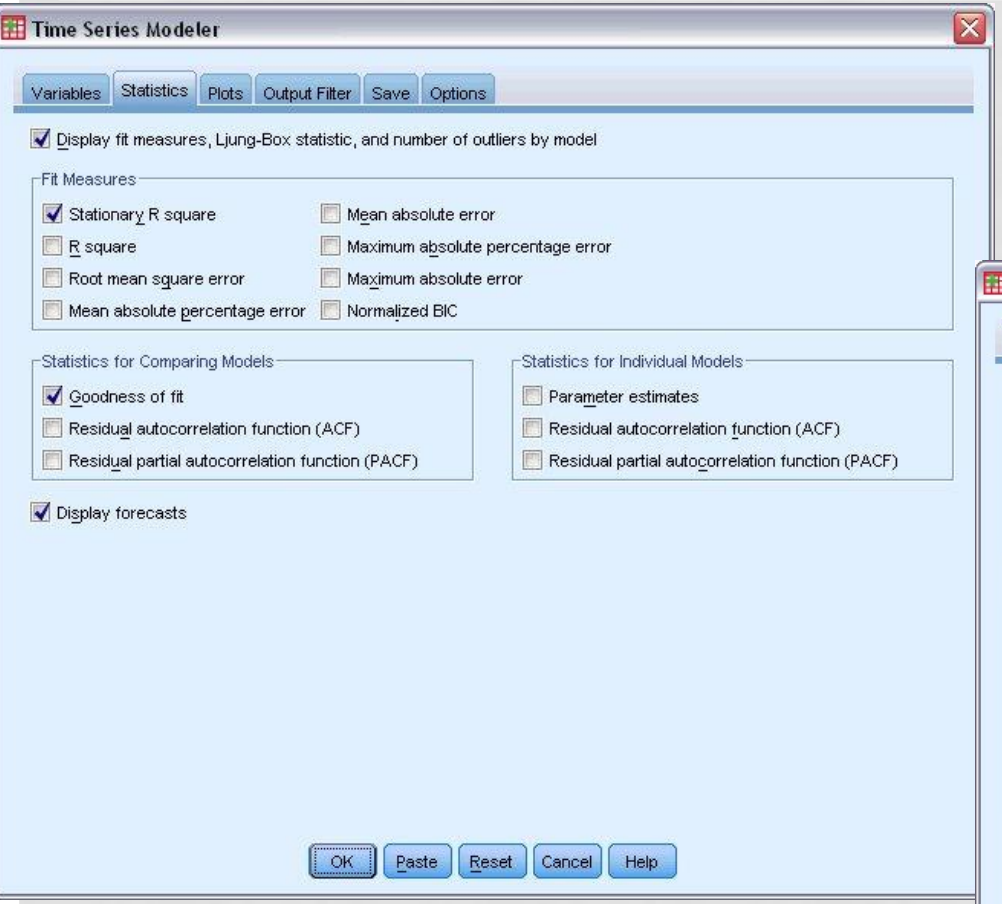


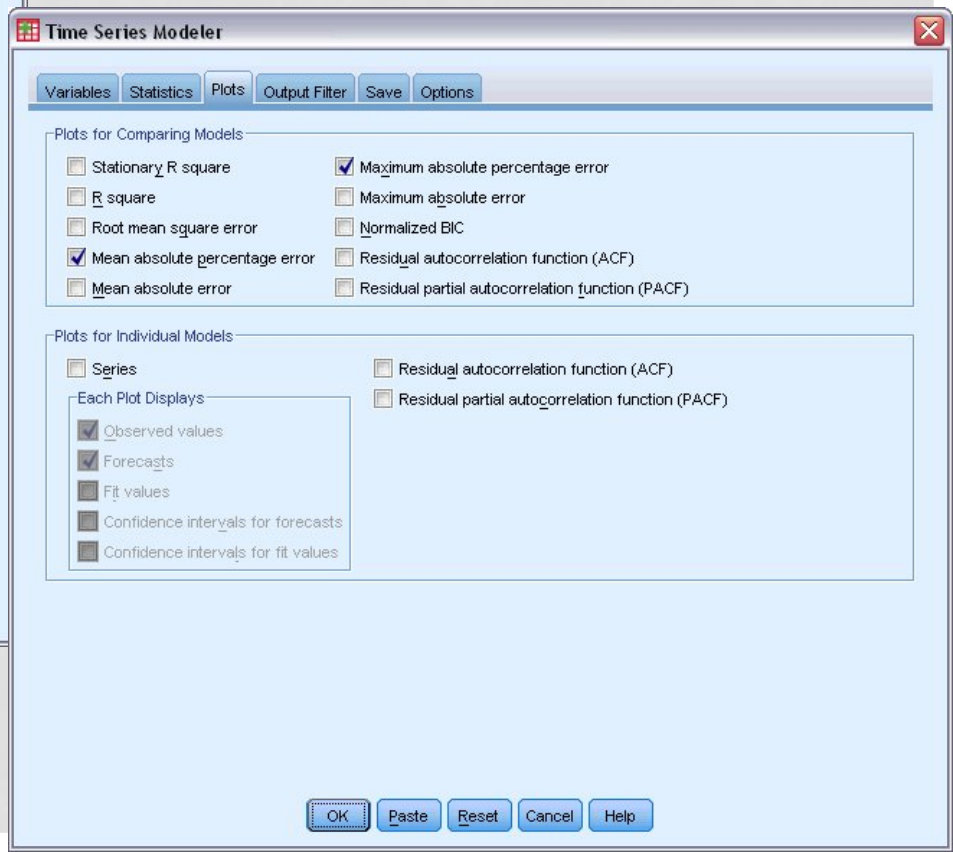
ΠΡΟΒΛΕΨΗ

- The model predictions are saved as new variables



ΠΡΟΒΛΕΨΗ


Time Series Modeler
Variables
Statistics
Plots
Output Filter
Save
Options
Display fit measures, Ljung-Box statistic, and number of outliers by model
Fit Measures
Stationary R square
R square
Root mean square error
Mean absolute percentage error
Mean absolute error
Maximum absolute percentage error
Maximum absolute error
Normalized BIC
Statistics for Comparing Models
Goodness of fit
Residual autocorrelation function (ACF)
Residual partial autocorrelation function (PACF)
Statistics for Individual Models
Parameter estimates
Residual autocorrelation function (ACF)
Residual partial autocorrelation function (PACF)
Display forecasts
OK
Paste
Reset
Cancel
Help


Time Series Modeler
Variables
Statistics
Plots
Output Filter
Save
Options
Plots for Comparing Models
Stationary R square
R square
Root mean square error
Mean absolute percentage error
Mean absolute error
Maximum absolute percentage error
Maximum absolute error
Normalized BIC
Residual autocorrelation function (ACF)
Residual partial autocorrelation function (PACF)
Plots for Individual Models
Series
Each Plot Displays
Observed values
Forecasts
Fit values
Confidence intervals for forecasts
Confidence intervals for fit values
Residual autocorrelation function (ACF)
Residual partial autocorrelation function (PACF)
OK
Paste
Reset
Cancel
Help

ΠΡΟΒΛΕΨΗ

- Ο Data Editor παρουσιάζει τις νέες μεταβλητές που περιέχουν τις προβλέψεις του μοντέλου.
- Αν και μόνο δύο φαίνονται εδώ, υπάρχουν 85 νέες μεταβλητές, μία για κάθε μία από τις 85 εξαρτώμενες σειρές.
- Τα ονόματα των μεταβλητών αποτελούνται από το προεπιλεγμένο πρόθεμα Predicted, που ακολουθείται από το όνομα της σχετικής εξαρτημένης μεταβλητής (για παράδειγμα, Market_1), ακολουθούμενο από ένα αναγνωριστικό μοντέλου (για παράδειγμα, Model_1)

YEAR_	MONTH_	DATE_	Predicted_Market_1_Model_1	Predicted_Market_2_Model_2
2003	10	OCT 2003	11820	51084
2003	11	NOV 2003	11857	51273
2003	12	DEC 2003	11687	53082
2004	1	JAN 2004	11503	54893
2004	2	FEB 2004	11447	55856
2004	3	MAR 2004	11390	56704

ΠΡΟΒΛΕΨΗ

- Τρεις νέες περιπτώσεις, που περιέχουν τις προβλέψεις για τον Ιανουάριο του 2004 έως τον Μάρτιο του 2004, έχουν προστεθεί στο σύνολο δεδομένων, μαζί με τις ετικέτες για την ημερομηνία που δημιουργήθηκαν αυτόματα.
- Κάθε μία από τις νέες μεταβλητές περιέχει τις προβλέψεις του μοντέλου για την περίοδο εκτίμησης (Ιανουάριος 1999 έως Δεκέμβριος 2003), η οποία σας επιτρέπει να δείτε πόσο καλά το μοντέλο ταιριάζει με τις γνωστές τιμές.)

YEAR_	MONTH_	DATE_	Predicted_Market_1_Model_1	Predicted_Market_2_Model_2
2003	10	OCT 2003	11820	51084
2003	11	NOV 2003	11857	51273
2003	12	DEC 2003	11687	53082
2004	1	JAN 2004	11503	54893
2004	2	FEB 2004	11447	55856
2004	3	MAR 2004	11390	56704

ΠΡΟΒΛΕΨΗ

- upper confidence limits (UCL) and lower confidence limits (LCL) for the forecasted values (95% by default)

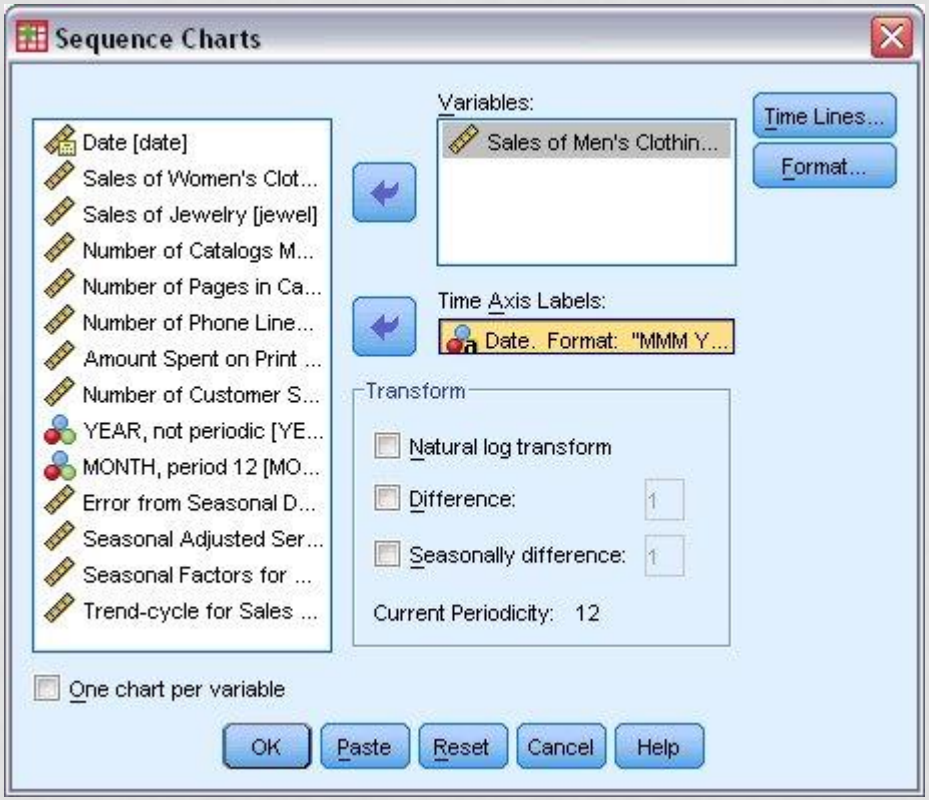
Model		JAN 2004	FEB 2004	MAR 2004
Subscribers for Market 1-Model_1	Forecast	11503	11447	11390
	UCL	11686	11767	11870
	LCL	11321	11126	10910
Subscribers for Market 2-Model_2	Forecast	54893	55856	56704
	UCL	55632	57195	58575
	LCL	54154	54518	54832
Subscribers for Market 3-Model_3	Forecast	59656	59305	58954
	UCL	60457	60753	61158
	LCL	58856	57857	56750
Subscribers for Market 4-Model_4	Forecast	18235	18424	18628
	UCL	18413	18731	19121
	LCL	18058	18116	18136

ΑΝΑΛΥΣΗ ΧΡΟΝΟΛΟΓΙΚΩΝ ΣΕΙΡΩΝ

- Μια εταιρεία, που ενδιαφέρεται για την ανάπτυξη ενός μοντέλου πρόβλεψης, έχει συλλέξει στοιχεία για τις μηνιαίες πωλήσεις των ανδρικών ενδυμάτων, μαζί με αρκετές σειρές που θα μπορούσαν να χρησιμοποιηθούν για να εξηγήσουν ορισμένες από τις διακυμάνσεις των πωλήσεων.
- Πιθανές προγνωστικοί παράγοντες περιλαμβάνουν τον αριθμό των καταλόγων που ταχυδρομήθηκαν, τον αριθμός των σελίδων του καταλόγου, τον αριθμό των τηλεφωνικών γραμμών που ήταν ανοικτές για την παραγγελία, το ποσό που δαπανήθηκε για την έντυπη διαφήμιση, και ο αριθμός των αντιπροσώπων εξυπηρέτησης πελατών.
- Είναι κάποιος από αυτούς τους προγνωστικούς παράγοντες χρήσιμος για την πρόβλεψη;

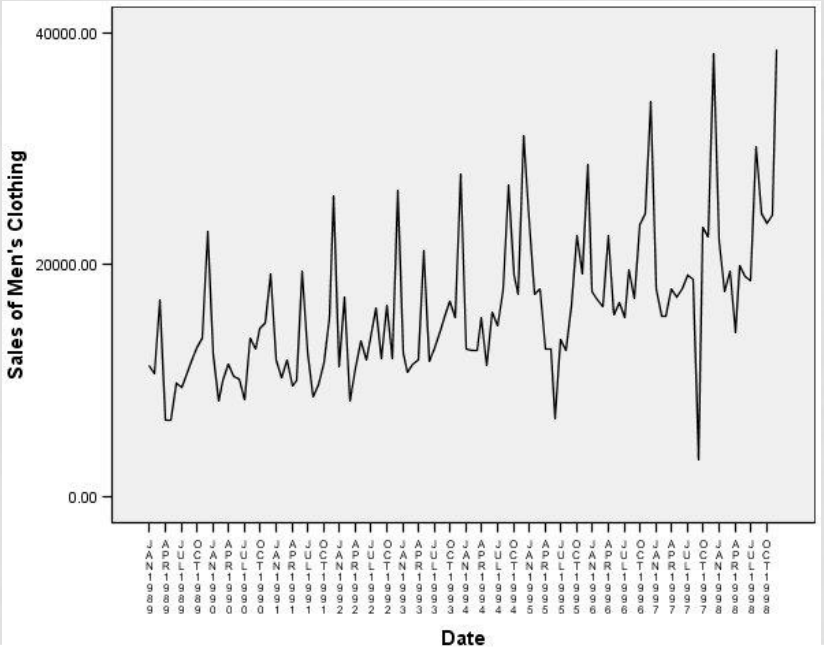
ΑΝΑΛΥΣΗ ΧΡΟΝΟΛΟΓΙΚΩΝ ΣΕΙΡΩΝ

- **Analyze > Forecasting > Sequence Charts...**



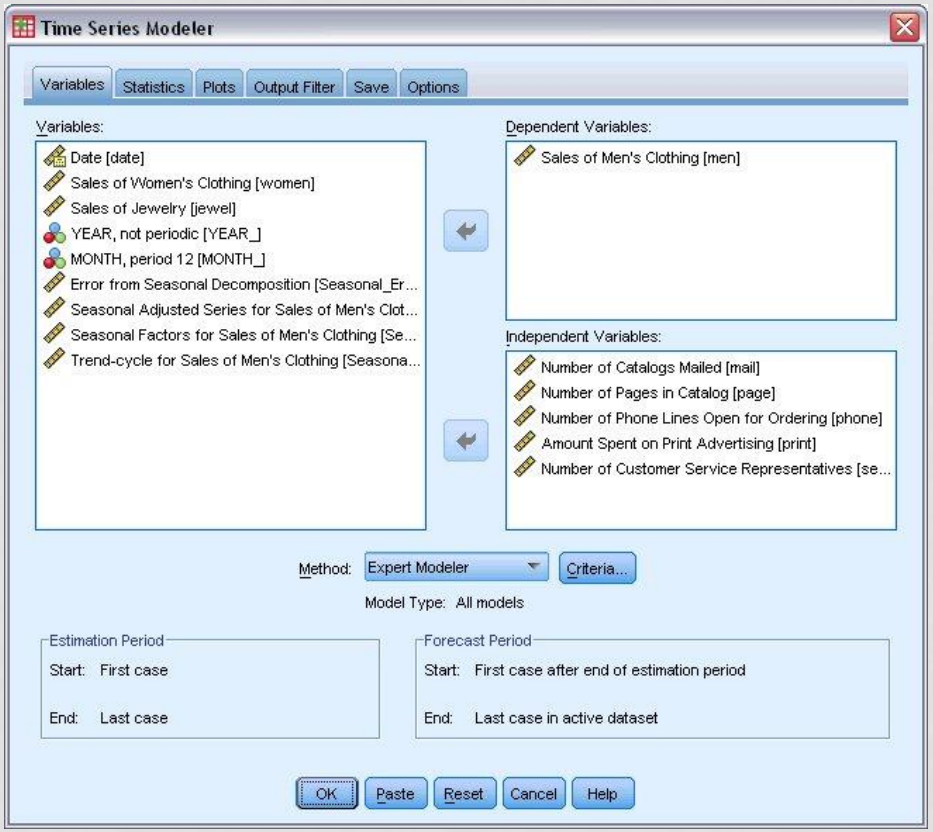
ΑΝΑΛΥΣΗ ΧΡΟΝΟΛΟΓΙΚΩΝ ΣΕΙΡΩΝ

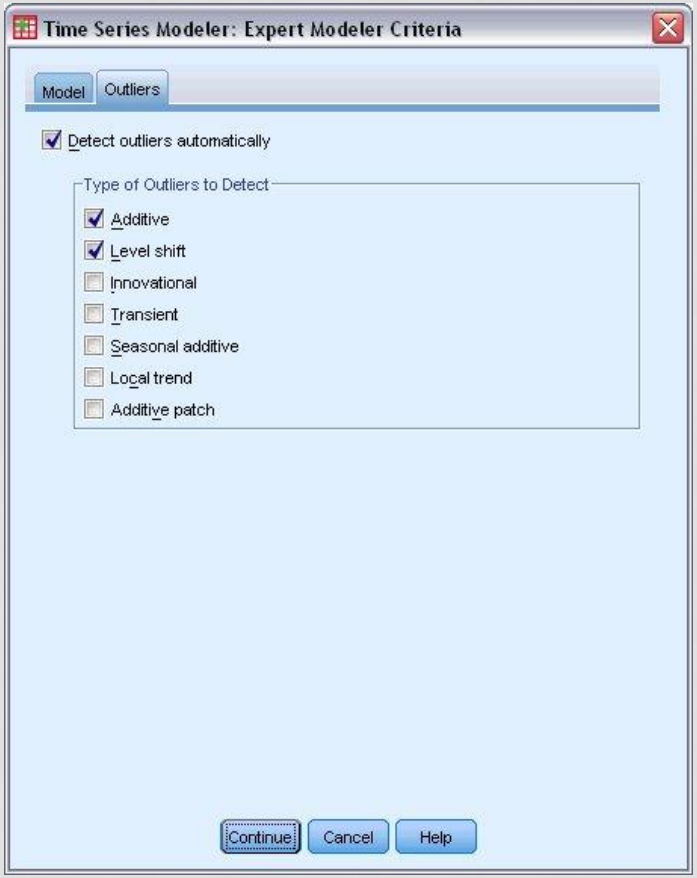
- Η σειρά παρουσιάζει πολλές κορυφές, πολλές από τις οποίες φαίνεται να ισαπέχουν, καθώς και μια σαφή ανοδική τάση. Οι ισαπέχουσες κορυφές υποδηλώνουν την παρουσία μιας περιοδικής συνιστώσας με τη χρονολογική σειρά



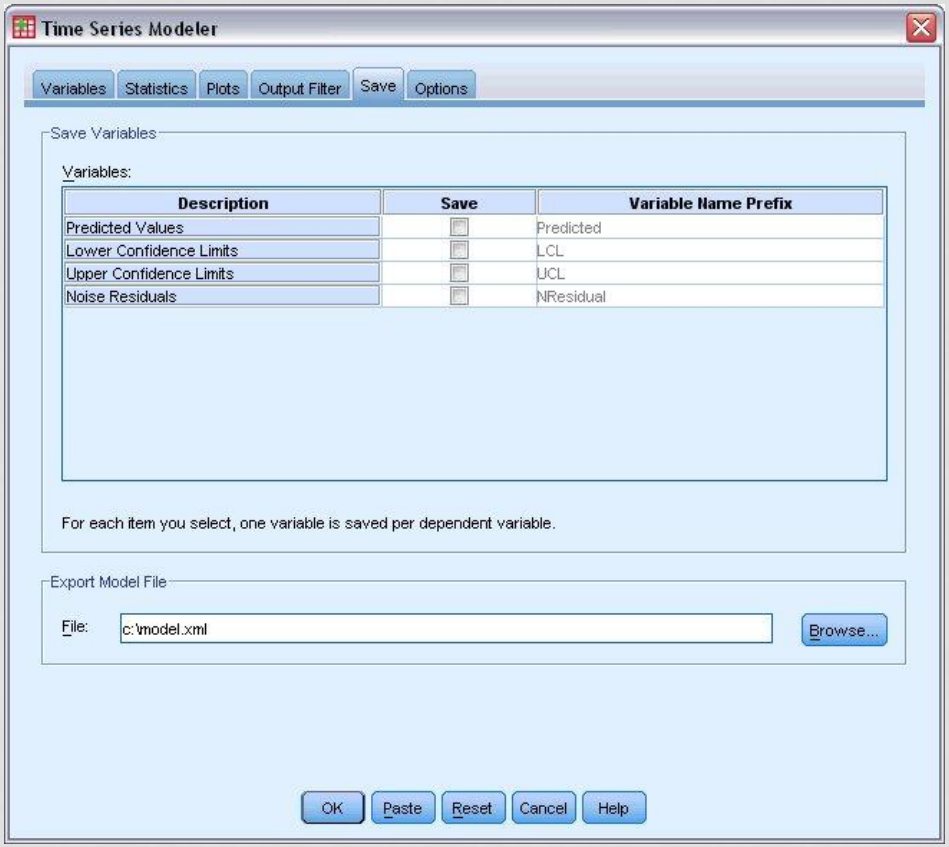
ΑΝΑΛΥΣΗ ΧΡΟΝΟΛΟΓΙΚΩΝ ΣΕΙΡΩΝ

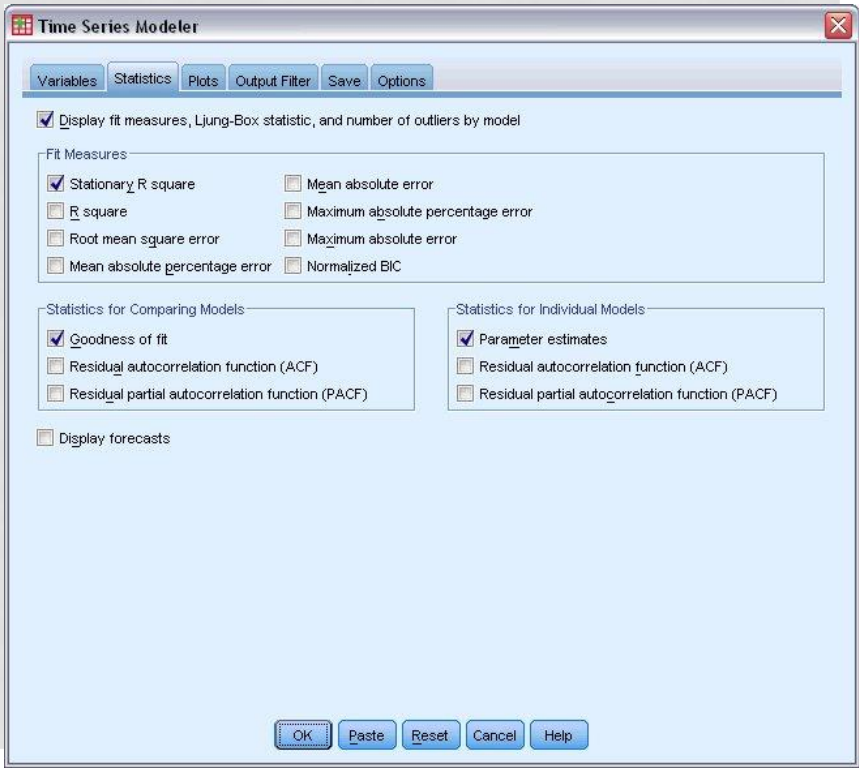
- **Analyze > Forecasting > Create Models...**



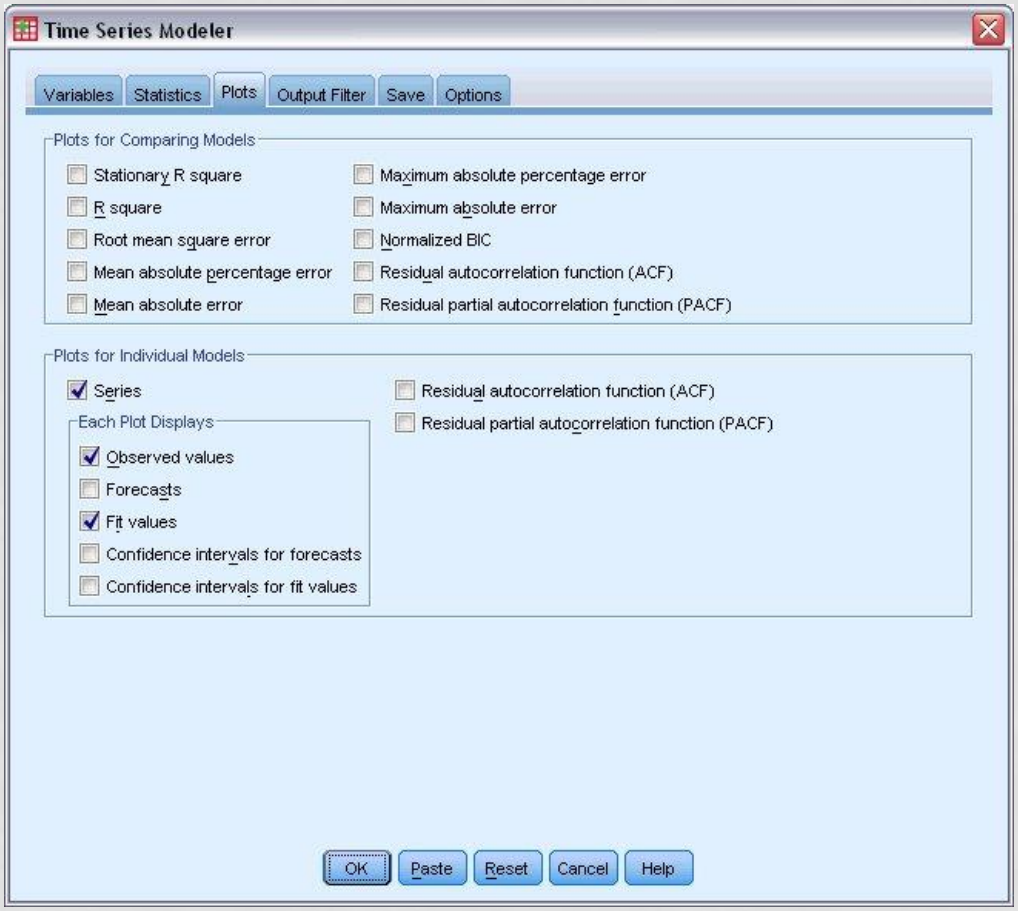


ΑΝΑΛΥΣΗ ΧΡΟΝΟΛΟΓΙΚΩΝ ΣΕΙΡΩΝ



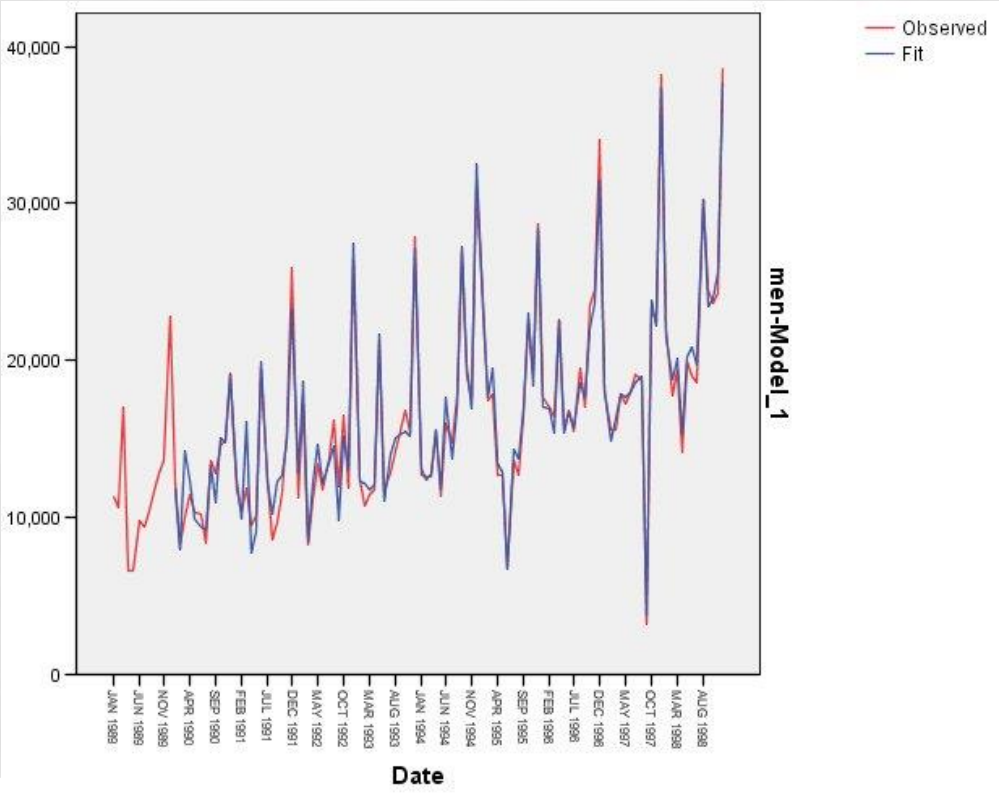


ΑΝΑΛΥΣΗ ΧΡΟΝΟΛΟΓΙΚΩΝ ΣΕΙΡΩΝ



ΑΝΑΛΥΣΗ ΧΡΟΝΟΛΟΓΙΚΩΝ ΣΕΙΡΩΝ

- Οι προβλεφθείσες τιμές δείχνουν καλή συμφωνία με τις παρατηρούμενες τιμές, υποδεικνύοντας ότι το μοντέλο έχει ικανοποιητική προγνωστική ικανότητα



ΑΝΑΛΥΣΗ ΧΡΟΝΟΛΟΓΙΚΩΝ ΣΕΙΡΩΝ

- Ο πίνακας που περιγράφει το μοντέλο περιέχει μια καταχώρηση για κάθε εκτιμώμενο μοντέλο και περιλαμβάνει model identifier και model type.
- Το αναγνωριστικό του μοντέλου αποτελείται από το όνομα (ή ετικέτα) της συνδεδεμένης εξαρτημένης μεταβλητής και ένα όνομα από το σύστημα.
- Στο παρόν παράδειγμα, η εξαρτημένη μεταβλητή είναι Sales of Men's Clothing και το σύστημα έδωσε το όνομα Model_1.

			Model Type
Model ID	Sales of Men's Clothing	Model_1	ARIMA(0,0,0)(0,1,0)

ΑΝΑΛΥΣΗ ΧΡΟΝΟΛΟΓΙΚΩΝ ΣΕΙΡΩΝ

- ARIMA(*p*,*d*,*q*)(*P*,*D*,*Q*),
- *p* is the order of autoregression,
- *d* is the order of differencing (or integration), and
- *q* is the order of moving-average, and (*P*,*D*,*Q*)
- Stationary *R*-squared: παρέχει μια εκτίμηση του ποσοστού της συνολικής διακύμανσης της σειράς που εξηγείται από το μοντέλο και είναι προτιμότερο από το συνηθισμένο R-squared όταν υπάρχει μια τάση ή εποχικές διακυμάνσεις
- Η στατιστική Ljung-Box, παρέχει μια ένδειξη για το αν το μοντέλο έχει καθοριστεί σωστά. Μια τιμή σημαντικότητας μικρότερη από 0,05 σημαίνει ότι υπάρχει δομή στην παρατηρούμενη σειρά που δεν εξηγείται από το μοντέλο

Model	Number of Predictors	Model Fit statistics	Ljung-Box Q(18)			Number of Outliers
		Stationary R-squared	Statistics	DF	Sig.	
Sales of Men's Clothing-Model_1	2	.948	7.589	18	.984	9

ΑΝΑΛΥΣΗ ΧΡΟΝΟΛΟΓΙΚΩΝ ΣΕΙΡΩΝ

- Ο πίνακας ARIMA εμφανίζει τις τιμές για όλες τις παραμέτρους του μοντέλου,
- Γνωρίζουμε ήδη από τον πίνακα στατιστικών στοιχείων μοντέλο ότι υπάρχουν δύο σημαντικοί παράγοντες πρόβλεψης. Ο πίνακας δείχνει ότι είναι οι μεταβλητές *Number of Catalogs Mailed* και *Number of Phone Lines Open for Ordering*.

					Estimate	SE	t	Sig.
Sales of Men's Clothing-Model_1	Sales of Men's Clothing	No Transformation	Seasonal Difference		1			
	Number of Catalogs Mailed	No Transformation	Numerator	Lag 0	1.549	.071	21.943	.000
			Seasonal Difference		1			
	Number of Phone Lines Open for Ordering	No Transformation	Numerator	Lag 0	315.262	15.298	20.607	.000
			Seasonal Difference		1			