Επίκουρος Καθηγητής **Ιωάννης Παραβάντης**
Τμήμα Διεθνών και Ευρωπαϊκών Σπουδών
ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ
Ιανουάριος 2011

Βασικές μέθοδοι πολυμεταβλητής στατιστικής (multivariate statistics)

1. Αντικείμενο

Στο παρόν φυλλάδιο παρουσιάζονται οι ακόλουθες μέθοδοι πολυμεταβλητής στατιστικής (multivariate statistics):

1. ανάλυση κύριων συνιστωσών (Principal Component Analysis)
2. ανάλυση παραγόντων (Factor Analysis), της οποίας εξετάζεται μόνο η συγγένεια με τη μέθοδο PCA
3. ανάλυση συστάδων (Cluster Analysis)
4. ανάλυση κανονικοποιημένης συσχέτισης (Canonical Correlation Analysis).

Το δείγμα (sample) που θα αναλυθεί σε όλες τις μεθόδους του φυλλαδίου, αποτελεί τμήμα του δείγματος που χρησιμοποιήθηκε από τους Paravantis και Iliadis (2009) σε έρευνα του Ψηφιακού Χάσματος (digital divide) και περιλαμβάνει πληροφορίες για την διείσδυση των τεχνολογιών πληροφορικής και επικοινωνιών (Information and Communication Technologies ή ICT) για 209 χώρες.

Οι μεταβλητές που εξετάζουμε παρουσιάζονται στον κατωτέρω πίνακα:

Πίνακας 1. Μεταβλητές

Αύξων αριθμός	*Όνομα μεταβλητής*	*Εξήγηση*
1	LOGBBSUB	λογάριθμος αριθμού συνδρομητών ευρείας ζώνης (broadband) ανά 100 κατοίκους
2	LOGINTBPP	λογάριθμος ταχύτητας διεθνούς Internet (σε bits ανά λεπτό)
3	LOGINTUSER	λογάριθμος αριθμού χρηστών Internet ανά 100 κατοίκους
4	TELSUB	αριθμός συνδρομητών σταθερής και κινητής τηλεφωνίας ανά 100 κατοίκους
5	LOGSERV	λογάριθμος αριθμού ασφαλών εξυπηρετητών (secure servers) Internet ανά ένα εκατομμύριο κατοίκους
6	LOGTEL	λογάριθμος αριθμού τηλεφωνικών γραμμών ανά 100 κατοίκους
7	LOGPC	λογάριθμος αριθμού υπολογιστών ανά 100 κατοίκων
8	LOGOUTG	λογάριθμος εξερχόμενης τηλεφωνικής κίνησης σε λεπτά της ώρας
9	CELLSUB	αριθμός συνδρομητών κινητής τηλεφωνίας ανά 100 κατοίκους

Κάποιες από τις μεταβλητές του ανωτέρω πίνακα έχουν μετασχηματισθεί λογαριθμικά για να αρθεί η στρεβλότητα τους ώστε να βελτιωθεί η κατανομή τους. Σε μερικές από αυτές, έχει προστεθεί η μονάδα ώστε να μην υπάρχουν μηδενικές τιμές στα πρωτογενή δεδομένα και να μπορούν να εξαχθούν οι λογάριθμοι.

Σε όλες τις ενότητες του φυλλαδίου, χρησιμοποιούμε το στατιστικό πακέτο Stagraphics Centurion (έκδοση 16), που είναι ιδιαίτερα φιλικό και πλήρες ως προς τις πολυμεταβλητές αναλύσεις που περιέχει (http://www.statgraphics.com).

2. Ανάλυση κύριων συνιστωσών (Principal Component Analysis)

Με τη μέθοδο αυτή, όπως και την ανάλυση παραγόντων, γίνεται προσπάθεια να μειωθεί ο αριθμός των μεταβλητών (variables), τις οποίες επιθυμούμε να αναλύσουμε. Η μείωση αυτή επιτυγχάνεται με την εξαγωγή ενός μικρού αριθμού γραμμικών συνδυασμών (linear combinations) των αρχικών μεταβλητών, που ονομάζονται κύριες συνιστώσες (principal components).

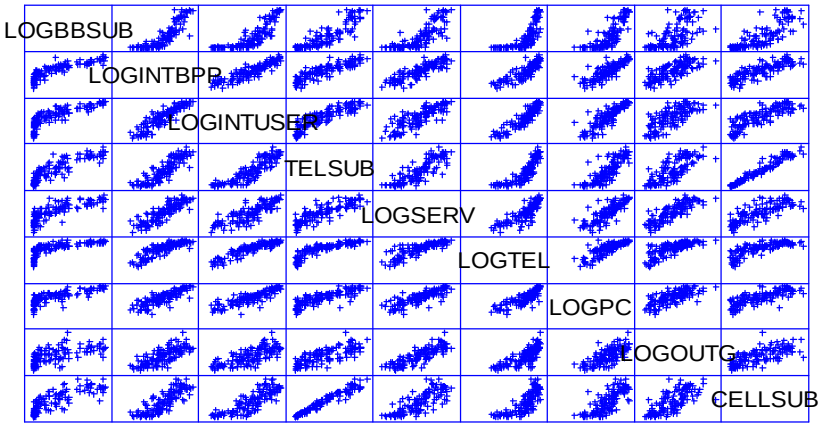
Για να γίνει πιο κατανοητό, στη συγκεκριμένη περίπτωση που έχουμε 9 αρχικές μεταβλητές (X_1, X_2 έως και την X_9), η πρώτη κύρια συνιστώσα (PC_1) που θα εξάγουμε θα είναι της μορφής:

$$PC_1 = a_1X_1 + a_2X_2 + \cdots + a_9X_9$$

όπου a_1, a_2 έως και a_9 είναι οι βαρυτικοί συντελεστές (weights) των αρχικών μεταβλητών στην εξίσωση της πρώτης κύριας συνιστώσας.

Οι μεταβλητές, από τις οποίες θέλουμε να εξάγουμε κύριες συνιστώσες, πρέπει να είναι μεταξύ τους συσχετισμένες. Αυτό μπορούμε να το ελέγξουμε με διαγράμματα διασποράς (scatterplots) και συντελεστές συσχέτισης (correlation coefficients) ανά δυο μεταβλητές. Εάν υπάρχουν μεταβλητές που δεν είναι συσχετισμένες με τις υπόλοιπες, αυτές δεν πρόκειται να συνεισφέρουν στις κύριες συνιστώσες που εξάγουμε.

Στην συγκεκριμένη περίπτωση, όλα τα διαγράμματα διασποράς φαίνονται στο παρακάτω σχήμα:



Σχήμα 1. Πίνακας διαγραμμάτων διασποράς (matrix plot)

Αν και κάποια διαγράμματα διασποράς (για παράδειγμα στην αριστερή κάθετη στήλη ή δεξιά – πάνω-πάνω) δείχνουν ότι κάποιες εκ των συσχετίσεων δεν είναι γραμμικές ούτε εξαιρετικά ισχυρές, θεωρούμε ότι η συσχέτιση των μεταβλητών είναι αρκετά ικανοποιητική για να προχωρήσουμε με την ανάλυση κύριων συνιστωσών. Λάβετε υπόψη σας ότι ο λογαριθμικός μετασχηματισμός κάποιων μεταβλητών έχει βελτιώσει όχι μόνο τις κατανομές τους αλλά και τις συσχετίσεις τους με τις άλλες μεταβλητές – πέραν αυτού, δεν μπορούμε να κάνουμε κάτι άλλο.

Το πλαίσιο διαλόγου (dialog box) που εμφανίζεται από το Statgraphics και δείχνει τις επιλογές που διατίθενται για την ανάλυση κύριων συνιστωσών (PCA) είναι:

Σχήμα 2. Πλαίσιο διαλόγου PCA

Πριν από την ανάλυση κύριων συνιστωσών, τα δεδομένα μετατρέπονται σε τυπικές τιμές (z scores) δηλαδή κανονικοποιούνται (standardized), που όπως θυμάστε από το μάθημα του προηγουμένου εξαμήνου, σημαίνει ότι αφαιρούμε από κάθε τιμή (x) τον μέσος όρος της μεταβλητής ($\overline{x}$) και μετά διαιρούμε με την τυπική απόκλιση της μεταβλητής (s):

$$z = \frac{x - \overline{x}}{s}$$

Παρατηρείστε ότι αυτό το έχουμε ήδη επιλέξει στο κάτω αριστερό μέρος του προηγούμενου παραθύρου διαλόγου (standardize).

Η επόμενη σημαντική απόφαση που πρέπει να πάρουμε πριν εκτελεστεί η PCA είναι το πώς θα χειριστούμε τα ελλιπή δεδομένα, δηλαδή εκείνες τις παρατηρήσεις (observations ή cases) που περιέχουν ελλιπείς τιμές (missing values) σε μια ή περισσότερες μεταβλητές.

Όπως φαίνεται από το παραπάνω πλαίσιο διαλόγου (πάνω αριστερά, εκεί που αναγράφεται Missing Value Treatment), υπάρχουν δυο επιλογές διαχείρισης των ελλιπών δεδομένων:

1. διαγραφή όλων των παρατηρήσεων που έχουν έστω και μια ελλιπή τιμή (listwise) σε κάποια μεταβλητή
2. διαγραφή των παρατηρήσεων που έχουν μια ελλιπή τιμή στη συγκεκριμένη φάση υπολογισμού της στατιστικής μεθόδου που εκτελείται (pairwise), για παράδειγμα εάν υπολογίζαμε τη μέση τιμή των μεταβλητών δεν θα είχε νόημα να διαγράψουμε μια παρατήρηση εάν είχε ελλιπή τιμή σε άλλη μεταβλητή από αυτή της οποίας υπολογίζεται η μέση τιμή!

Η δεύτερη μέθοδος (pairwise) εν γένει καταλήγει στη διαγραφή λιγότερων παρατηρήσεων από την πρώτη (listwise) αλλά είναι εντελώς ακατάλληλη γιατί δεν μπορούμε να είμαστε ποτέ σίγουροι επί ποίου υποσυνόλου έχει υπολογιστεί το κάθε κομμάτι μιας στατιστικής ανάλυσης που ζητήσαμε.

Ως εκ τούτου, στις πολυμεταβλητές μεθόδους πάντα επιλέγουμε τη διαγραφή παρατηρήσεων που περιέχουν ελλιπή δεδομένα με τη μέθοδο listwise.

Ερχόμαστε τώρα στις επιλογές που αφορούν την PCA. Η κυριότερη είναι η επιλογή της μεθόδου εξαγωγής των κύριων συνιστωσών, που την αφήνουμε στην ήδη επιλεγμένη μέθοδο των ιδιοτιμών (eigenvalues). Ας δούμε τα αποτελέσματα της ανάλυσης και εξηγούμε περισσότερα για τις ιδιοτιμές και άλλες λεπτομέρειες της μεθόδου στην πορεία.

Πίνακας 2. Προκαταρτικές πληροφορίες ανάλυσης κύριων συνιστωσών

```
Principal Components Analysis
Data variables:
     LOGBBSUB
     LOGINTBPP
     LOGINTUSER
     TELSUB
     LOGSERV
     LOGTEL
     LOGPC
     LOGOUTG
     CELLSUB

Data input: observations
Number of complete cases: 124
Missing value treatment: listwise
Standardized: yes
```

Στον Πίνακα 2 επιβεβαιώνεται

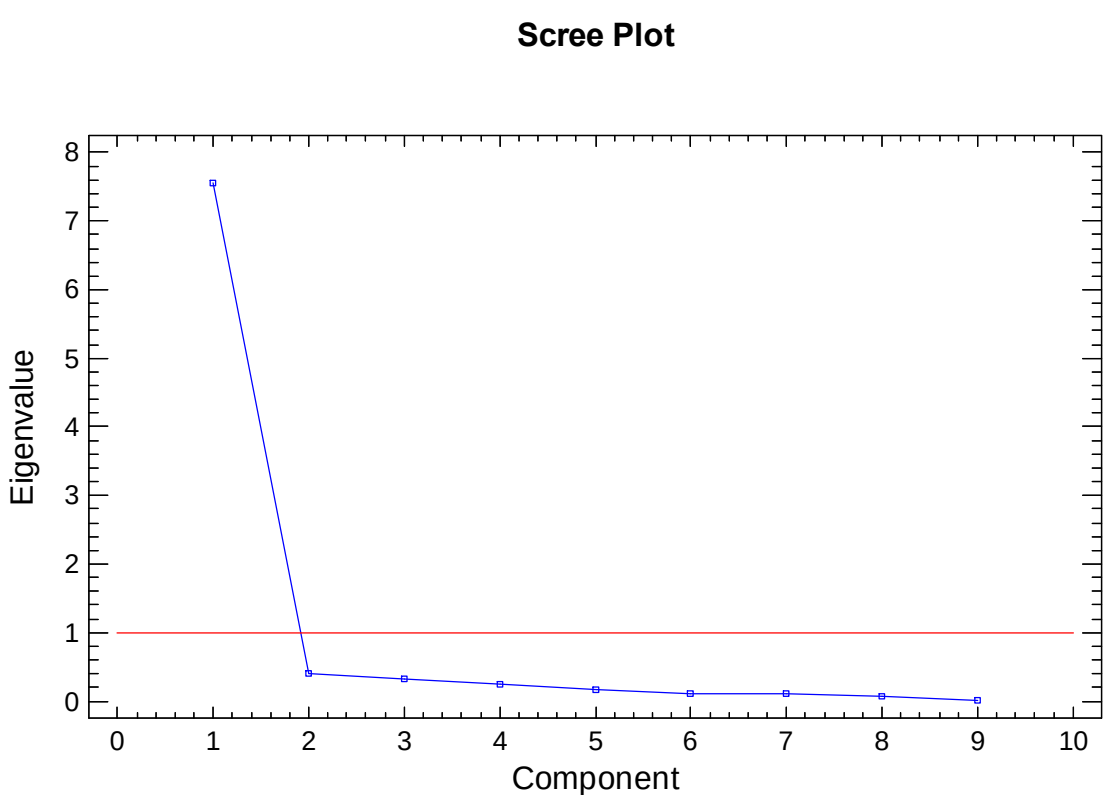
- η επιλογή της μεθόδου PCA
- οι μεταβλητές που έχουν εισαχθεί για ανάλυση
- το ότι οι ελλιπείς παρατηρήσεις διαγράφηκαν με τη μέθοδο listwise, γεγονός που μας περιόρισε από τις 209 παρατηρήσεις (δηλαδή χώρες) που είχαμε αρχικά, σε 124 χώρες
- και το ότι τυποποιήθηκαν οι τιμές πριν από την ανάλυση (Standardized: yes).

Το κύριο τμήμα της εξόδου (output) της ανάλυσης κύριων συνιστωσών είναι το εξής:

Πίνακας 3. Ιδιοτιμές κύριων συνιστωσών

Component Number	*Eigenvalue*	*Percent of Variance*	*Cumulative Percentage*
1	7.54703	83.856	83.856
2	0.409759	4.553	88.409
3	0.329638	3.663	92.071
4	0.246342	2.737	94.809
5	0.165991	1.844	96.653
6	0.118244	1.314	97.967
7	0.102871	1.143	99.110
8	0.0678222	0.754	99.863
9	0.0123025	0.137	100.000

Η ανάλυση συνοδεύεται και από διάγραμμα Scree, όπου στον άξονα Ψ βρίσκονται οι ιδιοτιμές (eigenvalues) και στον άξονα Χ βρίσκονται οι κύριες συνιστώσες (principal components):



Σχήμα 3. Διάγραμμα Scree

Για να αποφασίσουμε πόσες κύριες συνιστώσες (principal components) θα κρατήσουμε, χρησιμοποιούμε δυο κριτήρια:

1. Το κριτήριο του Kaiser (1960), σύμφωνα με το οποίο κρατάμε όσες κύριες συνιστώσες έχουν ιδιοτιμή μεγαλύτερη της μονάδας. Αξίζει να αναφέρουμε ότι κύριες συνιστώσες με ιδιοτιμές μικρότερες της μονάδες εξηγούν μικρότερη διασποράς των αρχικών δεδομένων από μια μεταβλητή, συνεπώς δεν έχει έννοια να τις διατηρήσουμε!
2. Το κριτήριο του Cattell (1966), σύμφωνα με το οποίο κοιτάμε το διάγραμμα Scree και κρατάμε εκείνες τις κύριες συνιστώσες που βρίσκονται πριν από την αλλαγή κλίσεως της τεθλασμένης γραμμής.

Ας εφαρμόσουμε το κάθε κριτήριο για να δούμε πόσες κύριες συνιστώσες κρατάμε στην περίπτωση της ανάλυσής μας.

Σύμφωνα με τον Πίνακα 3, μόνο η πρώτη κύρια συνιστώσα (component) έχει ιδιοτιμή μεγαλύτερη της μονάδας και, ως εκ τούτου, πληροί το κριτήριο Kaiser: η ιδιοτιμή αυτή ισούται με 7.54703 (δηλαδή είναι πολύ μεγαλύτερη της μονάδας), πράγμα που σημαίνει ότι η πρώτη κύρια συνιστώσα εξηγεί την διασπορά των αρχικών δεδομένων που προκαλείται από περίπου 7.5 αρχικές μεταβλητές! Πράγματι, σύμφωνα με τις δυο επόμενες στήλες του πίνακα που δείχνουν το ποσοστό της διασποράς που εξηγείται από (α) κάθε μια κύρια συνιστώσα και (β) αθροιστικά από όλες τις συνιστώσες μέχρι και την τρέχουσα, η πρώτη κύρια συνιστώσα εξηγεί το 83.856% του συνόλου της διασποράς που παρατηρείται στις μεταβλητές που εξετάζουμε (παρατηρείστε ότι $\frac{7.54703}{9} = 0.83856 = 83.856\%$).

Επιβεβαιώστε ότι και η εφαρμογή του κριτηρίου του Cattell στο διάγραμμα Scree, μας δίνει μια κύρια συνιστώσα (δείτε το κομμάτι της μπλε γραμμής που βρίσκεται αριστερά και πάνω-πάνω), πριν από το σημείο αλλαγής της καμπύλης (που βρίσκεται στο βύθισμα της γραμμής κάτω από την οριζόντια κόκκινη γραμμή). Μάλιστα, αυτή η οριζόντια κόκκινη γραμμή μας δείχνει γραφικά την εφαρμογή του κριτηρίου του Kaiser, δηλαδή δείχνει τη θέση του κάθετου άξονα στον οποίο έχουμε την τιμή της μονάδας για τις ιδιοτιμές.

Για να δούμε την εξίσωση της πρώτης κύριας συνιστώσας, που μόλις υπολογίσαμε:

Πίνακας 4. Βαρυτικοί συντελεστές πρώτης κύριας συνιστώσας

	Component 1
LOGBBSUB	0.325403
LOGINTBPP	0.342834
LOGINTUSER	0.34349
TELSUB	0.349357
LOGSERV	0.33715
LOGTEL	0.331005
LOGPC	0.329754
LOGOUTG	0.305296
CELLSUB	0.333712

Για να είμαστε σίγουροι ότι καταλαβαίνουμε αυτό τον πίνακα, γράφουμε κατωτέρω την εξίσωση της πρώτης κύριας συνιστώσας, PC1, βασιζόμενοι στην ανωτέρω πίνακα:

PC1 = 0.325403×LOGBBSUB + 0.342834×LOGINTBPP + 0.34349×LOGINTUSER +
0.349357×TELSUB + 0.33715×LOGSERV + 0.331005×LOGTEL + 0.329754×LOGPC +
0.305296×LOGOUTG + 0.333712×CELLSUB

Πράγματι λοιπόν, μια κύρια συνιστώσα αποτελεί γραμμικό συνδυασμό των τιμών των μεταβλητών στις οποίες βασίζεται. Μην ξεχνάμε όμως ότι οι τιμές αυτές είναι κανονικοποιημένες.

Παρατηρούμε ότι όλες οι μεταβλητές εμφανίζονται με θετικό πρόσημα και παρόμοιες τιμές βαρυτικών συντελεστών, στην εξίσωση αυτή της πρώτης κύριας συνιστώσας. Αυτό σημαίνει ότι όλες οι επιμέρους μεταβλητές του Πίνακα 1, που μετράνε διαφορετικά χαρακτηριστικά της διείσδυσης ψηφιακών τεχνολογιών στις κοινωνίες των κρατών που εξετάστηκαν, συνεισφέρουν ισόποσα στην διαμόρφωση της κύριας συνιστώσας, που αποτελεί ένα είδος δείκτη ψηφιακής διείσδυσης. Από δω και πέρα, αντί να αναλύουμε τις 9 επιμέρους μεταβλητές, μπορούμε να αναλύσουμε την μια κύρια συνιστώσα.

Συνοψίζοντας και καταλήγοντας, από τις 9 μεταβλητές που εξετάσαμε εξάγουμε μια μόνο κύρια συνιστώσα, η οποία εξηγεί την διασπορά των σημείων που προκαλείται από περίπου 7.5 από τις 9 αρχικές μεταβλητές.

Στην επόμενη ενότητα προβαίνουμε σε σύντομο σχολιασμό των ομοιοτήτων και διαφορών της μεθόδου ανάλυσης παραγόντων (Factor Analysis) με τη μέθοδο των κύριων συνιστωσών (Principal Component Analysis).

3. Ανάλυση παραγόντων (Factor Analysis)

απλή αναφορά ομοιοτήτων και διαφορών της μεθόδου με τη μέθοδο PCA

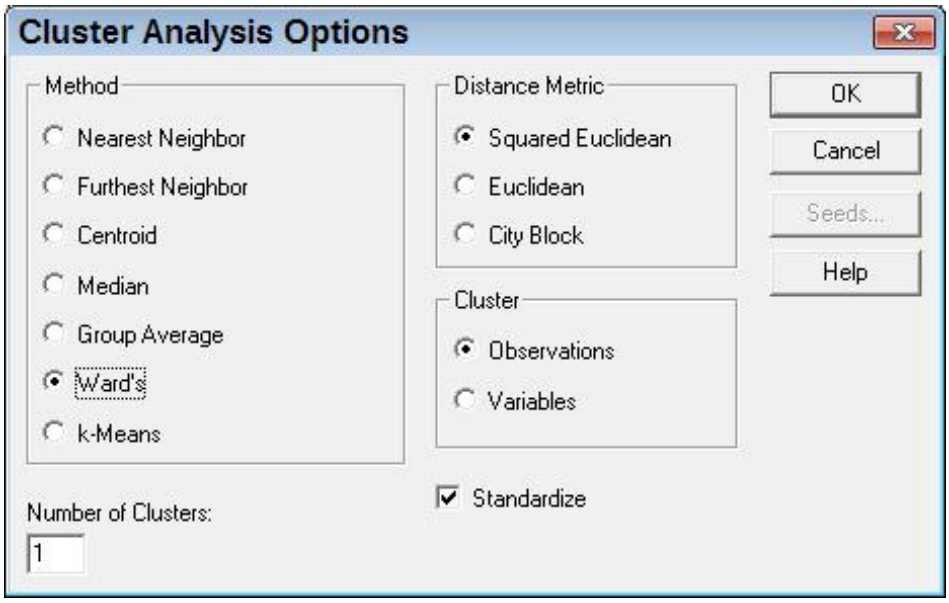
4. Ανάλυση συστάδων (Cluster Analysis)

Ο στόχος της ανάλυσης συστάδων είναι διαφορετικός από αυτόν της ανάλυσης κύριων συνιστωσών: με την ανάλυση συστάδων, ομαδοποιούνται οι παρατηρήσεις (observations ή cases) αντί να ομαδοποιούνται (σε γραμμικούς συνδυασμούς) οι μεταβλητές.

Ειδικά στην περίπτωση του ψηφιακού χάσματος, ενδιαφέρει να δούμε κατά πόσον οι χώρες που έχουμε στο δείγμα μας ομαδοποιούνται σε δυο συστάδες, που η μια θα περιείχε αναπτυγμένες χώρες με μεγάλη διείσδυση ψηφιακών τεχνολογιών ενώ η άλλη αναπτυσσόμενες χώρες με μικρή διείσδυση ψηφιακών τεχνολογιών.

Επιχειρούμε λοιπόν να ομαδοποιήσουμε τις ίδιες μεταβλητές που αναλύσαμε στην ενότητα των κύριων συνιστωσών.

Το πλαίσιο διαλόγου του Statgraphics για την ανάλυση συστάδων φαίνεται στο παρακάτω σχήμα:



Σχήμα 4. Πλαίσιο διαλόγου ανάλυσης συστάδων

Καταρχήν παρατηρήστε ότι δεν υπάρχει επιλογή για την διαγραφή των ελλιπών δεδομένων (pairwise ή listwise), όπως στην περίπτωση της ανάλυσης κύριων συνιστωσών. Αυτό γιατί λόγω της φύσης της μεθόδου ανάλυσης συστάδων, οι παρατηρήσεις που έχουν έστω και ένα ελλιπές δεδομένο διαγράφονται αυτόματα (listwise).

Παρατηρήστε ότι στο κάτω δεξιά μέρος του πλαισίου είναι επιλεγμένη η επιλογή Standardize (τυποποίηση ή κανονικοποίηση των δεδομένων). Πράγματι αποτελεί σωστή πρακτική πριν από την ανάλυση συστάδων να τυποποιούμε τα δεδομένα μας ώστε να μην μας οδηγήσουν σε παράδοξες επιλογές συσταδοποίησης ενδεχόμενες μεγάλες διαφορές ανάμεσα στις τιμές διαφορετικών μεταβλητών.

Πάνω από το Standardize υπάρχει μια επιλογή για την συσταδοποίηση παρατηρήσεων (observations) ή μεταβλητών (variables). Αγνοήστε αυτή την επιλογή – πάντα θα διαλέγουμε συσταδοποίηση παρατηρήσεων.

Τέλος, κάτω αριστερά μπορούμε να διαλέξουμε τον αριθμό των συστάδων (number of clusters). Αρχικά, αφήστε αυτό το νούμερο στο ένα (1) για να δούμε όλο το δενδρόγραμμα, και μετά αποφασίζουμε για τον βέλτιστο αριθμό συστάδων.

Πάμε τώρα στην ουσία της ανάλυσης συστάδων: τις επιλογές μεθόδου σύνδεσης (linkage method) και μέτρου απόστασης (distance metric).

Να θυμάστε ότι υπάρχουν συγκεκριμένοι συνδυασμοί μεθόδων σύνδεσης και μέτρων απόστασης, που η βιβλιογραφία θεωρεί κατάλληλους. Στα πλαίσια του μαθήματος μας, θα χρησιμοποιήσουμε ένα δοκιμασμένο τέτοιο συνδυασμό: τη σύνδεση των σημείων με τη μέθοδο του Ward με μέτρο απόστασης το τετράγωνο της Ευκλείδειας απόστασης (squared Euclidean). Η μέθοδος Ward τείνει

να σχηματίζει συστάδες με τον ίδιο αριθμό παρατηρήσεων ενώ η τετραγωνισμένη Ευκλείδεια απόσταση τείνει να «τιμωρεί» τα σημεία που είναι μακρινά και ασυνήθιστα (outliers).

Το πρώτο μέρος της εκτύπωσης περιέχει τα ακόλουθα:

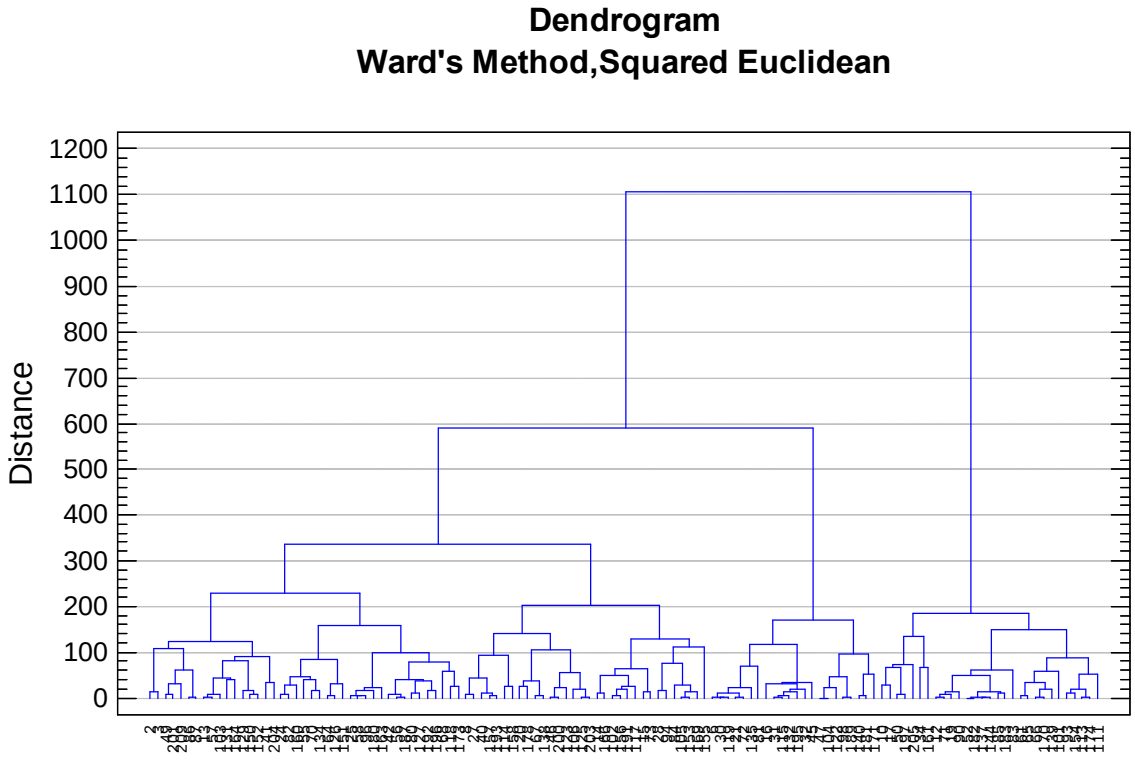
Πίνακας 5. Προκαταρκτικές πληροφορίες ανάλυσης συστάδων

```
Cluster Analysis
Data variables:
     LOGBBSUB
     LOGINTBPP
     LOGINTUSER
     TELSUB
     LOGSERV
     LOGTEL
     LOGPC
     LOGOUTG
     CELLSUB

Number of complete cases: 124
Clustering Method: Ward's
Distance Metric: Squared Euclidean
Clustering: observations
Standardized: yes
```

Επιβεβαιώνουμε τις μεταβλητές που αναλύσαμε, βλέπουμε ότι και πάλι έχουμε 124 παρατηρήσεις (μετά την διαγραφή των ελλιπών δεδομένων με τη μέθοδο listwise), επιβεβαιώνουμε ότι χρησιμοποιήσουμε τη μέθοδο Ward με το τετράγωνο της Ευκλείδειας απόστασης (Squared Euclidean) καθώς και ότι τυποποιήσαμε τις τιμές πριν από την ανάλυση (Standardized: yes).

Η υπόλοιπη εκτύπωση της ανάλυσης δεν έχει μεγάλη σημασία γιατί είχαμε επιλέξει μόνο μια συστάδα, προκειμένου να αξιολογήσουμε το δενδρόγραμμα:



Σχήμα 5. Δενδρόγραμμα (dendrogram)

Για να αποφασίσουμε ποιος είναι ο βέλτιστος αριθμός συστάδων στις οποίες θα εντάξουμε τα δεδομένα μας, κοιτάμε το δενδρόγραμμα και προσπαθούμε να διακρίνουμε σε πόσες συστάδες παρατηρείται η μεγαλύτερη αύξηση της απόστασης (distance, στον κάθετο άξονα) χωρίς να αλλάξει ο αριθμός των συστάδων.

Στο συγκεκριμένο δενδρόγραμμα, δείτε ότι από την τιμή 600 μέχρι την τιμή 1100 (περίπου), η ύπαρξη δυο κάθετων μπλε γραμμών μας δείχνει ότι εκεί μπορούμε να χωρίσουμε τα δεδομένα μας σε δυο συστάδες. Αν αντίθετα αποφασίζαμε να βασιστούμε στις τρεις κάθετες μπλε γραμμές που υπάρχουν από τη θέση 340 (περίπου) έως τη θέση 600 (περίπου), θα χωρίζαμε τα δεδομένα μας σε τρεις συστάδες, αυτό όμως δεν θα ήταν βέλτιστο γιατί η απόσταση από το 340 έως το 600 (=260) είναι μικρότερη από την απόσταση από το 600 έως το 1100 (=500)!

Αποφασίζουμε λοιπόν να διαλέξουμε δυο συστάδες και ξανατρέχουμε την ανάλυση συστάδων για να πάρουμε την εξής εκτύπωση (το αρχικό κομμάτι είναι ίδιο και δεν αναπαράγεται):

Πίνακας 6. Εκτύπωση ανάλυσης συστάδων

Cluster Summary

Cluster	Members	Percent
1	95	76.61
2	29	23.39

Centroids

Cluster	LOGBBSUB	LOGINTBPP	LOGINTUSER	TELSUB	LOGSERV	LOGTEL
1	0.21305	1.64702	0.968713	51.9748	0.413419	0.875458
2	1.14134	3.5474	1.71885	145.202	2.22857	1.67746

Cluster	LOGPC	LOGOUTG	CELLSUB
1	0.64818	1.13796	38.3006
2	1.59292	2.18026	96.0731

Από τα περιληπτικά στοιχεία της ανάλυσης συστάδων (Cluster Summary), βλέπουμε ότι οι 124 χώρες χωρίστηκαν σε δυο συστάδες, μια εκ των οποίων περιέχει 95 χώρες ενώ η άλλη 29 χώρες.

Στο κάτω μέρος της εκτύπωσης βλέπουμε τις τιμές των σημείων που είναι στα κέντρα των συστάδων (Centroids) για κάθε μεταβλητή. Παρατηρήστε ότι στο κέντρο της δεύτερης συστάδας, όλες οι μεταβλητές έχουν μεγαλύτερες τιμές από το κέντρο της πρώτης συστάδας, για παράδειγμα ο αριθμός συνδρομητών σταθερής και κινητής τηλεφωνίας είναι 145.202 στο κέντρο της δεύτερης συστάδας και 51.9748 στο κέντρο της πρώτης.

Πετύχαμε λοιπόν με την ανάλυση συστάδων να χωρίσουμε τις χώρες σε δυο συστάδες: 29 αναπτυγμένες ψηφιακά χώρες και 95 αναπτυσσόμενες! Η διάκριση των χωρών στις δυο αυτές συστάδες αναδεικνύει το ψηφιακό χάσμα σε ποσοτική και συγκεκριμένη βάση.

5. Ανάλυση κανονικοποιημένης συσχέτισης (Canonical Correlation Analysis)

Η ενότητα αυτή είναι **εκτός ύλης** για την εξεταστική περίοδο του **Φεβρουαρίου 2011**.

Βιβλιογραφικές αναφορές (references)

Cattell, R.B. (1966): «The Scree Test for the Number of Factors», *Multivariate Behavioral Research*, Vol. 1, pp. 245–276.

Kaiser, H.F. (1960): «The Application of Electronic Computer to Factor Analysis», *Educational and Psychological Measurement*, Vol.20, pp. 141–151.

Manly, B.F.J. (2005): Multivariate Statistical Methods: A Primer, Chapman and Hall/CRC.

Paravantis, J. and M. Iliadis (2009): *A Multivariate Cross-National Empirical Analysis of the Digital Divide*, working paper.

Stevens, J., (2002): *Applied Multivariate Statistics for the Social Sciences,* 4th edition, Lawrence Elrbaum Associates.