Πανεπιστήμιο Θεσσαλίας, Πολυτεχνική Σχολή
Τμήμα Μηχανικών Χωροταξίας, Πολεοδομίας και Περιφερειακής Ανάπτυξης

Δημιουργία, διαχείριση και οργάνωση βάσεων χωρικών δεδομένων
Η περίπτωση των δημοσίων υπαλλήλων της Ελλάδος

Establishment, management and organization of spatial databases
The case of public officials of Greece

Διπλωματική εργασία του
Αναστάσιου Κουτσούκου

Επιβλέπουσα καθηγήτρια: Μαρί Νοέλ Ντυκέν

Εξεταστική Επιτροπή:
Μαρί Νοέλ Ντυκέν (Αναπληρώτρια Καθηγήτρια)
Βύρων Κοτζαμάνης (Καθηγητής)
Δημήτριος Καλλιώρας (Επίκουρος Καθηγητής)

Βόλος, 2013

ΠΕΡΙΛΗΨΗ

Στην παρούσα μελέτη πραγματοποιείται η έρευνα για την δημιουργία, διαχείριση και οργάνωση βάσεων χωρικών δεδομένων. Αρχικά μελετώνται οι βάσεις δεδομένων γενικού στατιστικού περιεχομένου. Έπειτα, με την μελέτη περίπτωσης, αναπτύσσεται η εμπειρική προσέγγιση και η δημιουργία των βάσεων χωρικών δεδομένων.

Αναλυτικότερα στο πρώτο τμήμα παρουσιάζονται βασικές έννοιες και ορισμοί απαραίτητοι για την διεξαγωγή της μελέτης. Στο δεύτερο τμήμα εμπεριέχεται το θεωρητικό υπόβαθρο για τις στατιστικές βάσεις και τα δεδομένα τους. Στο τρίτο τμήμα πραγματοποιείται εφαρμογή του θεωρητικού υποβάθρου σε πραγματική βάση δεδομένων (η περίπτωση των δημοσίων υπαλλήλων της Ελλάδος) και παράγονται μέσω εμπειρικών πρακτικών το προφίλ των δημοσίων υπαλλήλων, καθώς και νέες πρωτότυπες χωρικές βάσεις και μεταβλητές. Τέλος, παρατίθενται τα συμπεράσματα και ο επίλογος.

ABSTRACT

During the following study, research is carried out for the establishment, management and organization of spatial databases. Primarily, databases of general statistical content are studied. Afterwards, during the case study, an empirical approach is developed for spatial databases. Also in this point, the creation of several spatial databases is complete.

Specifically, during the first section, basic concepts and definitions necessary for the conduct of the study are presented. The second part includes the theoretical background of statistical bases and their data. During the third section, the theoretical background is applied to the real database (the case of public officials of Greece). Also through empirical practice the profile of civil servants is completed and new original spatial databases and variables are produced. Finally the conclusions and the epilogue are presented.

ΠΕΡΙΕΧΟΜΕΝΑ

ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

ΚΑΤΑΛΟΓΟΣ ΔΙΑΓΡΑΜΜΑΤΩΝ

ΚΑΤΑΛΟΓΟΣ ΧΑΡΤΩΝ

ΕΥΧΑΡΙΣΤΙΕΣ

Θα ήθελα να ευχαριστήσω όλα τα άτομα που με βοήθησαν να ολοκληρώσω όχι μόνο την παρούσα εργασία, αλλά και τις σπουδές μου.

Αρχικά, θα ήθελα να ευχαριστήσω την οικογένειά μου και τους φίλους μου από την Σαλαμίνα για την πολύτιμη βοήθειά τους.

Επίσης, θα ήθελα να ευχαριστήσω την επιβλέπουσα καθηγήτριά μου, κυρία Μαρί Νοέλ Ντυκέν, για την βοήθειά και την καθοδήγησή της, καθόλη τη διάρκεια της διπλωματικής μου εργασίας.

Τέλος θα ήθελα να πω ένα μεγάλο ευχαριστώ στο παρεάκι μου από τον Βόλο και συγκεκριμένα στον Κωνσταντίνο Νικολάου που μέχρι και την τελευταία στιγμή με στήριξε όταν τον είχα ανάγκη.

ΕΙΣΑΓΩΓΗ

Στην σύγχρονη εποχή, η πληροφορία αποτελεί κυρίαρχο ρόλο στα πλαίσια της ανάπτυξης και της εργασίας. Όμως η πληθώρα στοιχείων και δεδομένων πρέπει να αντιμετωπίζεται προσεκτικά ώστε να ληφθούν ορθά συμπεράσματα για τις μελέτες των εκάστοτε φαινομένων. Με την βοήθεια της (περιγραφικής) στατιστικής μπορούμε να επεξεργαστούμε βάσεις δεδομένων ώστε να εξάγουμε νέα δεδομένα, νέες βάσης ή να παράγουμε μοντέλα και συστηματικές πρακτικές, για να επιτευχθεί η αποδοτικότερη μορφή των αναγκαίων στοιχείων της έρευνας. Κατά αυτόν τον τρόπο επιτυγχάνεται αντιπροσωπευτικότερη ερμηνεία των φαινομένων, ενισχύοντας τις εκάστοτε αποφάσεις μας ως σχεδιαστές και προγραμματιστές.

Ως αναλυτές μπορούμε να χρησιμοποιήσουμε και να δημιουργήσουμε βάσεις δεδομένων κατάλληλες, τόσο ως προς τις ερωτήσεις, αλλά και όσο προς τα δεδομένα του εκάστοτε κλάδου που μελετάμε. Έτσι μπορούμε να αντλήσουμε από τις βάσεις/μοντέλα/δεδομένα σχετικές πληροφορίες και να επεκτείνουμε το φάσμα ανάλυσης σε όποιο αντικείμενο μελετάται αντίστοιχα.

Εδώ όμως παρουσιάζεται ένα χάσμα μεταξύ του πεδίου της στατιστικής/οικονομετρίας/περιγραφικής στατιστικής, με το πεδίο των επιστημών του χώρου. Ο αναλυτής του χώρου χρειάζεται την «χωρική υπόσταση» των δεομένων ώστε να τα χρησιμοποιήσει στα πλαίσια των ερευνών του, η οποία πολλές φορές είτε βρίσκεται ως δευτερεύοντα δεδομένα εντός της βάσης, είτε η «αρχιτεκτονική» της βάσης δεν τον βοηθά στην εκμετάλλευσή της. Οπότε ο κάθε αναλυτής καλείται να «ξεπεράσει» το χάσμα αυτό με την εμπειρία και την εντριβή του με τα εκάστοτε δεδομένα. Αυτό όμως αποτελεί πιο εξειδικευμένο κλάδο της περιγραφικής στατιστικής, όπου κυριαρχεί η εμπειρική προσέγγιση παρά η καθαρή θεωρία.

Με γνώμονα τα παραπάνω, στην συνέχεια της παρούσας μελέτης ερευνάται η δυνατότητα χρήσης βάσεων δεδομένων για την παραγωγή νέων χωρικών βάσεων. Δηλαδή αναλύονται υπάρχουσες θεωρητικές προσεγγίσεις και εμπειρικές τεχνικές, ενώ μετέπειτα αναπτύσσονται συστηματικές μέθοδοι για την επεξεργασία, διαχείριση και παρουσίασή τους.

1. ΕΙΣΑΓΩΓΙΚΕΣ ΕΝΝΟΙΕΣ ΚΑΙ ΟΡΙΣΜΟΙ

Για την καλύτερη κατανόηση του υπολοίπου της παρούσας εργασίας, όπως επίσης και για την ξεκάθαρη ορολογία των βασικών εννοιών, τίθεται απαραίτητο να παρουσιαστούν αυτοί οι παρακάτω ορισμοί, οι οποίοι πλαισιώνουν το σύνολο της μελέτης.

1.1 ΣΤΑΤΙΣΤΙΚΗ

Για την Στατιστική επιστήμη έχουν δοθεί διάφοροι ορισμοί, οι οποίοι εξαρτώνται κάθε φορά από την θεωρητική προσέγγιση του κάθε μελετητή, ή από τους σκοπούς ανάλυσης του κάθε μελετητή. Βέβαια σε κάθε περίπτωση, η στατιστική απασχολείται με: α) την συγκέντρωση, επεξεργασία, παρουσίαση, αξιολόγηση και εξαγωγή δεδομένων και συμπερασμάτων, β) αβέβαια φαινόμενα και καταστάσεις του πραγματικού κόσμου από όπου προέρχονται τα παραπάνω δεδομένα γ) τα παραγόμενα μοντέλα και προτάσεις, μετά την εξέταση των προαναφερθέντων φαινομένων, καθώς και τέλος, δ) τα συμπεράσματα από την ολοκλήρωση της έρευνας, τα οποία με την σειρά τους αποτελούν την βάση για την λήψη νέων αποφάσεων και σχεδιασμού. (Κιντής, 1999) (Παπαδήμας Ο. και Κοΐλιας Χ., 1998)

1.1.1 ΑΝΤΙΚΕΙΜΕΝΟ

Παρά την κυρίαρχη χρήση της σε θεωρητικό επίπεδο η χρήση μαθηματικών μοντέλων είναι αναγκαία και για την επίλυση προβλημάτων στο επίπεδο εφαρμογών. Χρησιμοποιείται στο πλαίσιο όλων σχεδόν των άλλων γνωστικών πεδίων (ιατρική, χρηματοοικονομικά, χωροταξία, πολεοδομία, περιβάλλον κ.α.) και σύμφωνα με τον «σχετικό ορισμό» που δόθηκε στην αρχή του παρόντος κεφαλαίου, στο αντικείμενό της εμπεριέχονται τα εξής:

1. **Ανάπτυξη τεχνικών για συγκέντρωση, επεξεργασία και παρουσίαση πληροφοριών,** προερχόμενων από καταγραφές πραγματικών φαινομένων και συσχετιζόμενων με την συμπεριφορά και συσχέτιση αυτών.
2. **Διαμόρφωση μεθοδολογίας για ανάλυση των πληροφοριών.** Σημειώνεται πλήθος μεθοδολογιών, όπως για εκτίμηση παραμέτρων πληθυσμού, ή για έλεγχο υποθέσεων.
3. **Εξαγωγή συμπερασμάτων και λήψη αποφάσεων / πρακτικός σχεδιασμός και πολιτικές,** τα οποία προέρχονται από την ταύτιση ή όχι των αποτελεσμάτων των εμπειρικών δεδομένων με τις προβλέψεις των θεωρητικών μοντέλων.

(Κιντής, 1999) (Παπαδήμας Ο. και Κοΐλιας Χ., 1998)

Λόγω των παραπάνω, είναι λογικό η Στατιστική να αποτελεί ένα εργαλείο στα χέρια των αναλυτών το οποίο προσφέρει προσδιορίσιμης αξιοπιστίας συμπεράσματα για τους παράγοντες, την συμπεριφορά αλλά και τα ίδια τα φαινόμενα, τα οποία χρησιμοποιούνται στην εφαρμοσμένη κυρίως έρευνα, αλλά και στο πεδίο λήψεων αποφάσεων (π.χ. δημόσιοι οργανισμοί). (Κιντής, 1999)

Εδώ πρέπει να σημειωθεί ότι πλέον λόγω των προαναφερθέντων, οι αναλυτές από διάφορους κλάδους και τομείς στις μέρες μας εξοικειώνονται με την βασική στατιστική μεθοδολογία και τις τεχνικές ανάλυσης των δεδομένων, και είναι εφικτό να αναπτύσσονται εξειδικευμένες προσεγγίσεις και τεχνικές για την αντιμετώπιση προβλημάτων που σημειώνονται στην ειδικότητά τους. (Κιντής, 1999) Άλλωστε αυτό ερευνάται και στην παρούσα εργασία, όπου στην συνέχειά της, πραγματοποιείται προσπάθεια για την χρήση της στατιστικής και της οικονομετρίας για την μελέτη κοινωνικο-χωρικών στοιχείων, συνδυάζοντας έτσι το περιεχόμενο της χωροταξίας με την στατιστική έρευνα/μεθοδολογία.

1.1.2 ΜΕΘΟΔΟΛΟΓΙΕΣ

Όπως σημειώθηκε προηγουμένως, η χρήση της στατιστικής συσχετίζεται με ένα ευρύ φάσμα αντικειμένων και γνωστικών πεδίων. Για αυτό τον λόγο, η μεθοδολογική προσέγγιση της στατιστικής διερεύνησης (επιμέρους φαινομένων), διαχωρίζεται και να κατηγοριοποιείται σύμφωνα με τον επιδιωκόμενο σκοπό και το είδος των εκάστοτε πληροφοριών οι οποίες πρόκειται να αναλυθούν. Ο διαχωρισμός των υπάρχων μεθοδολογιών περιλαμβάνει:

- Περιγραφική στατιστική,
- Επαγωγική στατιστική - Στατιστική συμπερασματολογία,
- Ανάλυση παλινδρόμησης και διακύμανσης,
- Στοχαστική ανάλυση,
- Μπεϋζιανή ανάλυση,
- Πολυμεταβλητή ανάλυση – Ανάλυση κατηγορικών δεδομένων,
- Μη- παραμετρική στατιστική. (Κιντής, 1999) (Ζήμερας, 2003)

1.1.3 ΠΕΡΙΓΡΑΦΙΚΗ ΣΤΑΤΙΣΤΙΚΗ

Κύριο εργαλείο επεξεργασίας, και αντικείμενο έρευνας στην παρούσα μελέτη είναι η Περιγραφική Στατιστική. Σκοπός της είναι η δημιουργία τεχνικών για άθροιση, σύνοψη, επεξεργασία, αποθήκευση και παρουσίαση δεδομένων (πρωτογενές στατιστικό υλικό). Ειδικότερα αποτελεί ένα στατιστικό εργαλείο με σκοπό την συγκέντρωση και παρουσίαση πρωτογενών δεδομένων σε κατανοητή μορφή για μετέπειτα χρήση. Γίνεται με την χρήση πινάκων (συχνοτήτων, διπλής εισόδου), γραφημάτων (ραβδογράμματα, θηκογράμματα, διασποράς), και στατιστικών μέτρων (μέτρα κεντρικής τάσης, μέτρα κύμανσης, και μεταβλητότητας). Υπάρχει επίσης η δυνατότητα για εξαγωγή συμπερασμάτων περιγραφικού χαρακτήρα (για το ευρύτερο τμήμα του πληθυσμού) αλλά δεν παράγει προτάσεις/συμπεράσματα για γενικότερη εφαρμογή ή προβλέψεις για μελλοντικές πορείες των φαινομένων υπό έρευνα. (Κιντής, 1999) (Ζήμερας, 2003)

Η διαδικασία έρευνας των φαινομένων (οικονομικών, κοινωνικών, χωροταξικών κ.α.) ξεκινά με την συγκέντρωση και επεξεργασία μεγάλου αριθμού δεδομένων/παρατηρήσεων που αφορούν στην εξέλιξή τους (χρονικά / χωρικά) και ολοκληρώνεται όταν επιτευχθεί η μεγαλύτερη δυνατή συμπύκνωση της πληροφορίας σε δείκτες. Οι δείκτες αυτοί χαρακτηρίζονται ως ποσοτικής φύσεως (απλοί αριθμοί) και εξηγούν περιεκτικά τον κύριο όγκο πληροφορίας που αποσπάστηκε από το πληθυσμό ή το δείγμα της έρευνας. (Κιντής, 1999)

1.2 ΟΙΚΟΝΟΜΕΤΡΙΑ

Αρχικά, μπορεί να ειπωθεί πως α) ο έλεγχος της θεωρητικής προσέγγισης εμπειρικά, β) η άσκηση οικονομικής / χωρικής / κοινωνικής πολιτικής, γ) ο προγραμματισμός και ο σχεδιασμός, αλλά και δ) η αξιοποίηση των δεδομένων από τον υπαρκτό κόσμο (για την διατύπωση σχέσεων μεταξύ των αντίστοιχων μεταβλητών όπως και η κατασκευή προβλέψεων), προϋποθέτουν, την «ποσοποποίηση» των κοινωνικοοικονομικών σχέσεων και των αλληλεξαρτήσεων/αλληλεπιδράσεων. Για να επιτευχθεί αυτή η ενέργεια ποσοποποίησης, πρέπει συνδυαστικά ο μελετητής να κάνει χρήση της στατιστικής, της οικονομικής θεωρίας και των μαθηματικών. Συνδυάζοντας τις βασικές 3 επιστήμες, πραγματοποιήθηκε η ανάπτυξη μεθόδων και τεχνικών, κρίσιμων για την αντιμετώπιση έντονων προβλημάτων σε ότι αφορά την εκτίμηση ποσοτικών σχέσεων στο χώρο των οικονομικών, χωρικών και κοινωνικών επιστημών, παράγοντας έτσι την Οικονομετρία. (Κιντής, 2010) (Τραχανάς, 2003)

1.2.1 ΚΛΑΔΟΙ

Ως κλάδοι της οικονομετρίας σημειώνονται δύο, αυτός της Θεωρητικής Οικονομετρίας και αυτός της Εφαρμοσμένης Οικονομετρίας. Ο θεωρητικός κλάδος ασχολείται κυρίως με την ανάπτυξη μεθόδων και τεχνικών. Προέρχεται από τις προσπάθειες ερευνητών να προσαρμόσουν μεθόδους της στατιστικής, στην εμπειρική διερεύνηση των οικονομικών φαινομένων, δημιουργώντας έτσι τις προαναφερθείσες νέες τεχνικές και μεθόδους, ως τότε άγνωστες στους μαθηματικούς στατιστικούς. Ο κλάδος της εφαρμοσμένης οικονομετρίας αποτέλεσε ουσιαστικά το επόμενο βήμα. Βασιζόμενος στα «ευρήματα» του πρώτου, και χρησιμοποιώντας δεδομένα και καταγραφές από την πραγματικότητα, εκτιμά και ελέγχει σχέσεις και αλληλεξαρτήσεις που χρήζουν εξήγησης, ώστε να ληφθούν αποφάσεις και να παραχθεί σχεδιασμός/προγραμματισμός. (Κιντής, 2010)

1.2.2 ΣΚΟΠΟΙ

Ουσιαστικά οι σκοποί της οικονομετρίας μπορούν να ανέλθουν σε τέσσερεις, από τους οποίους κάθε έρευνα και μελέτη μπορεί να «υπακούει» από έναν έως όλους. Αναλυτικά:

1. **Έλεγχος εμπειρικής σημαντικότητας οικονομικών θεωριών.**

Στον παρόν έλεγχο πραγματοποιείται η έρευνα για την σημαντικότητα μίας θεωρίας, ή εναλλακτικά, η επιλογή ως ορθότερης μεταξύ δύο ή περισσοτέρων θεωριών. Αυτό πραγματοποιείται με σύγκριση και μελέτη των λογικών συμπερασμάτων της εκάστοτε θεωρητικής οικονομικής ανάλυσης και των καταγεγραμμένων δεδομένων από την

πραγματικότητα. Από την σύγκριση αυτή μπορούν τα λογικά συμπεράσματα (και κατ' επέκταση η θεωρία που τα υποστηρίζει) να αποκτήσουν ή όχι εγκυρότητα. Ο σκοπός αυτός κυρίως υποστηρίζεται από την Θεωρητική Οικονομετρία η οποία έχει αναπτύξει διάφορες μεθόδους και τεχνικές για ακριβώς αυτόν τον λόγο.

2. **Διατύπωση εμπειρικών νόμων αναφορικά με την συμπεριφορά οικονομικών φαινομένων.**

Αφήνοντας τις θεωρητικές προσεγγίσεις και εστιάζοντας στα δεδομένα και τις καταγραφές των φαινομένων της πραγματικότητας, ο ερευνητής έχει ως σκοπό να αποκαλύψει και να θεμελιώσει νέες σχέσεις μεταξύ των μεταβλητών. Αυτές οι νέες σχέσεις αποτελούν πλέον νέα επιστημονική γνώση, η οποία βοηθά στην περαιτέρω έρευνα των φαινομένων που μελετά η Οικονομική και όχι μόνο επιστήμη.

3. **Άσκηση αποτελεσματικής πολιτικής και αξιολόγηση συνεπειών εναλλακτικών πολιτικών.**

Η άσκηση πολιτικής (ή η επιλογή μεταξύ διαφόρων) και η κατανόηση των επιπτώσεών της, αποτελούν κυρίαρχο σκοπό για την οικονομετρία καθώς και για το σύνολο των επιστημών σχεδιασμού και προγραμματισμού. Αυτό βέβαια προϋποθέτει την ποσοτικοποίηση των μεταβλητών και των συσχετίσεών τους ώστε να μελετηθούν με ακρίβεια και να αποφέρουν ορθά αποτελέσματα. Με τροποποίηση των στοιχείων σε μετρίσιμες/ποσοτικοποιημένες πλέον μεταβλητές μπορούμε να βασίζουμε την επιλογή ή απόρριψη πολιτικής σε δυνατές βάσεις και αναγνωρισμένες επιστημονικά και εμπειρικά βάσεις. Συνδέεται άμεσα με τον προηγούμενο σκοπό, καθώς και με τον ακόλουθο.

4. **Διενέργεια προβλέψεων.**

Ένα από τα σημαντικότερα προτερήματα των οικονομικών υποδειγμάτων αποτελούν οι προβλέψεις που μπορούν να διεξαχθούν από αυτά. Στον οικονομικό χώρο, όπως και στα πλαίσια αντίστοιχων επιστημών (για την παρούσα μελέτη η χωροταξία/πολεοδομία) η επισκόπηση της εξέλιξης συγκεκριμένων οικονομικών και άλλων μεγεθών αποτελεί συχνό φαινόμενο και διαδραματίζει πρωτεύον ρόλο για τον σχεδιασμό του χώρου και της οικονομίας, αφού σχεδόν πάντα χαρακτηρίζεται ως βραχυπρόθεσμος. Όσο πιο αντιπροσωπευτικές γίνονται οι προβλέψεις τόσο πιο σταθερός και αποδοτικός γίνεται ο σχεδιασμός για το μέλλον. (Κιντής, 2010)

Σημειώνεται πως ο ορισμός της οικονομετρίας παρατίθεται για λόγους πληρότητας της μελέτης παρά για την ίδια την χρήση της οικονομετρίας στα ακόλουθα. Λόγος αυτού

αποτελεί το περιεχόμενο των στοιχείων που επεξεργάζονται στην μελέτη περίπτωσης, τα οποία δεν εμπεριέχουν οικονομική διάσταση. Παρά ταύτα η κατανόηση του όρου είναι απαραίτητη για τις περιπτώσεις ερευνών κοινωνικοοικονομικού και χωρικού περιεχομένου.

1.3 ΜΕΤΑΔΕΔΟΜΕΝΑ

1.3.1 ΟΡΙΣΜΟΣ

Ο ευρέως αποδεκτός ορισμός της έννοιας των μεταδεδομένων είναι εκείνος ο οποίος τα θεωρεί ως δεδομένα που περιγράφουν δεδομένα (data about data). Όσον αφορά τη στατιστική, ο ανωτέρω όρος χρησιμοποιείται για να δηλώσει κάθε πληροφορία σχετικά με πραγματικά στατιστικά δεδομένα. (Grossmann, 1997) (Sundgren, 1996)

Αποτελούν τη βάση για ανάλυση δεδομένων και ουσιαστικά, τα μεταδεδομένα περιγράφουν αριθμητικά δεδομένα και τις ιδιότητές τους και αποτελούν πληροφορία σχετικά με τις πραγματικές τιμές των μεταβλητών, όπως για παράδειγμα ποια είναι ακριβώς η έννοια των τιμών που παρουσιάζονται στις αναλύσεις των δεικτών, πώς προέκυψαν αυτά τα αποτελέσματα, κλπ. (Grossmann & Papageorgiou, 1997)

Ανάμεσα στις χρήσεις των μεταδεδομένων, μπορούμε να θεωρήσουμε ως αντιπροσωπευτικές τις παρακάτω:

1. **Την ερμηνεία των αποτελεσμάτων**
2. **Την εγκυρότητα των δεδομένων**
3. **Κατευθύνσεις για την ανάλυση των αποτελεσμάτων** (Hand, 1993)

Επιπλέον, τα μεταδεδομένα έχουν μεγάλη σημασία και για δευτερογενή ανάλυση πληροφορίας, η οποία πραγματοποιείται από άλλους ερευνητές διαφορετικών αυτών που έκαναν τη συλλογή και αρχική ανάλυση των δεδομένων (π.χ. χωροτάκτες, περιβαλλοντολόγοι κ.α.). Στις περιπτώσεις αυτές, τα μεταδεδομένα προσδίδουν αναγκαίες πληροφορίες για την επεξεργασία δεδομένων που έχουν συλλεχθεί από τις πρωτογενείς πηγές και δίνουν την δυνατότητα κατανόησης της κωδικοποίησης της πληροφορίας, ώστε να καθίσταται εφικτή η σύγκρισή της με άλλες αντίστοιχες πηγές δεδομένων σε διαφορετικά επίπεδα και κλάδους.

1.3.2 ΠΡΟΣΕΓΓΙΣΕΙΣ

1. Απλά (free-text) μεταδεδομένα υπό μορφή υποσημειώσεων

Παρουσιάζουν την πληροφορία τους υπό μορφή υποσημειώσεων (footnotes) στους πίνακες των αντίστοιχων δεδομένων που συνοδεύουν. Τα free-text μεταδεδομένα αποτελούσαν μία εύκολη μορφή παρουσίασης της πληροφορίας, αλλά δεν αποσκοπούν στην επεξεργασία αυτής από τα στατιστικά πληροφοριακά συστήματα. Σε αυτή την περίπτωση, όταν κάποιος χρήστης θέλει να συγκρίνει στοιχεία παρόμοιων ερευνών από διαφορετικές όμως έρευνες, χρονικές περιόδους και πηγές, συνήθως οι υποσημειώσεις των πινάκων δεν αναφέρονται στα ίδια μεταδεδομένα, αλλά σε αυτά που ο εκάστοτε μελετητής της πληροφορίας παρουσιάζει, προκαλώντας έτσι δυσχέρειες στην περεταίρω έρευνα και αύξηση στο κόστος ανάλυσης και μελέτης της πληροφορίας.

2. Πινακοποιημένη μορφή με χρήση ενιαίων φορμών (Templates)

Αποτελούν ενιαίες φόρμες – πίνακες (templates) όπου τα μεταδεδομένα, αποτελούν έναν συνδυασμό από τα δομημένα μεταδεδομένα (π.χ. λίστες κωδικών και περιγραφές καταγεγραμμένων στοιχείων) και απλών, μη δομημένων μεταδεδομένων (π.χ. περιγραφές μεταβλητών). (Sundgren, 1996, 1999, 2004)

Η μεταπληροφορία που αποθηκεύεται, κωδικοποιείται μέσω μιας τυποποιημένης φόρμας και τέλος όταν εξαχθεί αυτοματοποιημένα μπορεί επίσης να διευκολύνει το χρήστη στην κατανόηση των αποτελεσμάτων και των δεδομένων και να πραγματοποιήσει εύκολα συγκρίσεις. Με αυτόν τον τρόπο απλοποιείται η διαδικασία για τους μελετητές και μειώνεται το κόστος αφού κάθε μεταδεδομένο το οποίο κωδικοποιείται σε μία φόρμα μπορεί εύκολα να αποθηκευτεί σε ένα πληροφοριακό σύστημα με μία σχεσιακή βάση δεδομένων. Επιπρόσθετα αυτή η μεταπληροφορία μπορεί στη συνέχεια να ανακτηθεί από μηχανές αναζήτησης (όπως Google) ώστε να συντομευτεί η διαδικασία για τον εντοπισμό των πληροφοριών. Επομένως, οι φόρμες αυτές καταγράφουν με κωδικοποιημένο, ενιαίο τρόπο τα μεταδεδομένα για όλες τις δειγματοληπτικές έρευνες, επιτρέπουν την αποθήκευσή τους σε βάσεις δεδομένων με προκαθορισμένο τρόπο και, στη συνέχεια δίνουν τη δυνατότητα εξαγωγής των μεταδεδομένων όταν ζητηθεί από τον χρήστη. (Sundgren, 2000) (Malvestuto, 1993)

Άρα η τεχνική των templates αποτέλεσε μία σαφή βελτίωση από την απλή παρουσίαση των μεταδεδομένων σε υποσημειώσεις που συνόδευαν τους αντίστοιχους πίνακες (όπως

παρουσιάστηκε προηγουμένως), λόγω του ότι ενέχει κάποια σχετική δόμηση των μεταδεδομένων υπό κοινό σχεδιασμό. (Froeschl κ.α., 2002)

Παρά ταύτα η παρούσα τεχνική παρουσιάζει σχετικούς περιορισμούς. Αρχικά ο σημαντικότερος είναι ότι αυτές οι φόρμες χρησιμοποιούνται για την εισαγωγή των μεταδεδομένων αλλά χωρίς να προβλέπεται η αυτοματοποιημένη χρήση τους από πληροφοριακά συστήματα ανάλυσης δεδομένων. Επιπρόσθετα, τα πληροφοριακά συστήματα δεν έχουν την ικανότητα να αντιλαμβάνονται και να επεξεργάζονται το περιεχόμενο της αποθηκευμένης πληροφορίας, οπότε χρησιμοποιείται απλά σαν free-tex. Σε τέτοιες λοιπόν περιπτώσεις οι υπολογιστές δεν μπορούν να βοηθήσουν το χρήστη σε οποιαδήποτε ανάλυση δεδομένων, ούτε να τον ειδοποιήσουν για το σφάλμα (όπως π.χ. να προσθέσει τα χρονικά δεδομένα μίας στήλης με μονάδα μέτρησης σε «πενταετίες» και μίας άλλης στήλη σε «δεκαετίες»). (Papageorgiou κ.α., 2000)

3. Μοντέλα μεταδεδομένων (metadata models)

Έτσι για να επιλυθούν τα προβλήματα που δημιουργούνται κατά την χρήση των templates, αναπτύχθηκε η ιδέα της μοντελοποίηση των μεταδεδομένων, η οποία παράλληλα αποτελεί και μέσο επίτευξης δυναμικής ανάλυσής τους. Η μεταπληροφορία συλλαμβάνεται χρησιμοποιώντας ένα μοντέλο μεταδεδομένων, οπότε οι υπολογιστές μπορούν να χρησιμοποιήσουν τη μεταπληροφορία αυτή κατά τη διάρκεια ολόκληρης της ανάλυσης και συστηματοποίησης των δεδομένων. (Froeschl, 1999)

Τα οφέλη από τη χρήση μοντέλων μεταδεδομένων είναι ποικίλα. Μερικά από τα πιο σημαντικά συνοψίζονται ως εξής:

i. Μειώνει τις περιπτώσεις ανθρώπινου σφάλματος αφού περιορίζεται στο ελάχιστο η ανάγκη ανθρώπινης παρέμβαση στην ανάλυση των δεδομένων. Έτσι βελτιώνεται η ποιότητα των υπηρεσιών που προσφέρονται από τους εκάστοτε φορείς.

ii. Μειώνει το φόρτο εργασίας του ανθρώπινου δυναμικού.

iii. Εν συνεχεία με το προηγούμενο, μειώνει το κόστος της επεξεργασίας με την ποιοτική και αυτοματοποιημένη εξαγωγή των δεικτών από πληροφοριακά συστήματα.

iv. Επιτρέπει στους μελετητές που δεν έχουν ιδιαίτερες γνώσεις στατιστικής να χρησιμοποιήσουν με ευκολία τα αποτελέσματα αφού επιτρέπει αυτοματοποιη-

μένη ανάλυση και εξαγωγή ποιοτικών αποτελεσμάτων. Αυτή είναι μία σημαντική προσφορά αν σκεφτεί κανείς ότι οι διάφοροι μελετητές (π.χ. πολιτική μηχ., χωροτάκτες, πολεοδόμοι, αγρονόμοι) δεν έχουν επαρκή γνώση/μόρφωση επί του κλάδου.

v. Μειώνει σημαντικά το χρόνο που απαιτείται για την επεξεργασία των δεδομένων, αφού ο ανθρώπινος παράγοντας περιορίζεται.

vi. Ελαχιστοποιούνται τα σφάλματα από παρανόηση των επεξηγήσεων αφού τα δομημένα μεταδεδομένα είναι θεσπισμένα υπό εθνικούς ή διεθνείς ορισμούς.

(Froeschl, 1997)

1.4 ΜΕΤΑΒΛΗΤΕΣ

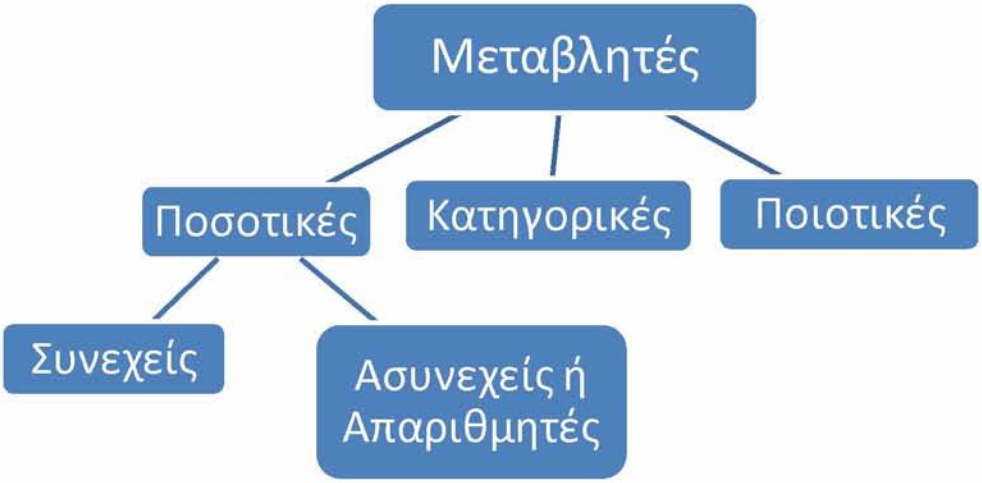
1.4.1 ΟΡΙΣΜΟΣ

Στα πλαίσια των μαθηματικών ως μεταβλητή εννοείται κάθε ποσότητα η οποία μπορεί να πάρει δύο ή περισσότερες διαφορετικές τιμές, άρα κατά αντιστοιχία στην στατιστική, ως τυχαία μεταβλητή ορίζεται κάθε χαρακτηριστικό ή ιδιότητα των αντίστοιχων στατιστικών μονάδων. Τα παραδείγματα ποικίλουν αφού ουσιαστικά καλύπτεται τεράστιο εύρος πειραμάτων και παρατηρήσεων. Ενδεικτικά μεταβλητή μπορεί να είναι το ύψος, ο αριθμός των ιδιοκτησιών ακινήτων ανά άτομο, ο συντελεστής δόμησης από κάποιο γενικό πολεοδομικό σχέδιο, ή ακόμα και το ποσοστό κατά το οποίο έχει γίνει πληρότητα του συντελεστή δόμησης. Σύμφωνα με τα χαρακτηριστικά της κάθε μεταβλητής μπορούμε την κατηγοριοποιήσουμε σε κάθε ένα από τα ακόλουθα είδη. (Παπαδήμας Ο. και Κοίλιας Χ., 1998)

1.4.2 ΕΙΔΗ

Ο διαχωρισμός των μεταβλητών φαίνεται στο παρακάτω διάγραμμα:

Διάγραμμα 1: Είδη μεταβλητών



Πηγή: Παπαδήμας Ο. και Κοίλιας Χ., 1998

Παρακάτω παρουσιάζεται η κάθε κατηγορία αναλυτικότερα :

1. **Ποσοτικές**

   i. **Συνεχείς**: χαρακτηρίζονται εκείνες οι μεταβλητές που η παρατηρούμενη μετρήσιμη αλλαγή τους είναι προοδευτική, με απροσδιόριστες ενδιάμεσες διαφορές όπως ποσοστά βαρέων μετάλλων στο περιβάλλον (1.25/1.78/2.56 %), η καταγραφή των βροχοπτώσεων, το ύψος των ατόμων, ο συντελεστής δόμησης (1/1.3/1.6)

   ii. **Ασυνεχείς ή απαριθμητές**: χαρακτηρίζονται εκείνες οι μεταβλητές που λαμβάνουν αριθμητικές τιμές από ενδιάμεσες διακοπές που έχουν προηγουμένως σαφώς καθοριστεί, όπως αριθμός αυτοκινήτων που πουλήθηκαν ανά έτος, αριθμός παικτών αθλητικής ομάδας, δήμοι με πραγματοποιημένο γενικό πολεοδομικό.

2. **Κατηγορικές:** λαμβάνουν τις τιμές τους από έννοιες που ιεραρχούνται όπως: επικλινές εδάφους (πεδινό/ημιορεινό/ορεινό), βλάστηση (άγονο/αραιή/πυκνή).

3. **Ποιοτικές**: λαμβάνουν τις τιμές τους από έννοιες που δεν ιεραρχούνται όπως: χρήση γης (βιομηχανία/κατοικία/εκπαίδευση), κάλυψη εδάφους (τεχνητό/φυσικό/θαλ. ύδατα/λιμνάζοντα ύδατα/καλλιέργειες)

(Παπαδήμας Ο. και Κοίλιας Χ., 1998)

2. ΘΕΩΡΗΤΙΚΟ ΥΠΟΒΑΘΡΟ

Αρχικά για να ικανοποιηθεί η ανάγκη για καταγραφή και επεξεργασία δεδομένων δημιουργήθηκαν στατιστικές υπηρεσίες κατάλληλων για την αποθήκευσης και αρχειοθέτηση δεδομένων καθώς και την αποδοτικότερη επεξεργασία και συλλογή αποτελεσμάτων. Στην συνέχεια η ανάγκη από διάφορους κλάδους, πέρα της στατιστικής, οδήγησε τους μελετητές να ασχολούνται και να επεκτείνουν τις γνώσεις τους, προσαρμόζοντας την στατιστική στον δικό τους χώρο, ώστε να επεξεργάζονται γρηγορότερα και αποδοτικότερα τα δεδομένα τους. Η διαδικασία της επεξεργασίας δεδομένων και συλλογής αποτελεσμάτων περιλαμβάνει ένα αρχικό στάδιο, το στάδιο της ανάλυσης των δεδομένων.

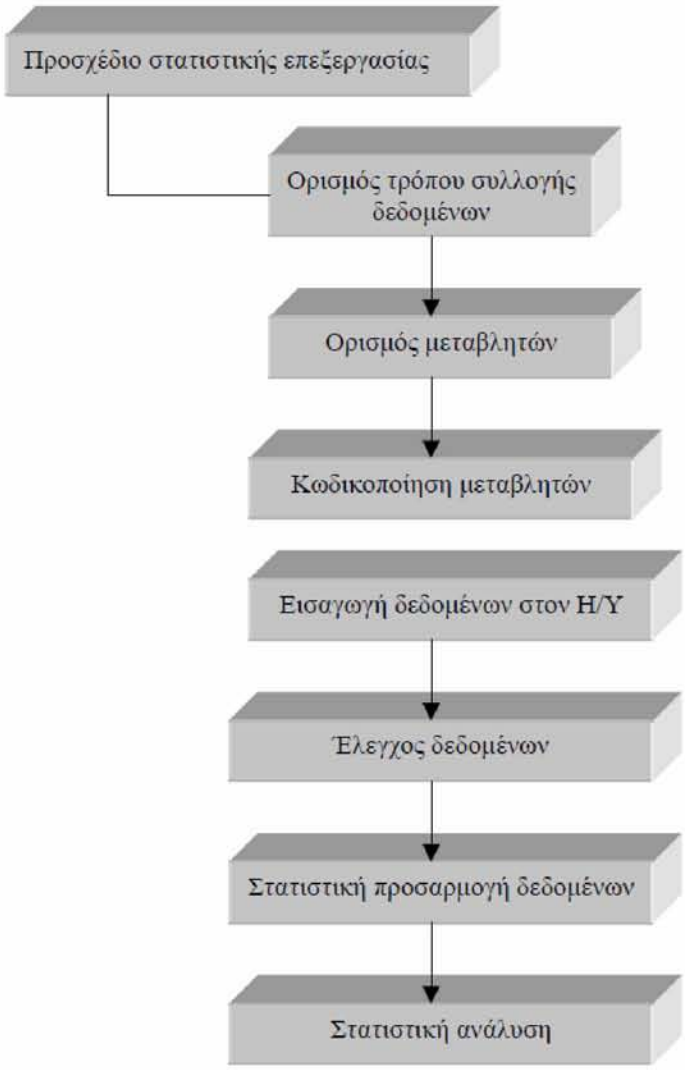
Ουσιαστικά στον όρο αυτό δεν περιλαμβάνονται μόνο οι τεχνικές και ο μέθοδοι επεξεργασίας πληροφοριών που προέκυψαν από πραγματικά/εικονικά πειράματα ή παρακολούθηση φαινομένων, αλλά και η θεσμοθέτηση κοινά παραδεκτών τεχνικών με σκοπό την εκτίμηση των χαρακτηριστικών του πληθυσμού και περεταίρω στόχο την χρήση τους σε συμπεράσματα και αναλύσεις. (Ζήμερας, 2003)

2.1 ΠΡΟΕΤΟΙΜΑΣΙΑ ΣΥΛΛΕΧΘΕΝΤΩΝ ΔΕΔΟΜΕΝΩΝ

Όσον αφορά την διαδικασία της προετοιμασίας των δεδομένων, αρχικά, επιλέγεται το θέμα έρευνας, συνεχίζει με τον σχεδιασμό συγκεκριμένης μεθοδολογίας που θα ακολουθηθεί και ολοκληρώνεται με την υλοποίηση και εφαρμογή των συγκεκριμένων στατιστικών μεθόδων. Πρέπει βέβαια να δοθεί ιδιαίτερη προσοχή ώστε να αποφευχθούν μεροληπτικά αποτελέσματα και λανθασμένες ερμηνείες λόγω ανεπαρκής ή ελλιπής προετοιμασία των δεδομένων, το οποίο οδηγεί σε στατιστικές αναλύσεις χαμηλής ποιότητας και αναξιοπιστία των αποτελεσμάτων/συμπερασμάτων.

Στην πρώτη φάση ελέγχεται το όργανο συλλογής των δεδομένων. Στην δεύτερη γίνεται η αντιστοίχηση των δεδομένων με τις μεταβλητές. Ακολουθεί η απόφαση για τον τρόπο κωδικοποίησης των μεταβλητών καθώς και ο τρόπος εισαγωγής των δεδομένων στον Η/Υ (λήψη απόφασης όσο αφορά το συγκεκριμένο στατιστικό πακέτο). Στην συνέχεια ελέγχεται η λογικότητα των δεδομένων και αποφασίζεται ο τρόπος χειρισμού των παρατηρήσεων που δεν έχουν καταγραφεί (ελλείπουσες τιμές). Τελικά στάδια είναι η στατιστική προσαρμογή των δεδομένων έτσι ώστε να υπάρξει η απαιτούμενη αντιπροσωπευτικότητα του πληθυσμού και ο ορισμός της στατιστικής ανάλυσης που θα ακολουθηθεί. Τα παραπάνω στάδια παρουσιάζονται στο διάγραμμα 2. (Ζήμερας, 2003)

Διάγραμμα 2: Γραφική παρουσίαση προετοιμασίας δεδομένων



Πηγή: Ζήμερας, 2003

Ειδικότερα τα στάδια προετοιμασίας των δεδομένων μπορούν να αναλυθούν στα επίπεδα που ακολουθούν στην συνέχεια του κεφαλαίου.

2.1.1 ΟΡΙΣΜΟΣ ΤΡΟΠΟΥ ΣΥΛΛΟΓΗΣ ΔΕΔΟΜΕΝΩΝ

Αποτελεί το αρχικό στάδιο ανάλυσης όπου η χρήση ενός συγκεκριμένου εργαλείου συλλογής δεδομένων πρέπει να χρησιμοποιηθεί. Τα σημαντικότερα και πλέον αποδεκτά μέσα συλλογής μπορεί να είναι: ερωτηματολόγια, πρόσβαση σε βάσεις δεδομένων, προσωπικές παρατηρήσεις, WEB, στατιστικές υπηρεσίες.

2.1.2 ΟΡΙΣΜΟΣ ΜΕΤΑΒΛΗΤΩΝ

Καθορισμός μεταβλητών από ερωτηματολόγια ή βάσεις δεδομένων. Ορισμός πρωταρχικών και δευτερευουσών μεταβλητών. Στην απλούστερη περίπτωση κάθε πεδίο ή ερώτηση από την συλλογή δεδομένων αποτελεί μια μεταβλητή. Το όνομα της μεταβλητής είναι καλό να είναι βολικό για μελλοντική ανάλυση, με λίγους χαρακτήρες και να αντιπροσωπεύει τα χαρακτηριστικά που καταγράφει η μεταβλητή. π.χ. ΦΥΛΟ – SEX ή ΗΛΙΚΙΑΚΗ ΟΜΑΔΑ – AGE_GROUP ή AGE.

2.1.3 ΚΩΔΙΚΟΠΟΙΗΣΗ ΜΕΤΑΒΛΗΤΩΝ

Εννοούμε την αντιστοίχηση κωδικών σε όλες τις πιθανές τιμές μίας μεταβλητής. Οι κωδικοί είναι συνήθως αριθμοί π.χ. ΦΥΛΟ – ΑΓΟΡΙ (1) , ΚΟΡΙΤΣΙ (2), αλλά μπορεί να είναι και χαρακτήρες π.χ. ΦΥΛΟ – ΑΓΟΡΙ (Α) , ΚΟΡΙΤΣΙ (Κ). Ίδιες τιμές δύο διαφορετικών χαρακτηριστικών πρέπει να αντιστοιχούν ακριβώς στον ίδιο κωδικό. Για παράδειγμα δεν μπορεί το φύλλο του ερωτούμενου σε ένα ερωτηματολόγιο να το κωδικοποιούμε αλλού με Α και αλλού με α για τον άνδρα/αγόρι και Γ/γ ή Κ/κ για την γυναίκα/κορίτσι. Οι ποσοτικές μεταβλητές είναι ήδη κωδικοποιημένες. Όλοι οι χρησιμοποιούμενοι κωδικοί μιας έρευνας συνήθως καταγράφονται σε έναν πίνακα που ονομάζεται πίνακας κωδικοποίησης .

2.1.4 ΕΙΣΑΓΩΓΗ ΔΕΔΟΜΕΝΩΝ ΣΤΟΝ Η/Υ

Το σημαντικότερο στάδιο της ανάλυσης δεδομένων είναι η εισαγωγή τους στο Η/Υ. Η διαδικασία τις περισσότερες φορές είναι επίπονη και κουραστική, καταναλώνοντας αρκετό χρόνο εξαρτώμενος από τον αριθμό και την κωδικοποίηση των δεδομένων.

2.1.5 ΕΛΕΓΧΟΣ ΔΕΔΟΜΕΝΩΝ

Πολλοί λόγοι μπορεί να οδηγήσουν στην ύπαρξη παράλογων τιμών. Όποιες τιμές εμφανίζονται ακραίες ή λανθάνουσες πρέπει να ελέγχονται σχολαστικά.

2.1.6 ΣΤΑΤΙΣΤΙΚΗ ΠΡΟΣΑΡΜΟΓΗ ΔΕΔΟΜΕΝΩΝ

Περιλαμβάνει στην κατασκευή νέων μεταβλητών που είναι απαραίτητες για την ανάλυση. Η δημιουργία βουβών μεταβλητών μπορεί να επαναπροσδιορίσει μόνο ποιοτικά δεδομένα. Οι πιθανές τιμές των μεταβλητών είναι τις περισσότερες φορές 0 ή 1.

2.2 ΑΝΤΙΜΕΤΩΠΙΣΗ ΕΛΛΙΠΩΝ ΔΕΔΟΜΕΝΩΝ

Σαν ελλείποντα δεδομένα χαρακτηρίζονται εκείνα τα δεδομένα τα οποία δεν έχουν καταγραφεί. Ο αντίκτυπος της απώλειας των δεδομένων, όσον αφορά την ισχύ των ερευνητικών συμπερασμάτων, εξαρτάται από τους μηχανισμούς που οδήγησαν στην απώλεια αυτών, το είδος τους και το ποσοστό τους επί του συνολικού δείγματος. Επιπροσθέτως, έχει αποδειχθεί ότι ο μηχανισμός και το είδος της απώλειας των δεδομένων ασκούν μεγαλύτερη επιρροή στα ερευνητικά αποτελέσματα σε σχέση με την έλλειψη μεγαλύτερου αριθμού δεδομένων. (Παπαδάκη, 2009)

2.2.1 ΑΙΤΙΕΣ ΑΠΩΛΕΙΑΣ ΤΟΝ ΔΕΔΟΜΕΝΩΝ

Τα προαναφερθέντα ζητήματα είναι κρίσιμα ζητήματα που ένας ερευνητής πρέπει να αντιμετωπίσει πριν επιλέξει μια διαδικασία για να εξετάσει τα χαμένα στοιχεία. Σύμφωνα με τους Little και Rubin (1987), τα στοιχεία μπορεί: να λείπουν εντελώς τυχαία (missing completely at random), να λείπουν τυχαία (missing at random), και να λείπουν για άγνωστη αιτία (non- gnorable). Αναλυτικά, τα στοιχεία μπορεί:

1. **Να λείπουν εντελώς τυχαία *(missing completely at random)***

Υπάρχουν διάφοροι λόγοι για τους οποίους ορισμένα στοιχεία μπορούν να λείπουν. Αυτό μπορεί να συμβεί είτε επειδή ο εξοπλισμός δυσλειτούργησε είτε επειδή παρουσιάστηκε κακοκαιρία και έτσι δεν πραγματοποιήθηκαν μετρήσεις κάποιες ορισμένες μέρες είτε επειδή τα στοιχεία δεν εισήχθησαν σωστά. Στην περίπτωση αυτή τα στοιχεία λείπουν εντελώς τυχαία (MCAR). Συγκεκριμένα, όταν λέμε ότι τα στοιχεία λείπουν εντελώς τυχαία, σημαίνει ότι η πιθανότητα μια παρατήρηση (Xi) να λείπει είναι ανεξάρτητη από την τιμή του Xi ή από την τιμή οποιωνδήποτε άλλων συμμεταβλητών. Κατά συνέπεια, στοιχεία που αφορούν το οικογενειακό εισόδημα δεν θα μπορούσαν να θεωρηθούν MCAR εάν οι άνθρωποι με τα χαμηλά εισοδήματα ήταν λιγότερο πιθανό να αναφέρουν το οικογενειακό εισόδημά τους από τους ανθρώπους με υψηλότερα εισοδήματα. Ομοίως, εάν οι Λευκοί Αμερικανοί πολίτες ήταν πιθανότερο από τους Αφρικανικούς Αμερικανούς πολίτες να παραλείψουν το εισόδημα, πάλι δεν θα είχαμε τα στοιχεία που να είναι MCAR επειδή στην συγκεκριμένη περίπτωση η απώλεια συσχετίζεται με το έθνος. Παρατηρήστε ότι είναι η αξία της παρατήρησης, και όχι η απώλεια (missingness) αυτή καθεαυτή που έχει σημασία. Εάν οι ερωτώμενοι που αρνήθηκαν να εκθέσουν το ατομικό τους εισόδημα είναι επίσης πιθανό να αρνηθούν να εκθέσουν και το οικογενειακό τους εισόδημα, τότε τα στοιχεία αυτά μπορούν να θεωρηθούν MCAR,

εφ' όσον κανένα από αυτά δεν έχει οποιαδήποτε σχέση με το ύψος του εισοδήματος αυτό καθεαυτό. (Παπαδάκη, 2009)

2. Να λείπουν τυχαία *(missing at random)*

Συχνά τα στοιχεία δεν λείπουν εντελώς τυχαία, αλλά αυτά μπορούν να είναι ταξινομήσιμα όπως ελλείποντα τυχαία (MAR). Για τα στοιχεία που λείπουν εντελώς τυχαία, η πιθανότητα το Xi στοιχείο να λείπει είναι ανεξάρτητη από την αξία του Xi. Αλλά τα στοιχεία μπορούν να θεωρηθούν ως να λείπουν τυχαία εάν τα στοιχεία αυτά ικανοποιούν την συνθήκη ότι η απώλεια (missingness) δεν εξαρτάται από την αξία του Xi αν αυτό δεσμευτεί για μια άλλη μεταβλητή. Παραδείγματος χάριν, οι άνθρωποι που είναι καταθλιπτικοί πιθανώς να είναι λιγότερο πρόθυμοι να εκθέσουν το εισόδημά τους και συνεπώς το αναφερόμενο εισόδημα σχετίζεται με την κατάθλιψη. Εντούτοις, εάν μέσα στους καταθλιπτικούς ασθενείς η πιθανότητα κάποιος να αναφέρει το εισόδημά του ήταν ανεξάρτητη από το εισοδηματικό επίπεδο, τότε τα στοιχεία θα εξετάζονταν σαν MAR. (Παπαδάκη, 2009)

3. Να λείπουν για άγνωστη αιτία *(Ignorability)*

Εάν τα στοιχεία είναι MCAR ή MAR, λέμε ότι η απώλεια (missingness) εμφανίστηκε για λόγο του οποίου η γνώση δεν είναι δυνατή. Η περίπτωση αυτή ονομάζεται «ignorable». Με τον όρο αυτό εννοούμε ότι δεν είναι απαραίτητο να μοντελοποιήσουμε την απώλεια. Ακόμη και στη σπάνια κατάσταση όπου ο όρος (2) δεν ικανοποιείται, ορισμένες μέθοδοι που υποθέτουν ότι το «ignorability» ισχύει μπορούν να εφαρμοστούν, αλλά το καλύτερο θα ήταν να διαμορφωθεί ένα μοντέλο για την αντιμετώπιση του προβλήματος των ελλειπόντων δεδομένων. Σε αντίθεση με τα στοιχεία που λείπουν τυχαία, τα ελλείποντα στοιχεία λείπουν για γνωστή αιτία, δηλαδή είναι «non- ignorable» εάν η πιθανότητα της απώλειας των δεδομένων εξαρτάται από τις τιμές αυτών. Αντίθετα από την προηγούμενη περίπτωση όπου αγνοούμε την αιτία που έχουμε ελλείποντα στοιχεία, στην συγκεκριμένη περίπτωση το μοντέλο που περιγράφει τα ελλείποντα στοιχεία πρέπει να διευκρινιστεί από τον ερευνητή και να ενσωματωθεί στην ανάλυση των δεδομένων, προκειμένου να παραχθούν οι αμερόληπτες εκτιμήσεις των παραμέτρων. Είναι εύκολο να καταλάβουμε ότι τα «non-ignorable» απολεσθέντα στοιχεία στην έρευνα είναι πλέον πιθανό να εμφανιστούν στις μελέτες που επιδιώκουν να συγκεντρώσουν τις ευαίσθητες ή προσωπικές πληροφορίες των διαφόρων ατόμων. Αξίζει να σημειώσουμε ότι κανένα στατιστικό τεστ δεν υπάρχει προς το παρόν που να εξετάζει εάν αυτός ο όρος ικανοποιείται. (Παπαδάκη, 2009)

2.2.2 ΣΗΜΑΝΤΙΚΑ ΣΗΜΕΙΑ ΠΕΡΙ ΑΙΤΙΩΝ ΑΠΩΛΕΙΑΣ ΔΕΔΟΜΕΝΩΝ

Στο σημείο αυτό θα πρέπει να σημειώσουμε ότι ο όρος MCAR μπορεί να εξεταστεί χρησιμοποιώντας το πολυμεταβλητό test του Little το οποίο εξετάζει εάν ο όρος MCAR ισχύει για τα συγκεκριμένα δεδομένα.. Η περίπτωση του όρου MAR μπορεί να εξεταστεί με ένα απλό t-test των μέσων διαφορών μεταξύ του γκρουπ με τα πλήρη στοιχεία και αυτού με τα απολεσθέντα στοιχεία. Εντούτοις, προειδοποιούμε τους αναγνώστες ότι τα αποτελέσματα αυτών των ελέγχων δεν παρέχουν αναμφισβήτητες αποδείξεις ότι βρισκόμαστε είτε στην περίπτωση MCAR είτε στην MAR. Σχετικά με την ερώτηση που αφορά το μέγιστο ποσοστό των απολεσθέντων δεδομένων που μπορεί μια μέθοδος να δώσει αξιόπιστα αποτελέσματα, δεν υπάρχει καμία συγκεκριμένη απάντηση που να βρίσκει σύμφωνους όλους τους στατιστικούς αυτή τη στιγμή. Στην περίπτωση που λείπουν μόνο λίγες τιμές σε ένα τυχαίο υπόδειγμα από ένα μεγάλο σύνολο δεδομένων (δηλ. στην περίπτωση MCAR), το πρόβλημα ελλειπόντων δεδομένων είναι λιγότερο σοβαρό και σχεδόν κάθε μέθοδος για τα χαμένα δεδομένα παράγει παρόμοια αποτελέσματα. Εντούτοις, αν ένα μεγάλο ποσοστό δεδομένων λείπει, τότε το πρόβλημα μπορεί να είναι πολύ σοβαρό. (Παπαδάκη, 2009)

2.2.3 ΕΠΙΣΚΟΠΗΣΗ ΤΩΝ ΠΕΝΤΕ ΕΙΔΙΚΩΝ ΜΕΘΟΔΩΝ ΕΛΛΕΙΠΟΝΤΩΝ ΔΕΔΟΜΕΝΩΝ

Η ιστορία της ανάπτυξης των μεθόδων των ελλειπόντων δεδομένων μπορεί να διαιρεθεί σε τρεις χρονικές περιόδους (Schafer, 1997). Στην πρώτη περίοδο, που αφορά το χρονικό διάστημα πριν από το έτος 1980, οι περισσότερες ευρέως εφαρμοσμένες μέθοδοι που εξέταζαν τα ελλείποντα δεδομένα ήταν ειδικές. Αυτές περιλαμβάνουν το LD, το PD, την Αντικατάσταση με τον Μέσο (Mean Substitution), την απλή Hot-Deck μέθοδο και τις διάφορες βασισμένες στην Παλινδρόμηση μεθόδους (Regression- Based Methods). Οι προαναφερθείσες μέθοδοι είναι εύκολες στην χρήση αλλά παράγουν μεροληπτικούς εκτιμητές.

Στη δεύτερη περίοδο, κατά προσέγγιση αρχίζοντας με τη δημοσίευση του άρθρου των Little και Rubin (1987), οι μέθοδοι, όπως η Πλήρης Πληροφορίας Μέγιστη Πιθανοφάνεια (Full Information Maximum Likelihood ή FIML) και ο Αλγόριθμος EM (Expectation Maximization), άρχισαν να εμφανίζονται. Αυτές οι μέθοδοι είναι γενικά ανώτερες από τις ειδικές μεθόδους δεδομένου ότι είναι στατιστικά επαρκείς και παράγουν εκτιμητές των παραμέτρων με αποδεκτά τυπικά σφάλματα. Αν και αυτές οι μέθοδοι βασίζονται σε συγκεκριμένα μοντέλα και μπορεί να είναι δύσκολο να εφαρμόσουν,

αντιμετωπίστηκαν ως πολύ σημαντικές και συνάμα εντυπωσιακές ανακαλύψεις στην ιστορία των μεθόδων των απολεσθέντων δεδομένων. (Παπαδάκη, 2009)

Η τρίτη περίοδος στην ανάπτυξη των μεθόδων των ελλειπόντων δεδομένων άρχισε προς το τέλος της δεκαετίας του '80 και στις αρχές της δεκαετίας του ΄90 και χαρακτηρίστηκε από την εισαγωγή των μεθόδων πολλαπλής απόδοσης (multiple imputation methods) για να υπερνικηθούν οι περιορισμοί των απλών μεθόδων απόδοσης. Οι μελέτες προσομοίωσης έχουν δείξει ότι η μέθοδος Πολλαπλής Απόδοσης είναι ευπροσάρμοστη και ότι παράγει εκτιμητές με μικρότερα τυπικά σφάλματα από εκείνα που λαμβάνονται από τις άλλες διαδικασίες. Αν και η μέθοδος πολλαπλής απόδοσης αποτελεί την πιο πρόσφατη προσπάθεια από τους μελετητές σχετικά με την εξέταση του προβλήματος των ελλειπόντων δεδομένων, δεν έχει υιοθετηθεί ευρέως από τους ερευνητές. (Παπαδάκη, 2009)

Σε αυτό το μέρος της εργασίας, αναφέρουμε τις πέντε ειδικές μεθόδους για τα χαμένα δεδομένα λόγω της επικράτησής τους στο στατιστικό λογισμικό. Αυτές περιλαμβάνουν τις δύο μεθόδους που περιγράψαμε νωρίτερα (δηλ. την Listwise και Pairwise deletion), μαζί με την Αντικατάσταση από τον Μέσο (Mean Substitution), την απλή Hot-Deck μέθοδο, και την Παλινδρόμηση. Τα πλεονεκτήματα και τα μειονεκτήματα κάθε μεθόδου συζητούνται από την άποψη της εκτίμησης των παραμέτρων και των ελέγχων υποθέσεων. Ακόμη, παραθέτουμε και τις μεθόδους: Principled Single Imputation- FIML, Principled Single Imputation- EM και Multiple Imputation- MI. (Παπαδάκη, 2009)

1. Listwise διαγραφή (LD)

Όπως σημειώθηκε και νωρίτερα, το LD αφαιρεί από τη στατιστική ανάλυση τα υποκείμενα που έχουν τις ελλείπουσες πληροφορίες για μία ή περισσότερα μεταβλητές. Όπως ο Kim και ο Curry (1977) παρουσιάζουν, το 59% των δεδομένων μπορεί να χαθεί χρησιμοποιώντας την μέθοδο LD εάν μόνο 10% των δεδομένων χαθούν τυχαία από κάθε μεταβλητή σε ένα σύνολο δεδομένων με πέντε μεταβλητές.

Το LD είναι η ευκολότερη και η πιο κοινή μέθοδος για τα χαμένα δεδομένα, ενώ αυτή η δημοτικότητά της οφείλεται κυρίως στο γεγονός ότι **Listwise Διαγραφή** αποτελεί προεπιλογή σε περιπτώσεις πολυμεταβλητών και διαφόρων μονομεταβλητών στατιστικών διαδικασιών στα δημοφιλή στατιστικά πακέτα, όπως το SPSS®, το SYSTAT® και το SAS®.

Αν και το LD γενικά δεν συστήνεται, μπορεί να χρησιμοποιηθεί ακίνδυνα εάν τόσο οι συσχετίσεις μεταξύ τεσσάρων ή λιγότερων μεταβλητών είναι μικρές όσο και το ποσοστό των απολεσθέντων δεδομένων είναι μικρό. Εφ' όσον ισχύει η υπόθεση MCAR για τα χαμένα δεδομένα, ο Allison βεβαίωσε ότι μεταξύ των συμβατικών μεθόδων για τα ελλιπή δεδομένα, η Listwise Deletion είναι η λιγότερο προβληματική μέθοδος. Αν και η Listwise Deletion μπορεί να απορρίψει ένα ουσιαστικό μέρος των δεδομένων, δεν υπάρχει κανένας λόγος να αναμείνει κανείς μεροληψία εκτός αν τα στοιχεία δεν λείπουν εντελώς τυχαία. Επιπλέον, τα τυπικά σφάλματα θα είναι αρκετά καλές εκτιμήσεις των αληθινών τυπικών σφαλμάτων των διαφόρων εκτιμητών. (Παπαδάκη, 2009)

Επιπλέον, εάν εκτιμούμε ένα μοντέλο γραμμικής παλινδρόμησης, η Listwise Deletion είναι αρκετά ευσταθής σε καταστάσεις όπου υπάρχουν ελλειπόντα δεδομένα στην ανεξάρτητη μεταβλητή και η πιθανότητα απώλειας εξαρτάται από την αξία εκείνης της μεταβλητής. Εάν εκτιμούμε ένα λογιστικό μοντέλο παλινδρόμησης, τότε η Listwise Deletion μπορεί να εφαρμοστεί έστω και αν υπάρχει μη τυχαία απώλεια στην εξαρτημένη μεταβλητή ή μη τυχαία απώλεια στις ανεξάρτητες μεταβλητές (αλλά όχι και στις δύο). (Παπαδάκη, 2009)

Με άλλα λόγια, όταν έχουμε περίπτωση MCAR, το LD δεν παράγει τις μεροληπτικές εκτιμήσεις και οι συνήθεις στατιστικές διαδικασίες μπορούν να εφαρμοστούν. Ακόμα, η απώλεια στη στατιστική ισχύ κάθε ελέγχου και στην ακρίβεια της εκτίμησης δεν πρέπει να αγνοηθεί, ούτε μπορεί να αντισταθμιστεί. Εάν η υπόθεση MCAR δεν ισχύει, τότε το LD παράγει μεροληπτικές εκτιμήσεις παραμέτρων και μεροληπτικά στατιστικά τεστ και τα αποτελέσματα είναι μη αντιπροσωπευτικά του πληθυσμού από τον οποίο επιλέχθηκε το δείγμα (Cohen & Cohen, 1983). Το LD επομένως δεν αποτελεί μια ικανοποιητική λύση στο πρόβλημα των ελλειπόντων δεδομένων. (Παπαδάκη, 2009)

2. Pairwise διαγραφή (PD)

Η Pairwise διαγραφή διατηρεί όλα τα διαθέσιμα δεδομένα που παρέχονται από ένα υποκείμενο. Όταν αυτή η προσέγγιση εφαρμόζεται στην ανάλυση των δεδομένων, οι διάφορες μέθοδοι της περιγραφικής στατιστικής και μερικές της επαγωγικής στατιστικής, όπως το t-test, το z-test, το chi-square καθώς και άλλα υπολογίζονται από τα μη - ελλείποντα δεδομένα ανά μεταβλητή. Αν η Pairwise διαγραφή είναι τόσο εύκολη όσο η Listwise Διαγραφή, δεν χρησιμοποιείται τόσο ευρέως όσο η προαναφερθείσα. Σύμφωνα με τους Kim και Curry (1977), η Pairwise Διαγραφή είναι μια ελκυστική εναλλακτική

λύση όταν υπάρχει μικρός αριθμός ελλειπόντων δεδομένων σε κάθε μεταβλητή σε σχέση με το συνολικό μέγεθος δείγματος καθώς και μεγάλος αριθμός μεταβλητών.

Έναντι της μεθόδου Listwise Deletion, η Pairwise Deletion χρησιμοποιεί τις πληροφορίες που λαμβάνονται από τις μερικώς πλήρεις παρατηρήσεις. Το μειονέκτημά της είναι ότι το μέγεθος του δείγματος αλλάζει από μεταβλητή σε μεταβλητή. Αυτή η μεταβλητότητα στο μέγεθος δείγματος δημιουργεί πρακτικά προβλήματα, όπως ο προσδιορισμός του συνολικού μεγέθους δείγματος και ο προσδιορισμός του αριθμού των βαθμών ελευθερίας. Όπως η Listwise Deletion, η Pairwise Deletion παράγει μεροληπτικές εκτιμήσεις των παραμέτρων και μεροληπτικά στατιστικά τεστ εκτός αν η υπόθεση MCAR ισχύει. Για αυτούς τους λόγους, η Pairwise Deletion δεν είναι μια ικανοποιητική λύση στο πρόβλημα των ελλειπόντων δεδομένων. (Παπαδάκη, 2009)

3. Αντικατάσταση με τον Μέσο (Mean Substitution)

Η προσέγγιση MS «λύνει» το πρόβλημα των ελλειπόντων δεδομένων με την αντικατάσταση των ελλειπόντων τιμών με το μέσο όρο της κάθε μεταβλητής. Αυτό το βήμα ολοκληρώνεται στην αρχή της ανάλυσης δεδομένων. Επομένως η συγκεκριμένη μέθοδος υποθέτει ότι ο μέσος όρος μίας μεταβλητής είναι η καλύτερη εκτίμηση για οποιαδήποτε παρατήρηση που λείπει από την συγκεκριμένη μεταβλητή. Σε αντίθεση με την Listwise Deletion και Pairwise Deletion, η μέθοδος MS δεν αλλάζει το μέσο όρο της μεταβλητής και δεν απορρίπτει οποιεσδήποτε πληροφορίες έχουν ήδη συλλεχθεί. Μια παραλλαγή του MS είναι η περίπτωση όπου μία χαμένη τιμή αντικαθίσταται με τον μέσο της υποομάδας στην οποία ανήκει. Παραδείγματος χάριν, εάν η παρατήρηση με μια ελλείπουσα τιμή ανήκει σε ένα άτομο που είναι Δημοκρατικός, ο μέσος όρος για όλους τους Δημοκρατικούς υπολογίζεται και αντικαθιστά την ελλείπουσα αυτή τιμή. Αυτή η διαδικασία δεν είναι τόσο συντηρητική όσο να παρεμβάλλαμε το γενικό μέσο όρο της μεταβλητής. (Παπαδάκη, 2009)

Ανεξάρτητα από το είδος της μεθόδου MS που εφαρμόζεται, αυτή η μέθοδος έχει πολλούς περιορισμούς. Σύμφωνα με τους Little και Rubin (1987), οι περιορισμοί είναι ότι: (α) Το μέγεθος του δείγματος υπερεκτιμάται, (β) η διασπορά υποτιμάται, (γ) οι συσχετίσεις είναι αρνητικά μεροληπτικές, και (δ) η κατανομή των νέων τιμών είναι μια ανακριβής αντιπροσώπευση των τιμών του πληθυσμών, επειδή η μορφή της κατανομής διαστρεβλώνεται με την προσθήκη των τιμών οι οποίες ταυτίζονται με το μέσο όρο. Η μεροληψία που εισάγεται στη διασπορά του πληθυσμού, στην συσχέτιση και την κατα-

νομή των μεταβλητών εξαρτάται από το ποσοστό των δεδομένων που λείπουν. Οι Little και Rubin συνιστούν να μην χρησιμοποιείται ποτέ η μέθοδος MS.

4. Απλή Hot-Deck (HD)

Η μέθοδος Hot-Deck αντικαθιστά κάθε ελλείπουσα τιμή με μια τυχαία επιλεγμένη τιμή από το δείγμα στην ίδια μεταβλητή. Οι παράμετροι που υπολογίζονται μ' αυτό τον τρόπο έχουν τις μεγαλύτερες διασπορές έναντι εκείνων που υπολογίζονται από την MS, αλλά μικρότερες διασπορές από εκείνες του πλήρους δείγματος. Το σοβαρότερο μειονέκτημα αυτής της μεθόδου είναι η διαστρέβλωση των συσχετίσεων και των συνδιακυμάνσεων. Συνεπώς, αυτή η μέθοδος δεν πρέπει να χρησιμοποιηθεί όταν πρόκειται να χρησιμοποιηθούν στατιστικές μέθοδοι βασισμένες είτε στις συσχετίσεις είτε στις συνδιακυμάνσεις. (Παπαδάκη, 2009)

5. Εκτίμηση Παλινδρόμησης (Regression Estimation)

Η Εκτίμηση Παλινδρόμησης αντικαθιστά τις ελλείπουσες τιμές με τις προβλεφθείσες τιμές που προσδιορίζονται από μια εξίσωση παλινδρόμησης που βασίζεται στις μεταβλητές που δεν έχουν καμία χαμένη τιμή. Οι μεταβλητές με χαμένα δεδομένα αντιμετωπίζονται ως κριτήρια και οι τιμές τους προβλέπονται από όλες τις μεταβλητές που έχουν όλα τους τα στοιχεία. Εάν τα ελλείποντα στοιχεία επιδεικνύουν μία μονοτονία και μπορούμε να κάνουμε την ρεαλιστικότερη υπόθεση, δηλαδή ότι λείπουν εντελώς *τυχαία*, η Εκτίμηση Παλινδρόμησης μπορεί να χρησιμοποιηθεί για να απλοποιήσει την εκτίμηση των παραμέτρων του πληθυσμού. Έναντι άλλων ήδη αναφερθέντων μεθόδων, η Εκτίμηση Παλινδρόμησης είναι πιο πληροφοριακή επειδή χρησιμοποιεί τις πληροφορίες που ήδη υπάρχουν σε ένα δείγμα. (Παπαδάκη, 2009)

Παρομοίως με την μέθοδο MS, η μέθοδος Εκτίμησης Παλινδρόμησης έχει το πλεονέκτημα ότι στις περιπτώσεις που έχουμε απώλεια δεδομένων το αρχικό μέγεθος δείγματος διατηρείται. Τα μειονεκτήματα της Εκτίμησης Παλινδρόμησης περιλαμβάνουν: (α) ένα μοντέλο παλινδρόμησης που πρέπει να καθοριστεί (β) οι καθ' υπολογισμό τιμές πάντα προβλέπονται σωστά από το μοντέλο παλινδρόμησης ενώ οι συσχετίσεις και οι συνδιακυμάνσεις είναι αναπόφευκτα αυξημένες (γ) μπορεί να είναι δύσκολο να εφαρμοστεί η μέθοδος της Εκτίμησης Παλινδρόμησης σε πολυμεταβλητά σύνολα δεδομένων όταν περισσότερες από μια μεταβλητές έχουν ελλείπουσες τιμές (δ) οι προβλεφθείσες τιμές μπορούν να υπερβούν το λογικό εύρος τιμών των αποτελεσμάτων για τα χαμένα δεδομένα (ε) μπορεί να απαιτήσει μεγάλα δείγματα για να παράγει σταθερές

εκτιμήσεις (στ) μπορούν να παραγάγουν τις λεπτόκυρτες (leptokurtic) κατανομές, δηλαδή τις κατανομές που έχουν μεγαλύτερο συντελεστή κύρτωσης από την κανονική κατανομή, (ζ) εάν οι καλοί εκτιμητές των ελλειπόντων δεδομένων δεν είναι διαθέσιμοι στο συνολικό δείγμα δεδομένων, τότε οι προβλεφθείσες τιμές δεν είναι καλύτερες από το μέσο όρο. Με άλλα λόγια, η μέθοδος της Εκτίμησης Παλινδρόμησης και η μέθοδος MS καταλήγουν σε παρόμοια αποτελέσματα αν ένα αποτελεσματικό μοντέλο παλινδρόμησης δεν μπορεί να προσδιοριστεί.

Για να υπερνικήσουν τον περιορισμό (β) που αναγέρθηκε παραπάνω, οι στατιστικοί έχουν προτείνει μία τροποποιημένη μέθοδο Εκτίμησης Παλινδρόμησης. Στην μέθοδο αυτή στις καθ' υπολογισμό τιμές προστίθεται ένα τυχαίο σφάλμα. Το τυχαίο σφάλμα παράγεται τυχαία από μια κανονική κατανομή με μέσο όρο 0 και τυπικό σφάλμα ίσο με την τετραγωνική ρίζα του μέσου τετραγωνικού σφάλματος του μοντέλου παλινδρόμησης. (Παπαδάκη, 2009)

Όπως αναφέρθηκε και νωρίτερα, οι μέθοδοι MS, Εκτίμησης Παλινδρόμησης καθώς και οι ειδικές μέθοδοι «λύνουν» το πρόβλημα των ελλειπόντων δεδομένων με το να αποδώσουν τις ελλείπουσες τιμές μία φορά και κατά συνέπεια αναφέρονται ως απλές μέθοδοι απόδοσης. Η απλή αυτή προσέγγιση δυστυχώς δεν απεικονίζει την αβεβαιότητα που πηγάζει από την εκτίμηση των απολεσθέντων δεδομένων. Δηλαδή το σφάλμα στην εξίσωση εκτίμησης (οποιασδήποτε μορφής) που χρησιμοποιείται για να αποδώσει κάποια τιμή στις ελλείπουσες τιμές τίθεται μηδέν. Επιπλέον, το μέγεθος του δείγματος είναι πολύ μεγαλύτερο από το πραγματικό. Επιπροσθέτως, τα διαστήματα εμπιστοσύνης για τις εκτιμώμενες παραμέτρους είναι πάρα πολύ στενά, και το σφάλμα τύπου I είναι πάρα πολύ υψηλό (Little & Rubin, 1987).

Για να παρουσιάσουμε καλύτερα αυτές τις ανεπάρκειες των μεθόδων MS και Εκτίμησης Παλινδρόμησης ας υποθέσουμε ότι το 30% των στοιχείων λείπουν και ότι επιθυμούμε να παραγάγουμε ένα διάστημα εμπιστοσύνης για έναν συντελεστή παλινδρόμησης. Η χρήση οποιεσδήποτε από αυτές τις δύο μεθόδους για να παραχθεί ένα πλήρες σετ δεδομένων σημαίνει ότι τα διαστήματα με τις ονομαστικές τιμές επιπέδων εμπιστοσύνης 90%, 95%, και 99% στην πραγματικότητα έχουν επίπεδα εμπιστοσύνης 77%, 85%, και 94%, αντίστοιχα (Rubin, 1996). Ομοίως, σε ένα πλαίσιο ελέγχου υποθέσεων που περιλαμβάνει μια μηδενική υπόθεση με δέκα παραμέτρους στις οποίες τα ελλείποντα στοιχεία έχουν αποδοθεί χρησιμοποιώντας μία από τις προαναφερθείσες μεθόδους, το στατιστικό τεστ που εκτελείται σε επίπεδο σημαντικότητας α με τιμές 10%, 5%, ή

1% στην πραγματικότητα εκτελείται σε επίπεδο σημαντικότητας 57%, 45%, και 25%, αντίστοιχα, ενώ το πρόβλημα αυτό επιδεινώνεται όσο αυξάνεται το ποσοστό των χαμένων δεδομένων στο σύνολο ή όσο αυξάνεται ο αριθμός των παραμέτρων του μοντέλου (Rubin, 1996).

2.2.4 ΑΠΛΗ ΑΠΟΔΟΣΗ ΜΕ ΑΡΧΕΣ- FIML(PRINCIPLED SINGLE IMPUTATION- FIML)

Η συντομογραφία FIML σημαίνει Full Information Maximum Likelihood ή Πλήρους Πληροφορίας Μέγιστη Πιθανοφάνεια και αντιπροσωπεύει μία πρωταρχική μέθοδο για την εκτίμηση των μέσων τιμών και των συνδιακυμάνσεων όταν αυτά βασίζονται σε ελλιπή δεδομένα και μπορούμε να υποθέσουμε ότι βρισκόμαστε στην περίπτωση MAR. Στη μέγιστη πιθανοφάνεια (ML), οι εκτιμήσεις των παραμέτρων παράγονται έτσι ώστε η πιθανοφάνεια να μεγιστοποιείται. Η μέθοδος FIML για την εκτίμηση των παραμέτρων παρουσία ελλείπων δεδομένων κάνει εκτενή χρήση του εκτιμητή μέγιστης πιθανοφάνειας. (Παπαδάκη, 2009)

Η μέθοδος FIML έχει τις ρίζες της στην εργασία των Hartley και Hocking (1971). Λαμβάνοντας υπόψη τις q ομάδες, μια για κάθε τύπο των απολεσθέντων δεδομένων, το FIML υπολογίζει αρχικά την πιθανοφάνεια για κάθε μια από τις q ομάδες. Εάν δεν υπάρχει κανένα ελλείπον στοιχείο, τότε θέτουμε $q=1$. Η πρόθεση του FIML είναι να χρησιμοποιήσει την πληροφορία από κάθε τύπο απολεσθέντων δεδομένων για να εκτιμήσει τις παραμέτρους. Στις περιπτώσεις που ισχύει η υπόθεση της πολυμεταβλητής κανονικότητας, οι q πιθανοφάνειες αθροίζονται. Το αποτέλεσμα της άθροισης των πιθανοφανειών χρησιμεύει ως η βάση για την εκτίμηση των παραμέτρων με χρήση μέγιστης πιθανοφάνειας. Η FIML μέθοδος είναι εννοιολογικά παρόμοια με την μέθοδο των q-γκρουπ των Hartley και Hocking (1971), εκτός από το ότι η πιθανοφάνεια υπολογίζεται για κάθε παρατήρηση, χρησιμοποιώντας οποιαδήποτε δεδομένα είναι διαθέσιμα για την συγκεκριμένη παρατήρηση. Στις περιπτώσεις της πολυμεταβλητής κανονικής κατανομής οι μέθοδοι MAR και FIML υπολογίζουν μια πιθανοφάνεια για κάθε περίπτωση χρησιμοποιώντας όλα τα διαθέσιμα δεδομένα. Αφότου έχουν υπολογιστεί N πιθανοφάνειες, αθροίζονται και ο εκτιμητής μέγιστης πιθανοφάνειας χρησιμοποιείται για να υπολογίσει τους μέσους, τις διασπορές και τις συνδιακυμάνσεις βασισμένες στην αθροισμένη πιθανοφάνεια. Αυτές οι εκτιμήσεις των παραμέτρων θα είναι αμερόληπτες και επαρκείς παρουσία ελλειπόντων δεδομένων υπό τις υποθέσεις MAR και πολυμεταβλητής κανονικότητας.

2.2.5 ΑΠΛΗ ΑΠΟΔΟΣΗ ΜΕ ΑΡΧΕΣ- ΕΜ ΑΛΓΟΡΙΘΜΟΣ (PRINCIPLED SINGLE IMPUTATION- EM ALGORITHM)

Ο αλγόριθμος ΕΜ αποτελεί την μεγιστοποίηση της προσδοκίας και είναι μία μέθοδος με αρχές για την επεξεργασία των ελλειπόντων στοιχείων. Ο όρος ΕΜ καθιερώθηκε από τους Dempster, Laird και Rubin στην δημοσιευμένη εργασία τους που σχετιζόταν με αυτή την μέθοδο το έτος 1977. Σύμφωνα τους Little και Rubin (1987), ο αλγόριθμος ΕΜ τυποποιεί μια σχετικά παλαιά «ad hoc» ιδέα για το χειρισμό των ελλειπόντων στοιχείων ως εξής:

1. **αντικαθιστά τις ελλείπουσες τιμές με τις κατ' εκτίμηση τιμές,**
2. **εκτιμά τις παραμέτρους,**
3. **επανεκτιμά τις ελλείπουσες τιμές υποθέτοντας ότι οι νέες εκτιμήσεις των παραμέτρων είναι σωστές,**
4. **επανεκτιμά τις παραμέτρους και ούτω καθ' εξής, επαναλαμβάνοντας την προαναφερθείσα διαδικασία μέχρι να έχουμε σύγκλιση.**

Κάθε επανάληψη του αλγορίθμου ΕΜ αποτελείται από δύο βήματα: ένα βήμα Ε (Expectation- Προσδοκία) που ακολουθείται από ένα βήμα Μ (Maximization- Μεγιστοποίηση). Στο βήμα Ε, η αναμενόμενη τιμή του λογαρίθμου της πιθανοφάνειας του πλήρους σετ δεδομένων προκύπτει, λαμβάνοντας υπόψη τα παρατηρηθέντα στοιχεία και τις κατ' εκτίμηση παραμέτρους από μια προηγούμενη επανάληψη. Στο βήμα Μ, η δεσμευμένη αναμενόμενη τιμή του λογαρίθμου της πιθανοφάνειας του πλήρους σετ δεδομένων μεγιστοποιείται. Η τιμή αυτή αυξάνεται έως ότου επιτυγχάνεται ένα στάσιμο σημείο (Dempster κ.α., 1977). Με άλλα λόγια, ο αλγόριθμος συνεχίζεται έως ότου η παρατηρηθείσα πιθανοφάνεια που παράγεται σε δύο διαδοχικές επαναλήψεις είναι σχεδόν ίδια.

2.2.6 ΠΟΛΛΑΠΛΗ ΑΠΟΔΟΣΗ (MULTIPLE IMPUTATION- MI)

Για να υπερνικηθούν οι περιορισμοί των μεθόδων που αποτυγχάνουν να λάβουν υπόψη την αβεβαιότητα που συνδέεται με τις καθ' υπολογισμό τιμές, ο Rubin και οι συνεργάτες του ανέπτυξαν την πολλαπλή μέθοδο απόδοσης (MI) στη δεκαετία του '80. Η μέθοδος MI είναι μια έγκυρη μέθοδος για τον χειρισμό των χαμένων στοιχείων στην περίπτωση MAR. Είναι μια γενικής χρήσης μέθοδος, ιδιαίτερα αποδοτική ακόμη και για μικρά μεγέθη δείγματος. Η μέθοδος MI γενικά αποτελείται από τρία βήματα: απόδοση, ανάλυση, και συγκέντρωση. Στην πρώτη φάση (απόδοση), κάθε ελλείπουσα τιμή αντικαθίσταται από όχι μια, αλλά $m > 1$ προσομοιωμένες τιμές. Οι καθ' υπολογισμό τιμές

προέρχονται από μια κατανομή που προσδιορίζεται από τον ερευνητή. Στο τέλος του πρώτου βήματος *m* πλήρη σετ δεδομένων δημιουργούνται. Στο δεύτερο βήμα (ανάλυση), κάθε ένα από τα *m* πλήρη σετ δεδομένων αναλύεται με τις τυποποιημένες μεθόδους ανάλυσης δεδομένων. Τέλος, τα αποτελέσματα των *m* αναλύσεων είναι ενσωματωμένα στο τρίτο βήμα (συγκέντρωση) για να παραγάγουν ένα τελικό αποτέλεσμα όπως μια εκτίμηση διαστήματος εμπιστοσύνης για μια παράμετρο του πληθυσμού ή για έναν έλεγχο υποθέσεων ή για ένα LRT τεστ (Likelihood Ratio Test). (Παπαδάκη, 2009)

2.2.7 Η ΣΗΜΑΣΙΑ ΤΩΝ ΣΤΑΤΙΣΤΙΚΩΝ ΥΠΟΘΕΣΕΩΝ ΓΙΑ ΤΙΣ ΜΕΘΟΔΟΥΣ ΜΕ ΑΡΧΕΣ

Οι μέθοδοι FIML, EM, και MI που περιγράφονται στα προηγούμενα τμήματα της παρούσης εργασίας στηρίζονται σε δύο βασικές υποθέσεις. Η πρώτη υπόθεση είναι ότι τα ελλείποντα στοιχεία λείπουν τυχαία. Εάν ένας ερευνητής αναμένει ότι θα έχει απώλεια δεδομένων, θα πρέπει να συλλέξει επιπρόσθετα δεδομένα που να αφορούν μεταβλητές που θεωρούνται ότι συνδέονται με τα απολεσθέντα δεδομένα ως στερεότυπο μέρος του σχεδίου μιας μελέτης, ακόμα κι αν αυτές οι μεταβλητές μπορεί να μην σχετίζονται άμεσα με την συγκεκριμένη έρευνα. Ακόμη, ένα μοντέλο ελλειπόντων δεδομένων που εξηγεί ένα μέρος της απώλειας μειώνει την μεροληψία και είναι πιθανό να παραγάγει καλύτερους εκτιμητές από αυτούς που βασίζονται στην LD. Οι μέθοδοι FIML, EM, και MI (όπως εφαρμόζονται στο πακέτο SAS® PROC MI) επίσης υποθέτουν ότι τα πλήρη στοιχεία ακολουθούν μια πολυμεταβλητή κανονική κατανομή. Πάλι, υπάρχει πολύ λίγη βιβλιογραφία που να εξετάζει τα αποτελέσματα της μη κανονικότητας στις εκτιμήσεις των παραμέτρων. (Παπαδάκη, 2009)

Στο σημείο αυτό θα πρέπει να αναφέρουμε ότι, ενώ μπορεί να φανεί λογικό πως οι δραστικές αποκλίσεις από αυτήν την υπόθεση θα οδηγήσουν σε μεροληπτικά συμπεράσματα μερικοί επιστήμονες, όπως οι Graham και Schafer (1999) διαφωνούν. Οι προαναφερθέντες υποστήριξαν ότι η μη κανονικότητα ασκεί λίγη επίδραση στα συμπεράσματα που προέρχονται από την μέθοδο MI, επειδή οι καθ' υπολογισμό τιμές θα μοιάζουν με τα παρατηρηθέντα δεδομένα στις πρώτες και δεύτερες ροπές (δηλ. στον μέσο όρο και την διασπορά). Δεδομένου ότι οι περισσότερες αναλύσεις δεδομένων για την εκπαιδευτική έρευνα είναι βασισμένες στις πρώτες και δεύτερες ροπές, μπορούμε να πούμε ότι το πρόβλημα που προκαλείται από την μη κανονικότητα δεν είναι συνήθως αρκετό για να δημιουργήσει μεροληψία στα συμπεράσματα. Παρά όμως την αισιοδοξία των Graham και Schafer, φαίνεται συνετό να χρησιμοποιηθούν οι μέθοδοι για ελλείπο-

ντα δεδομένα προσεκτικά όταν υπάρχουν ισχυρές ενδείξεις ώστε να απορρίψουμε την υπόθεση της κανονικότητας.

2.3 ΔΙΑΧΕΙΡΙΣΗ ΣΥΝΟΛΙΚΩΝ ΔΕΔΟΜΕΝΩΝ

Από τα ερωτηματολόγια, τα δεδομένα καταγράφονται και κωδικοποιούνται σε φύλλα δεδομένων ή φύλλα κωδικοποίησης και στην συνέχεια ακολουθεί η μεταφορά τους σε ειδικά πακέτα στατιστικής ανάλυσης (SPSS, MINITAB κ.α.) ή στην απλούστερη περίπτωση σε προγράμματα λογιστικών φύλλων (EXCEL). Αν ο όγκος των δεδομένων είναι μεγάλος ή η δομή τους σύνθετη τότε χρησιμοποιείται προγράμματα δημιουργίας βάσης δεδομένων όπως ACCESS.

Τα στάδια για την εισαγωγή των δεδομένων στον Η/Υ είναι τα ακόλουθα:

1. **Κωδικοποίηση** – την σωστή μεταφορά των δεδομένων στον Η/Υ.
2. **Πληκτρολόγηση** – μεταφορά δεδομένων σε ηλεκτρονική μορφή.
3. **Έλεγχος σφαλμάτων** – ανάλυση για τον εντοπισμό και την διόρθωση λαθών κατά την πληκτρολόγηση.

2.3.1 ΚΩΔΙΚΟΠΟΙΗΣΗ

Συνήθως σε μελέτες μεγάλου μεγέθους η κωδικοποίηση και πληκτρολόγηση γίνεται από ανεξάρτητες ομάδες ατόμων που καθοδηγούνται από ένα ειδικευόμενο αναλυτή. Για να αποφευχθούν τα σφάλματα και να διευκολυνθεί ο εκάστοτε αναλυτής, πριν την κωδικοποίηση πρέπει να αποφασιστούν τα ακόλουθα:

1. **Ποίες μεταβλητές θα εισαχθούν, ονόματα και σειρά**

Τα ονόματα είναι σύντομα. Επιπλέον πρέπει να θυμίζουν το περιεχόμενο του χαρακτηριστικού – μεταβλητή. Καλό θα ήταν να τεθεί ένας συγκεκριμένος κανόνας και να ακολουθηθεί σε όλες τις μεταβλητές. Τέλος θα ήταν χρήσιμο η σειρά των μεταβλητών να είναι ίδια με το ερωτηματολόγιο.

2. **Τρόποι κωδικοποίησης ποιοτικών, ποσοτικών και κατηγορικών μεταβλητών**

Πρώτα ορίζουμε τους κωδικούς – νούμερα για κάθε επίπεδο της μη ποσοτικής μεταβλητής. Όταν έχουμε δίτιμες μεταβλητές χρησιμοποιούμε την κωδικοποίηση 0-1 (επιτυχία στη στατιστική 1 – ΝΑΙ, 0-ΟΧΙ). Όταν έχουμε πολλές κατηγορικές μεταβλητές χρησιμοποιούνται κωδικοί που ξεκινούν από το 1 για το μικρότερο επίπεδο και αυξάνουν κατά μια μονάδα κάθε ανώτερο επίπεδο. Εναλλακτικά μπορούμε να χρησιμοποι-

ούμε κωδικούς συμμετρικούς στο μηδέν έτσι ώστε αρνητικές τιμές να αποδίδουν αρνητικές γνώμες.

3. Αγνοούμενες τιμές ή μη καταχωρημένες τιμές

Οι μη καταχωρημένες ή αγνοούμενες τιμές όπως αναλύθηκαν παραπάνω, μπορούν να παραμείνουν ως έχουν σε αυτό το στάδιο, οπότε και κωδικοποιούνται με τιμές 9.99 ή 999, ώστε να διαφοροποιούνται από της υπόλοιπες καταγεγραμμένες τιμές.

4. Δημιουργία μεταβλητής αναγνώρισης ερωτηματολογίου

Είναι σημαντικό να δημιουργηθεί μια μεταβλητή η οποία να αντιστοιχεί μοναδικά σε κάθε ερωτηματολόγιο – μεταβλητή. Συνήθως ονομάζεται Α.Α (Αύξων αριθμός) και ξεκινά από το 1. Ο ίδιος αριθμός θα πρέπει να σημειώνεται στα ερωτηματολόγια. Η χρησιμότητα του Α.Α. Εντοπίζεται στον έλεγχο σφαλμάτων, στην αρχειοθέτηση ερωτηματολογίων και στην διασταύρωση στοιχείων. (Ζήμερας, 2003)

2.3.2 ΕΛΕΓΧΟΣ ΣΦΑΛΜΑΤΩΝ ΠΛΗΚΤΡΟΛΟΓΗΣΗΣ

Μετά την εισαγωγή των δεδομένων στον Η/Υ, ακολουθεί ο έλεγχος της πιστότητας των δεδομένων για τυχόν εντοπισμό σφαλμάτων. Σαν ύποπτες τιμές ορίζονται εκείνες οι οποίες μπορεί να οφείλονται σε εσφαλμένη πληκτρολόγηση και μπορούν να χαρακτηρισθούν οι ακόλουθες περιπτώσεις:

1. Ακραίες τιμές

Ορίζονται ως οι τιμές οι οποίες είναι ασυνεπείς με το σύνολο των υπολοίπων τιμών. Μπορεί να οφείλονται σε λανθασμένη πληκτρολόγηση αλλά μπορεί και όχι. Στην δεύτερη περίπτωση μπορεί να χαρακτηρίζει την μη κανονικότητα της. Όταν έχουμε κατηγορικές μεταβλητές οι ακραίες τιμές οφείλονται σε λάθος πληκτρολόγηση. Όταν εντοπίσουμε ακραίες τιμές επιστρέφουμε στα ερωτηματολόγια ελέγχοντας την ορθότητά τους.

2. Αντιστροφή ψηφίων

Αντί για 34 μπορεί να πληκτρολογήσουμε 43. Τέτοια σφάλματα είναι δύσκολο να εντοπισθούν στις ποσοτικές μεταβλητές. Στις ποιοτικές ο εντοπισμός τους μπορεί να είναι πιο εύκολος ειδικά αν ο αντίστροφος κωδικός δεν ανήκει στους προεπιλεγμένους κωδικούς της μεταβλητής. Συνήθως τα σφάλματα αυτού του τύπου εντοπίζονται ως ακραίες τιμές.

3. Επαναλήψεις τιμών (Διπλοεγγραφές)

Επαναλήψεις των ίδιων αριθμών ή κωδικών είναι σύνηθες φαινόμενο. Ο εντοπισμός τους μπορεί να γίνει μόνο οπτικά και διασταυρώνονται με τα ερωτηματολόγια.

4. Λάθος καταχωρήσεις

Αναφέρονται σε καταχωρήσεις που γίνονται σε λάθος στήλες (λάθος μεταβλητή). Αυτά τα σφάλματα μπορούν να εντοπισθούν ως ακραίες τιμές αν το εύρος των τιμών των τιμών δύο διαδοχικών μεταβλητών είναι διαφορετικό.

Έλεγχος των λαθών μπορεί να γίνει με:

vii. Εκτύπωση μέσης τιμής, τυπικής απόκλισης, ελάχιστης και μέγιστης τιμής μίας μεταβλητής.

viii. Κατανομές συχνοτήτων για κάθε μεταβλητή.

ix. Εκτύπωση και έλεγχο των πληκτρολογούμενων δεδομένων κάθε μεταβλητής για ύπαρξη επαναλαμβανόμενων τιμών

x. Διασταύρωση ερωτηματολογίων και πληκτρολογούμενων δεδομένων στην περίπτωση ακραίων τιμών.

xi. Δειγματοληπτική διασταύρωση με ερωτηματολόγια.

(Ζήμερας, 2003)

2.4 ΧΡΗΣΗ ΤΕΛΙΚΩΝ ΔΕΔΟΜΕΝΩΝ

Η χρήση και η παρουσίαση των τελικών δεδομένων εξαρτάται από των εκάστοτε ερευνητή και τον κλάδο εντός του οποίου πραγματοποιείται η έρευνα. Για τα χωρικά δεδομένα και όλη την διαδικασία επεξεργασίας και παρουσίασης γίνεται εκτεταμένη αναφορά στο επόμενο κεφάλαιο της μελέτης περίπτωσης.

3. ΜΕΛΕΤΗ ΠΕΡΙΠΤΩΣΗΣ (CASE STUDY)

Σε αυτό το κεφάλαιο πραγματοποιείται η «μελέτη περίπτωσης», όπου γίνεται εφαρμογή και προσαρμογή του θεωρητικού υποβάθρου σε πραγματικά δεδομένα. Στόχος δεν είναι μόνο η εφαρμογή της θεωρίας (όπως αναλύθηκε σε προηγούμενα κεφάλαια), αλλά και ο εμπλουτισμός αυτής, με εμπειρικά συμπεράσματα και σημειώσεις, ή η παραγωγή εκ νέου εμπειρικών λύσεων για προβλήματα που εμφανίζονται στο καθαυτό πεδίο μελέτης.

3.1 ΠΡΩΤΑΡΧΙΚΑ ΔΕΔΟΜΕΝΑ ΚΑΙ ΜΕΤΑΔΕΔΟΜΕΝΑ (FREE-TEXT)

Για την ακόλουθη έρευνα χρησιμοποιήθηκε έτοιμη βάση δεδομένων καθώς η συλλογή των αξιοποιούμενων στοιχείων θα ήταν αρκετά χρονοβόρα και με υψηλό κόστος για τα πλαίσια της παρούσας διπλωματικής εργασίας. Η βάση δεδομένων περιέχει στοιχεία για μέρος των δημοσίων υπαλλήλων (από εδώ και εμπρός ΔΥ) της Ελλάδας, υπό την μορφή αρχείου EXCEL. Ακολουθεί εκτενής ανάλυση στην συνέχεια του κεφαλαίου.

3.1.1 ΠΗΓΗ ΔΕΔΟΜΕΝΩΝ

Η βάση προέρχεται από το Υπουργείο Διοικητικής Μεταρρύθμισης και Ηλεκτρονικής Διακυβέρνησης (Γενική Διεύθυνση Κατάστασης Προσωπικού Ομάδα Απογραφής) και παραχωρήθηκε στο Εργαστήριο Δημογραφικών και Κοινωνικών Αναλύσεων (ΕΔΚΑ) του Τμήματος Μηχανικών Χωροταξίας, Πολεοδομίας και Περιφερειακής Ανάπτυξης (ΤΜΧΠΠΑ) Πανεπιστημίου Θεσσαλίας, με σκοπό την περεταίρω έρευνα και αξιοποίηση των στοιχείων που συλλέχθηκαν. (Η ηλεκτρονική ή γραπτή συνολική βάση δεν παρατίθεται για λόγους δικαιωμάτων. Οι αναφορές σε αρχεία πραγματοποιείται για την διευκόλυνση των καθηγητών της τριμελής εξεταστικής επιτροπής.) Αναλυτικά η βάση παρουσιάζεται στο αντίστοιχο αρχείο «PT_all». Πραγματοποιήθηκε από το ίδιο το υπουργείο σε συνεργασία με το Υπουργείο Οικονομικών και τους ίδιους τους εργαζόμενους. Η καταγραφή των ΔΥ αποτελούσε βασικό βήμα *«για την αποτελεσματική αντιμετώπιση της παθογένειας στη Δημόσια Διοίκηση»*. (Ρέππας, 2012)

3.1.2 ΣΥΛΛΟΓΗ ΔΕΔΟΜΕΝΩΝ

Η τεχνική συλλογής δεδομένων που χρησιμοποιήθηκε ήταν on-line ερωτηματολόγια, σχεδιασμένα από τα υπουργεία, για την αποκλειστική συμπλήρωση των εργαζομένων του ελληνικού δημοσίου. Σημαντική παρατήρηση αποτελεί ότι η αξιοπιστία της βάσης είναι υψηλή, όχι μόνο επειδή πραγματοποιήθηκε από αξιόπιστους φορείς (τα αναφερόμενα υπουργεία), αλλά επίσης επειδή τα ερωτηματολόγια συμπληρώνονταν από τους ίδιους τους υπαλλήλους όντας σχεδιασμένα και προσβάσιμα ειδικά για αυτούς. Άρα

κατ' επέκταση το ποσοστό λαθών και σφαλμάτων πλησιάζει το μηδέν. Επίσης το δείγμα είναι αντιπροσωπευτικό αφού λαμβάνει σχεδόν ολόκληρο το πλήθος των ΔΥ. Τέλος σημειώνεται η ημερομηνία εξαγωγής τους η οποία είναι η 11-7-2012. (Ρέππας, 2012)

3.1.3 ΜΕΤΒΛΗΤΕΣ ΚΑΙ ΠΕΡΙΕΧΟΜΕΝΟ ΒΑΣΗΣ

Το περιεχόμενο της βάσης αποτελεί μέρος των συλλεγμένων δεδομένων, καθώς ευαίσθητα στοιχεία όπως το όνομα, το επώνυμο, ο αριθμός μητρώου κ.α. δεν επιτρέπεται να παραχωρηθούν αλλά επίσης δεν είναι και απαραίτητα για την παρούσα μελέτη. Οι εγγραφές ανέρχονται στις 136484. Η μεταβλητές που παραχωρούνται, ανέρχονται στις οχτώ και είναι οι ακόλουθες:

1. Κατηγορία ηλικίας: αποτελείται από ηλικιακές ομάδες ανά πέντε έτη. Είναι ποσοτική απαριθμητή και οι ομάδες της ανέρχονται στις 11. Παρουσιάζει missing-data (άρα 12 στο σύνολο).
2. Νομός/Δήμος: ουσιαστικά περιέχει τους δήμους (Καλλικρατικούς) κατοικίας των απογραφέντων. Είναι ποιοτική και όπως είναι λογικό οι επιλογές της ανέρχονται στις 325.
3. Κατηγορία φορέα: περιέχει τα ονόματα των κατηγοριών των αντίστοιχων φορέων όπου εργάζονται οι ΔΥ. Επίσης ποιοτική με 7 διαφορετικές τιμές.
4. Εργασιακή σχέση: ομοίως με την προηγούμενη αλλά με 4 διαφορετικές τιμές.
5. Φύλο: ποιοτική με 2 τιμές.
6. Οικογενειακή κατάσταση: ομοίως αλλά παρουσιάζει 5 διαφορετικές τιμές
7. Κατηγορία εκπαίδευση: ομοίως, όμως με 6 τιμές. Παρουσιάζει missing-data (7 στο σύνολο.
8. Άθροισμα: αυτή η μεταβλητή παρουσιάζει το πλήθος των ατόμων που παρουσιάζουν ακριβώς τα ίδια χαρακτηριστικά. Ποσοτική απαριθμητή.

Πίνακας 3.1: Τμήμα αρχικής μορφής βάσης δεδομένων

	A	B	C	D	E	F	G	H	I
1	ΚΑΤΗΓΟΡΙΑ ΗΛΙΚΙΑΣ	ΝΟΜΟΣ/ΔΗΜΟΣ	ΚΑΤΗΓΟΡΙΑ ΦΟΡΕΑ	ΕΡΓΑΣΙΑΚΗ ΣΧΕΣΗ	ΦΥΛΟ	ΟΙΚΟΓΕΝΕΙΑΚΗ ΚΑΤΑΣΤΑΣΗ	ΚΑΤΗΓΟΡΙΑ ΕΚΠΑΙΔΕΥΣΗΣ	ΑΘΡΟΙΣΜΑ	
2	20-24	ΔΗΜΟΣ ΔΙΣΤΟΜΟ'	ΦΟΡΕΙΣ ΔΗΜΟΣΙΩΝ '	ΜΟΝΙΜΟΙ ΥΠΑΛΛΗ	ΓΥΝΑΙΚΑ	ΑΓΑΜΟΣ/Η	ΔΕ	1	
3	20-24	ΔΗΜΟΣ ΔΟΜΟΚΟ'	ΦΟΡΕΙΣ ΔΗΜΟΣΙΩΝ '	ΜΟΝΙΜΟΙ ΥΠΑΛΛΗ	ΑΝΔΡΑΣ	ΑΓΑΜΟΣ/Η	ΑΝΕΥ ΚΑΤΗΓΟΡΙΑΣ ΕΚΠ/ΣΗ	1	
4	20-24	ΔΗΜΟΣ ΔΟΜΟΚΟ'	ΦΟΡΕΙΣ ΔΗΜΟΣΙΩΝ '	ΜΟΝΙΜΟΙ ΥΠΑΛΛΗ	ΑΝΔΡΑΣ	ΑΓΑΜΟΣ/Η	ΔΕ	4	
5	20-24	ΔΗΜΟΣ ΔΟΜΟΚΟ'	ΦΟΡΕΙΣ ΔΗΜΟΣΙΩΝ '	ΜΟΝΙΜΟΙ ΥΠΑΛΛΗ	ΑΝΔΡΑΣ	ΑΓΑΜΟΣ/Η	ΕΙΔΙΚΩΝ ΘΕΣΕΩΝ	8	
6	20-24	ΔΗΜΟΣ ΔΟΜΟΚΟ'	ΦΟΡΕΙΣ ΔΗΜΟΣΙΩΝ '	ΜΟΝΙΜΟΙ ΥΠΑΛΛΗ	ΓΥΝΑΙΚΑ	ΑΓΑΜΟΣ/Η	ΕΙΔΙΚΩΝ ΘΕΣΕΩΝ	2	
7	20-24	ΔΗΜΟΣ ΔΟΜΟΚΟ'	ΦΟΡΕΙΣ ΝΠΔΔ	ΜΟΝΙΜΟΙ ΥΠΑΛΛΗ	ΓΥΝΑΙΚΑ	ΑΓΑΜΟΣ/Η	ΤΕ	1	
8	20-24	ΔΗΜΟΣ ΔΟΞΑΤΟΥ	ΦΟΡΕΙΣ ΔΗΜΟΣΙΩΝ '	ΜΟΝΙΜΟΙ ΥΠΑΛΛΗ	ΑΝΔΡΑΣ	ΑΓΑΜΟΣ/Η	ΑΝΕΥ ΚΑΤΗΓΟΡΙΑΣ ΕΚΠ/ΣΗ	1	
9	20-24	ΔΗΜΟΣ ΔΟΞΑΤΟΥ	ΦΟΡΕΙΣ ΔΗΜΟΣΙΩΝ '	ΜΟΝΙΜΟΙ ΥΠΑΛΛΗ	ΑΝΔΡΑΣ	ΑΓΑΜΟΣ/Η	ΔΕ	10	
10	20-24	ΔΗΜΟΣ ΔΟΞΑΤΟΥ	ΦΟΡΕΙΣ ΔΗΜΟΣΙΩΝ '	ΜΟΝΙΜΟΙ ΥΠΑΛΛΗ	ΑΝΔΡΑΣ	ΑΓΑΜΟΣ/Η	ΕΙΔΙΚΩΝ ΘΕΣΕΩΝ	9	
11	20-24	ΔΗΜΟΣ ΔΟΞΑΤΟΥ	ΦΟΡΕΙΣ ΔΗΜΟΣΙΩΝ '	ΜΟΝΙΜΟΙ ΥΠΑΛΛΗ	ΓΥΝΑΙΚΑ	ΑΓΑΜΟΣ/Η	ΕΙΔΙΚΩΝ ΘΕΣΕΩΝ	2	
12	20-24	ΔΗΜΟΣ ΔΡΑΜΑΣ	ΦΟΡΕΙΣ ΔΗΜΟΣΙΩΝ '	ΜΟΝΙΜΟΙ ΥΠΑΛΛΗ	ΑΝΔΡΑΣ	ΑΓΑΜΟΣ/Η	ΑΝΕΥ ΚΑΤΗΓΟΡΙΑΣ ΕΚΠ/ΣΗ	6	
13	20-24	ΔΗΜΟΣ ΔΡΑΜΑΣ	ΦΟΡΕΙΣ ΔΗΜΟΣΙΩΝ '	ΜΟΝΙΜΟΙ ΥΠΑΛΛΗ	ΑΝΔΡΑΣ	ΑΓΑΜΟΣ/Η	ΔΕ	41	
14	20-24	ΔΗΜΟΣ ΔΡΑΜΑΣ	ΦΟΡΕΙΣ ΔΗΜΟΣΙΩΝ '	ΜΟΝΙΜΟΙ ΥΠΑΛΛΗ	ΑΝΔΡΑΣ	ΑΓΑΜΟΣ/Η	ΕΕΠ	1	
15	20-24	ΔΗΜΟΣ ΔΡΑΜΑΣ	ΦΟΡΕΙΣ ΔΗΜΟΣΙΩΝ '	ΜΟΝΙΜΟΙ ΥΠΑΛΛΗ	ΑΝΔΡΑΣ	ΑΓΑΜΟΣ/Η	ΕΙΔΙΚΩΝ ΘΕΣΕΩΝ	47	
16	20-24	ΔΗΜΟΣ ΔΡΑΜΑΣ	ΦΟΡΕΙΣ ΔΗΜΟΣΙΩΝ '	ΜΟΝΙΜΟΙ ΥΠΑΛΛΗ	ΑΝΔΡΑΣ	ΑΓΑΜΟΣ/Η	ΠΕ	2	
17	20-24	ΔΗΜΟΣ ΔΡΑΜΑΣ	ΦΟΡΕΙΣ ΔΗΜΟΣΙΩΝ '	ΜΟΝΙΜΟΙ ΥΠΑΛΛΗ	ΑΝΔΡΑΣ	ΑΓΑΜΟΣ/Η	ΤΕ	1	
18	20-24	ΔΗΜΟΣ ΔΡΑΜΑΣ	ΦΟΡΕΙΣ ΔΗΜΟΣΙΩΝ '	ΜΟΝΙΜΟΙ ΥΠΑΛΛΗ	ΑΝΔΡΑΣ	ΕΓΓΑΜΟΣ/Η	ΔΕ	1	
19	20-24	ΔΗΜΟΣ ΔΡΑΜΑΣ	ΦΟΡΕΙΣ ΔΗΜΟΣΙΩΝ '	ΜΟΝΙΜΟΙ ΥΠΑΛΛΗ	ΑΝΔΡΑΣ	ΕΓΓΑΜΟΣ/Η	ΕΙΔΙΚΩΝ ΘΕΣΕΩΝ	3	
20	20-24	ΔΗΜΟΣ ΔΡΑΜΑΣ	ΦΟΡΕΙΣ ΔΗΜΟΣΙΩΝ '	ΜΟΝΙΜΟΙ ΥΠΑΛΛΗ	ΓΥΝΑΙΚΑ	ΑΓΑΜΟΣ/Η	ΔΕ	5	
21	20-24	ΔΗΜΟΣ ΔΡΑΜΑΣ	ΦΟΡΕΙΣ ΔΗΜΟΣΙΩΝ '	ΜΟΝΙΜΟΙ ΥΠΑΛΛΗ	ΓΥΝΑΙΚΑ	ΑΓΑΜΟΣ/Η	ΕΙΔΙΚΩΝ ΘΕΣΕΩΝ	14	
22	20-24	ΔΗΜΟΣ ΔΡΑΜΑΣ	ΦΟΡΕΙΣ ΝΠΔΔ	ΙΔΙΩΤΙΚΟΥ ΔΙΚΑΙΟΥ	ΑΝΔΡΑΣ	ΑΓΑΜΟΣ/Η	ΔΕ	1	
23	20-24	ΔΗΜΟΣ ΔΡΑΜΑΣ	ΦΟΡΕΙΣ ΝΠΔΔ	ΙΔΙΩΤΙΚΟΥ ΔΙΚΑΙΟΥ	ΓΥΝΑΙΚΑ	ΑΓΑΜΟΣ/Η	ΠΕ	1	
24	20-24	ΔΗΜΟΣ ΔΥΤΙΚΗΣ Α	ΦΟΡΕΙΣ ΔΗΜΟΣΙΩΝ '	ΙΔΙΩΤΙΚΟΥ ΔΙΚΑΙΟΥ	ΑΝΔΡΑΣ	ΑΓΑΜΟΣ/Η	ΔΕ	1	
25	20-24	ΔΗΜΟΣ ΔΥΤΙΚΗΣ Α	ΦΟΡΕΙΣ ΔΗΜΟΣΙΩΝ '	ΜΟΝΙΜΟΙ ΥΠΑΛΛΗ	ΑΝΔΡΑΣ	ΑΓΑΜΟΣ/Η	ΑΝΕΥ ΚΑΤΗΓΟΡΙΑΣ ΕΚΠ/ΣΗ	1	
26	20-24	ΔΗΜΟΣ ΔΥΤΙΚΗΣ Α	ΦΟΡΕΙΣ ΔΗΜΟΣΙΩΝ '	ΜΟΝΙΜΟΙ ΥΠΑΛΛΗ	ΑΝΔΡΑΣ	ΑΓΑΜΟΣ/Η	ΔΕ	13	
27	20-24	ΔΗΜΟΣ ΔΥΤΙΚΗΣ Α	ΦΟΡΕΙΣ ΔΗΜΟΣΙΩΝ '	ΜΟΝΙΜΟΙ ΥΠΑΛΛΗ	ΑΝΔΡΑΣ	ΑΓΑΜΟΣ/Η	ΕΙΔΙΚΩΝ ΘΕΣΕΩΝ	17	
28	20-24	ΔΗΜΟΣ ΔΥΤΙΚΗΣ Α	ΦΟΡΕΙΣ ΔΗΜΟΣΙΩΝ '	ΜΟΝΙΜΟΙ ΥΠΑΛΛΗ	ΑΝΔΡΑΣ	ΑΓΑΜΟΣ/Η	ΠΕ	2	
29	20-24	ΔΗΜΟΣ ΔΥΤΙΚΗΣ Α	ΦΟΡΕΙΣ ΔΗΜΟΣΙΩΝ '	ΜΟΝΙΜΟΙ ΥΠΑΛΛΗ	ΑΝΔΡΑΣ	ΑΓΑΜΟΣ/Η	ΤΕ	2	

Πηγή: Υπουργείο Διοικητικής Μεταρρύθμισης και Ηλεκτρονικής Διακυβέρνησης, 2012

Όλες οι μεταβλητές αποτελούνται από κείμενο με μόνη εξαίρεση την μεταβλητή «Άθροισμα». Συγκεντρωτικά παρουσιάζονται οι βασικές τιμές των κυρίως μεταβλητών:

Πίνακας 3.2: Τιμές μεταβλητών

Κατηγορία ηλικίας	Δήμος κατοικίας	Τομέας	Εργασιακή σχέση	Φύλο	Οικογενειακή κατάσταση	Κατηγορία εκπαίδευσης
20-24	(Οι 325 δήμοι του προγράμματος Καλλικράτης)	ΝΠΔΔ ΤΩΝ ΟΤΑ	ΙΔΙΩΤΙΚΟΥ ΔΙΚΑΙΟΥ ΑΟΡΙΣΤΟΥ ΧΡΟΝΟΥ	ΑΝΔΡΑΣ	ΑΓΑΜΟΣ/Η	ΔΕ
25-29		ΟΤΑ Α ΒΑΘΜΟΥ (ΚΑΛΛΙΚΡΑΤΗΣ)	ΙΔΙΩΤΙΚΟΥ ΔΙΚΑΙΟΥ ΟΡΙΣΜΕΝΟΥ ΧΡΟΝΟΥ	ΓΥΝΑΙΚΑ	ΕΓΓΑΜΟΣ/Η	ΕΕΠ
30-34		ΟΤΑ Β ΒΑΘΜΟΥ	ΜΟΝΙΜΟΙ ΥΠΑΛΛΗΛΟΙ ΤΟΥ ΔΗΜΟΣΙΟΥ /ΔΙΚΑΣΤΙΚΟΙ ΛΕΙΤΟΥΡΓΟΙ /ΔΗΜΟΣΙΟΙ ΛΕΙΤΟΥΡΓΟΙ		ΔΙΑΖΕΥΓΜΕΝΟΣ/Η	ΕΙΔΙΚΩΝ ΘΕΣΕΩΝ
35-39		ΥΠΟΥΡΓΕΙΑ	ΣΥΜΒΑΣΙΟΥΧΟΙ ΕΡΓΟΥ		ΣΕ ΔΙΑΣΤΑΣΗ	ΠΕ
40-44		ΦΟΡΕΙΣ ΔΗΜΟΣΙΩΝ ΥΠΗΡΕΣΙΩΝ			ΧΗΡΟΣ/Α	ΤΕ
45-49		ΦΟΡΕΙΣ ΝΠΔΔ				ΥΕ
50-54		ΦΟΡΕΙΣ ΟΤΑ				
55-59						
60-64						
65-69						
70-74						

Πηγή: Υπουργείο Διοικητικής Μεταρρύθμισης και Ηλεκτρονικής Διακυβέρνησης, 2012

Σημαντική παρατήρηση: από την τελευταία μεταβλητή γίνεται αμέσως αντιληπτό πως οι 136484 εγγραφές δεν είναι το πλήθος των ΔΥ. Η βάση είναι ομαδοποιημένη και τα πραγματικά άτομα που καταγράφηκαν (άθροισμα της μεταβλητής «Άθροισμα») είναι 640759. Η μεταβλητή «Άθροισμα» ουσιαστικά αποτελεί το στατιστικό βάρος των υπολοίπων μεταβλητών.

3.1.4 ΈΛΕΓΧΟΣ ΔΕΔΟΜΕΝΩΝ

Λόγω της αξιόπιστης πηγής των δεδομένων και του είδους των δεδομένων (κείμενο), δεν σημειώνονται σφάλματα όπως αυτά αναλύθηκαν στα προηγούμενα κεφάλαια (π.χ. ακραίες τιμές, λάθος καταχωρίσεις). Όμως παρατηρούνται missing-values. Εμφανίζονται μόνο σε δύο μεταβλητές, στην κατηγορία ηλικίας και στην κατηγορία εκπαίδευσης. Σημειώνουν όμως μηδαμινά ποσοστά επί του συνόλου έκαστης (εκπαίδευση: 3,7 και ηλικία : 0,7) και για αυτόν τον λόγο η μέθοδος που επιλέχθηκε για την αντιμετώπιση τους περιλαμβάνει την διατήρηση των κενών σημείων σαν υπάρχουσες τιμές μεταβλητών και η επεξεργασία τους ως τέτοιες.

3.2 ΚΩΔΙΚΟΠΟΙΗΣΗ ΔΕΔΟΜΕΝΩΝ

Ακολουθώντας την εμπειρία και την ανάλυση που πραγματοποιήθηκε στο κεφάλαιο δύο του παρόντος, παρουσιάζεται η κωδικοποίηση της βάσης δεδομένων. Αυτό το στάδιο είναι αναγκαίο όχι μόνο για την ευκολότερη χρήση των δεδομένων από τους ερευνητές και τα στατιστικά προγράμματα, αλλά επίσης μειώνει τις απαιτήσεις σε χωρητικότητα και υπολογιστική ισχύ, επιταχύνοντας κατά αυτόν τον τρόπο την επεξεργασία τους. Αρχικά η ίδια η κωδικοποίηση είναι αρκετά χρονοβόρα, αλλά τα προτερήματά της στην μετέπειτα μελέτη και επεξεργασία των δεδομένων είναι εκπληκτικά.

3.2.1 ΒΑΣΙΚΕΣ ΑΡΧΕΣ ΚΩΔΙΚΟΠΟΙΗΣΗΣ

Βασικές αρχές που ακολουθήθηκαν κατά την κωδικοποίηση είναι οι ακόλουθες:

1. Τα ονόματα των μεταβλητών αποτελούνται από αριθμούς και λατινικούς χαρακτήρες, ώστε να αναγνωρίζονται από τα διάφορα προγράμματα επεξεργασίας. Συνήθως τέτοια προγράμματα δεν έχουν υλοποιηθεί από ελληνικές εταιρίες και λίγες φορές έχουν επεκτάσεις για την ελληνική ή άλλες γλώσσες πέρα της αγγλικής.
2. Οι ονομασίες είναι είτε μικρές/περιορισμένες λέξεις/φράσεις είτε συντομογραφίες για διευκόλυνση του χρήστη κατά την επεξεργασία.
3. Αποφεύγονται τα κενά και οι περίεργοι συμβολισμοί (όπως &, # κ.α.), ομοίως για τους παραπάνω λόγους (τα προγράμματα τις περισσότερες φορές δεν τα αναγνωρίζουν και οι χρήστες καθυστερούνται κατά την επεξεργασία και πληκτρολόγηση περίπλοκων αυτοματοποιημένων εντολών).
4. Οι τιμές των μεταβλητών κωδικοποιούνται με αριθμούς ώστε να επεξεργάζονται ευκολότερα από τα στατιστικά προγράμματα.
5. Αποφεύγεται η χρήση του 0 (μηδέν) καθώς θα μπορούσε να «αναγνωριστεί» από προγράμματα ή μελετητές ως missing value.
6. Σε περίπτωση επίσημων κωδικών από οργανισμούς, κρατικά μέσα (κ.α.) είναι σημαντικό να ακολουθούμε αυτήν την μορφή αφού αργότερα τα δεδομένα και η μελέτη γίνονται ταχύτερα αντιληπτά από άλλους μελετητές ή το ευρύ κοινό (για την παρούσα περίπτωση οι κωδικοί των δήμων από το πρόγραμμα Καλλικράτης, πηγή Υπουργείο Εσωτερικών, 2012).

3.2.2 ΤΕΛΙΚΕΣ ΤΙΜΕΣ ΚΑΙ ΕΠΙΠΡΟΣΘΕΤΑ ΜΕΤΑΔΕΔΟΜΕΝΑ (FREE-TEXT)

Οι τελικές μορφές των μεταβλητών και των τιμών τους μετά την κωδικοποίηση παρουσιάζονται συγκεντρωτικά στον παρακάτω πίνακα:

Πίνακας 3.3: Κωδικοποιημένες και μη μεταβλητές και τιμές αυτών

age	ota_c	org	empl	sex	fam	edu
Κατηγορία ηλικίας	Δήμος κατοικίας	Τομέας	Εργασιακή σχέση	Φύλο	Οικογενειακή κατάσταση	Κατηγορία εκπαίδευσης
20-24=1	(Οι 325 δήμοι του προγράμματος Καλλικράτης με την επίσημη κωδικοποίηση)	ΝΠΔΔ ΤΩΝ ΟΤΑ=1	ΙΔΙΩΤΙΚΟΥ ΔΙΚΑΙΟΥ ΑΟΡΙΣΤΟΥ ΧΡΟΝΟΥ=1	ΑΝΔΡΑΣ=1	ΑΓΑΜΟΣ/Η=1	ΔΕ=1
25-29=2		ΟΤΑ Α ΒΑΘΜΟΥ (ΚΑΛΛΙΚΡΑΤΗΣ)=2	ΙΔΙΩΤΙΚΟΥ ΔΙΚΑΙΟΥ ΟΡΙΣΜΕΝΟΥ ΧΡΟΝΟΥ=2	ΓΥΝΑΙΚΑ=2	ΕΓΓΑΜΟΣ/Η=2	ΕΕΠ=2
30-34=3		ΟΤΑ Β ΒΑΘΜΟΥ=3	ΜΟΝΙΜΟΙ ΥΠΑΛΛΗΛΟΙ ΤΟΥ ΔΗΜΟΣΙΟΥ /ΔΙΚΑΣΤΙΚΟΙ ΛΕΙΤΟΥΡΓΟΙ /ΔΗΜΟΣΙΟΙ ΛΕΙΤΟΥΡΓΟΙ=3		ΔΙΑΖΕΥΓΜΕΝΟΣ/Η=3	ΕΙΔΙΚΩΝ ΘΕΣΕΩΝ=3
35-39=4		ΥΠΟΥΡΓΕΙΑ=4	ΣΥΜΒΑΣΙΟΥΧΟΙ ΕΡΓΟΥ=4		ΣΕ ΔΙΑΣΤΑΣΗ=4	ΠΕ=4
40-44=5		ΦΟΡΕΙΣ ΔΗΜΟΣΙΩΝ ΥΠΗΡΕΣΙΩΝ=5			ΧΗΡΟΣ/Α=5	ΤΕ=5
45-49=6		ΦΟΡΕΙΣ ΝΠΔΔ=6				ΥΕ=6
50-54=7		ΦΟΡΕΙΣ ΟΤΑ=7				Missing-value=99
55-59=8						
60-64=9						
65-69=10						
70-74=11						
Missing-value=99						

Πηγή: Υπουργείο Διοικητικής Μεταρρύθμισης και Ηλεκτρονικής Διακυβέρνησης, 2012 και Υπουργείο Εσωτερικών, 2012

Αναλυτικά η κωδικοποιημένη βάση παρουσιάζεται στο αρχείο «bd_omad_cod». Ακολουθεί τμήμα αυτής ως παράδειγμα.

Πίνακας 3.4: Τμήμα κωδικοποιημένης βάσης

	A	B	C	D	E	F	G	H
1	age	ota_c	org	empl	sex	fam	edu	sum
2	1	9159	5	3	1	1	99	1
3	1	9001	5	3	1	1	99	1
4	1	9002	5	3	1	1	99	6
5	1	9131	5	3	1	1	99	1
6	1	9253	5	3	1	1	99	1
7	1	9015	5	3	1	1	99	2
8	1	9097	5	3	1	1	99	1
9	1	9179	5	3	2	1	99	1
10	1	9194	5	3	1	1	99	2
11	1	9194	5	3	1	2	99	1
12	1	9180	5	3	1	1	99	3
13	1	9180	5	3	2	1	99	4
14	1	9123	5	3	1	1	99	6
15	1	9123	5	3	2	1	99	1
16	1	9186	1	4	1	1	99	1
17	1	9186	2	4	1	1	99	1
18	1	9186	5	3	1	1	99	37
19	1	9186	5	3	1	2	99	1
20	1	9186	5	3	2	1	99	12
21	1	9181	5	2	1	1	99	2

Πηγή: Υπουργείο Διοικητικής Μεταρρύθμισης και Ηλεκτρονικής Διακυβέρνησης, 2012 και Υπουργείο Εσωτερικών, 2012

3.3 ΑΞΙΟΠΟΙΗΣΗ ΣΤΟΙΧΕΙΩΝ ΒΑΣΗΣ

Σύμφωνα με τα στοιχεία που περιέχει η βάση δεδομένων είναι δυνατόν να δημιουργηθεί το προφίλ των ΔΥ και να αξιοποιηθούν οι χωρικές πληροφορίες της για την παρουσίαση χωρικών φαινομένων και χαρακτηριστικών του δείγματος. Όμως, στην παρούσα μορφή της, δηλαδή ομαδοποιημένη σύμφωνα με το πλήθος των ατόμων που παρουσιάζουν κοινά χαρακτηριστικά (μεταβλητή «sum»), δεν είναι δυνατή η πραγματοποίηση των παραπάνω. Για αυτό και αναπτύσσεται η παρακάτω εμπειρική μεθοδολογία.

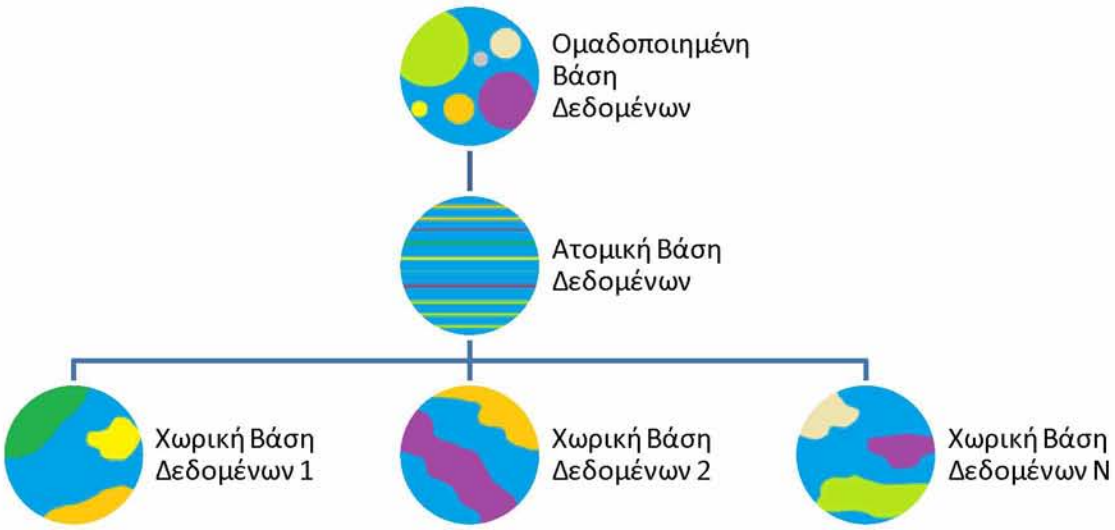
3.3.1 ΑΝΑΠΤΥΞΗ ΝΕΑΣ ΕΜΠΕΙΡΙΚΗΣ ΜΕΘΟΔΟΛΟΓΙΑΣ ΚΑΙ ΠΡΑΚΤΙΚΩΝ

Αρχικά, με στόχο την δημιουργία του προφίλ των δημοσίων υπαλλήλων, η βάση θα μετατραπεί από «ομαδοποιημένη» σε «ατομική». Αυτό περιλαμβάνει την ανάπτυξη του πλήθους από 136484 εγγραφές σε 640759. Δηλαδή ουσιαστικά θα καταργηθεί η μεταβλητή sum και πλέον κάθε παρατήρηση θα περιέχει μόνο ένα καταγεγραμμένο άτομο. Οι υπόλοιπες μεταβλητές θα παραμείνουν ως έχουν.

Αποτελεί το πιο χρονοβόρο βήμα της μελέτης αλλά είναι απαραίτητο για αυτήν. Η νέα ατομική βάση δεδομένων πλέον θα είναι έτοιμη για χρήση, ώστε να παρουσιαστεί το προφίλ των ΔΥ.

Αυτό όμως δεν επαρκεί για τις εξειδικευμένες μελέτες άλλων κλάδων πέρα της στατιστικής και της οικονομετρίας. Σε ότι αφορά την μελέτη της βάσης για σκοπούς των επιστημών του χώρου, θα πρέπει να πραγματοποιηθεί ακόμα ένα βήμα. Το τελικό βήμα είναι η επανα-ομαδοποίηση της βάσης ώστε η χωρική της υπόσταση να είναι χρήσιμη. Άρα θα ομαδοποιηθεί αυτήν την φορά σύμφωνα με την μεταβλητή του «Δήμου κατοικίας/ota_c». Αυτό θα έχει σαν αποτέλεσμα η νέα «χωρική βάση» να έχει 325 εγγραφές (ο αριθμός των δήμων σύμφωνα με το πρόγραμμα Καλλικράτη, δίχως το Άγιο Όρος). Εδώ όμως σημειώνεται ότι λόγω των πολλαπλών μεταβλητών ουσιαστικά δεν θα παραχθεί μία χωρική βάση, αλλά υπάρχει η δυνατότητα για άπυρους συνδυασμούς των προηγούμενων μεταβλητών σε νέες, άρα και κατ' επέκταση για άπειρες χωρικές βάσεις.. Κατά αυτόν τον τρόπο θα είναι δυνατόν να δίνονται απαντήσεις σε πολλά ερωτήματα χωρικού περιεχομένου με παραγωγή των αντίστοιχων χωρικών βάσεων και την παρουσίαση αυτών σε χάρτες.

Διάγραμμα 3.1: Μεθοδολογία χρήσης βάσης



Αναλυτικότερη επεξήγηση ακολουθεί παρακάτω.

3.3.2 ΑΤΟΜΙΚΗ ΒΑΣΗ ΚΑΙ ΠΡΟΦΙΛ ΔΗΜΟΣΙΩΝ ΥΠΑΛΛΗΛΩΝ

Για να δημιουργηθεί η ατομική βάση γίνεται χρήση ενός προγράμματος στατιστικής ή όπως επιλέχθηκε εδώ, υπολογιστικών φύλων (EXCEL). Με απλές εντολές αντιγραφής ή αυτόματης συμπλήρωσης μετατρέπουμε τις 136484 εγγραφές σε 640759 και καταργούμε την μεταβλητή sum. (αρχείο ατομικής βάσης «atomikh_bd»)

Πίνακας 3.5: Τμήμα ατομικής βάσης

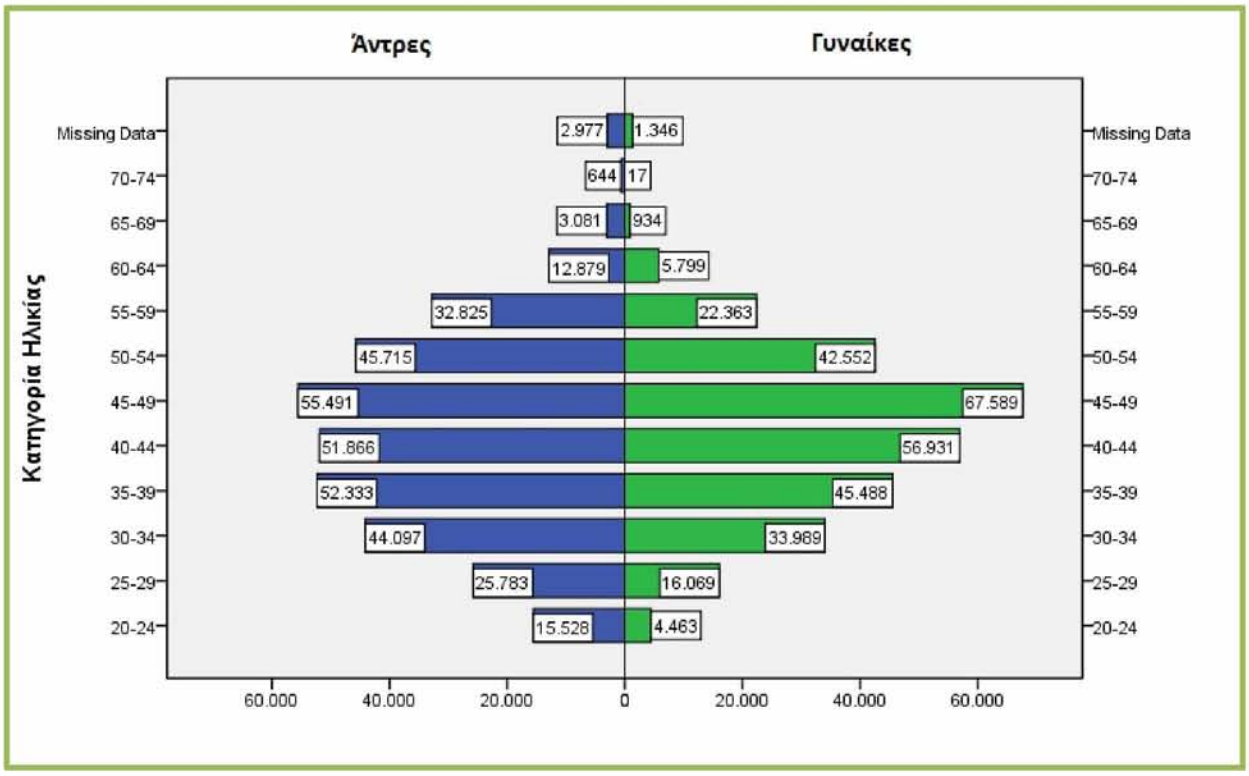
	A	B	C	D	E	F	G
1	age	ota_c	org	empl	sex	fam	edu
2	1	9001	5	3	1	1	99
3	2	9001	5	3	1	1	99
4	3	9001	5	2	2	2	99
5	3	9001	5	3	1	1	99
6	4	9001	5	3	1	1	99
7	4	9001	5	3	2	2	99
8	7	9001	5	2	1	2	99
9	7	9001	5	3	1	2	99
10	8	9001	2	1	2	2	99
11	8	9001	5	3	1	2	99
12	9	9001	6	3	1	2	99
13	6	9001	5	3	1	2	99
14	6	9001	5	3	1	2	99
15	3	9001	5	3	1	2	99
16	5	9001	5	3	1	2	99
17	3	9001	5	3	1	2	99
18	5	9001	5	3	1	2	99
19	3	9001	5	3	1	2	99
20	5	9001	5	3	1	2	99
21	3	9001	1	3	2	2	6
22	3	9001	2	1	1	2	6
23	3	9001	2	3	2	2	6
24	4	9001	1	1	2	1	6
25	4	9001	1	3	1	2	6
26	4	9001	2	1	1	2	6
27	4	9001	2	1	2	3	6
28	4	9001	2	1	2	2	6
29	4	9001	2	3	1	3	6
30	4	9001	5	3	1	1	6

Πηγή: Υπουργείο Διοικητικής Μεταρρύθμισης και Ηλεκτρονικής Διακυβέρνησης, 2012 και Υπουργείο Εσωτερικών, 2012

Σύμφωνα με τα στοιχεία που υπάρχουν σε αυτήν την φάση ανάλυσης είναι εφικτό να δημιουργήσουμε το προφίλ των δημοσίων υπαλλήλων. Με χρήση στατιστικού προγράμματος (εδώ SPSS) μπορούμε να συμπεράνουμε τα ακόλουθα:

- Πάνω από το 50% των εργαζομένων είναι άντρες (53,6%), οπότε υπάρχει μία σχετική ισορροπία μεταξύ αυτών και των γυναικών (υπόλοιπο 46,4%).
- Όσον αφορά την οικογενειακή κατάσταση το συντριπτικό ποσοστό του 67,6% αποτελεί τους έγγαμους/ες, ακολουθεί ο άγαμος/μη με σχεδόν 26% (ακριβές: 25,9%) και τέλος ακολουθούν οι κατηγορίες διαζευγμένος/νη, χήρος/α και «σε διάσταση», αντίστοιχα με 4% , 1,1% και 1%.
- Ένα σχετικά θετικό σημείο εμφανίζεται στην εκπαίδευση, αφού σχεδόν ο ένας στους δύο (40,5%) έχει πανεπιστημιακή εκπαίδευση. Έπειτα ακολουθούν οι απόφοιτη της δευτεροβάθμιας εκπαίδευσης (27,3%), και οι υπόλοιπες κατηγορίες με σειρά: ΤΕ (9,4%), ΥΕ (8,1%), ΕΕΠ (0,1%) και όσοι δεν δήλωσαν (3,7%).
- Επίσης, τα ποσοστά των κατηγοριών τομέα στον οποίο εργάζονται οι ΔΥ έχουν την ακόλουθη κατάταξη: φορείς δημόσιων υπηρεσιών (54,9%), φορείς ΝΠΔΔ (23,5%), ΟΤΑ Α βαθμού – Καλλικράτης (12,5%), υπουργεία (4,4%), ΟΤΑ Β βαθμού (2,3%), ΝΠΔΔ των ΟΤΑ (2,2%), τέλος, φορείς ΟΤΑ (0,1%).
- Συνεχίζοντας, εξετάζεται η εργασιακή σχέση των υπαλλήλων, όπου οι μόνιμοι υπάλληλοι του δημοσίου/δικαστικοί λειτουργεί/δημόσιοι λειτουργοί όπως ήταν αναμενόμενο κατέχουν την πρώτη θέση και με μεγάλη μάλιστα διαφορά (86,4%), συνεχίζουν οι ιδιωτικού δικαίου αορίστου χρόνου (8,1%), οι ιδιωτικού δικαίου ορισμένου χρόνου (4,7%) και τελευταίοι οι συμβασιούχοι έργου (0,7%).
- Ολοκληρώνοντας το προφίλ των ΔΥ παρατίθεται η παρακάτω πληθυσμιακή πυραμίδα για τις ηλικίες:

Διάγραμμα 3.2: Ηλικιακή πυραμίδα ΔΥ



Πηγή: Υπουργείο Διοικητικής Μεταρρύθμισης και Ηλεκτρονικής Διακυβέρνησης, 2012

Στο διάγραμμα αυτό παρατηρούμε πως ο πληθυσμός είναι σχετικά μεγάλος σε ηλικία αλλά όχι γερασμένος (πιθανόν λόγω ορίων συνταξιοδότησης). Επίσης υπάρχουν μικρές προσλήψεις νέων, πιθανόν λόγω της νέας πολιτικής στα πλαίσια της «κρίσης» για περιορισμό των ΔΥ. Τέλος, στους άντρες έχουμε μεγαλύτερα ποσοστά σε σχέση με τις γυναίκες με μόνη εξαίρεση τις ηλικιακές ομάδες 40-44 και 45-49. Υποθέτοντας πως υπάρχουν ανώτατα όρια για προσλήψεις, εδώ αιτία ίσως αποτελούν νέες διατάξεις ή τροποποίηση των υπαρχόντων ώστε να ευνοούν της γυναίκες υπαλλήλους, όπως πραγματοποιήθηκαν πριν από τα αντίστοιχα έτη.

Η ανάλυση όμως δεν τελειώνει εδώ. Ως αναλυτές του χώρου πρέπει να προχωρήσουμε στο επόμενο βήμα, δηλαδή στις χωρικές βάσεις δεδομένων.

3.3.3 ΧΩΡΙΚΕΣ ΒΑΣΕΙΣ ΚΑΙ ΖΗΤΗΜΑΤΑ ΤΟΥ ΧΩΡΟΥ

Για να παρουσιάσουμε τα χωρικά δεδομένα και για να μπορούμε να απαντήσουμε σε ερωτήματα που αφορούν στην χωρική υπόσταση των αναλύσεων, μπορούμε πλέον να περάσουμε από την ατομική βάση σε νέες χωρικές βάσεις. Οι βάσεις αυτές μπορούν να διαμορφωθούν έτσι ώστε να αναδεικνύουν χωρικά στοιχεία και να απαντούν σε συγκεκριμένα ερωτήματα για τα χωρικά φαινόμενα.

Για να αναδείξουμε την χωρική πληροφορία, ουσιαστικά ομαδοποιούμε τις μεταβλητές από την ατομική βάση σύμφωνα με τις τιμές της μεταβλητής "ota_c" , δηλαδή δημιουργούμε νέες μεταβλητές, οι οποίες πλέον δεν παίρνουν τιμές όπως οι παλιές, αλλά αριθμό ατόμων που εντοπίζονται στον αντίστοιχο δήμο "ota_c" (π.χ. 9305=Ηράκλειο) και παρουσιάζουν κάποιο χαρακτηριστικό, είτε παλιάς μεταβλητής όπως το "sex" (π.χ. 1=Άντρας), ή νέας σύνθετης μεταβλητής όπως το "sex_age_fin_sex2_age_10" (παράγωγο δύο διαφορετικών μεταβλητών) που παρουσιάζει το πλήθος τον ατόμων γένους θηλυκού στην ηλικιακή ομάδα 65-69. Ακολουθούν εκτενέστερα πιο αναλυτικά παραδείγματα.

Στις νέες αυτές βάσεις όπως αναφέρθηκε παραπάνω μπορούμε να έχουμε είτε μία από τις παλιές μεταβλητές είτε περισσότερες τροποποιημένες σε νέες μεταβλητές σύμφωνα με το πλήθος των τιμών έκαστης. Έτσι μπορούμε να δημιουργήσουμε άπειρες χωρικές βάσεις όσων «διαστάσεων» χρειαζόμαστε. Στα πλαίσια αυτής της μελέτης πραγματοποιείται η παραγωγή χωρικής βάσης με «μονοδιάστατες, δισδιάστατες και τρισδιάστατες» μεταβλητές. Σε κάθε κατηγορία παρουσιάζονται παραδείγματα. Οι ίδια μεθοδολογία μπορεί να επαναλαμβάνεται συστηματικά για την δημιουργία και άλλων χωρικών βάσεων.

- **1η Κατηγορία: Χωρική βάση δεδομένων με απλές μεταβλητές (ποσοστά και πλήθη ανά φύλο).**

Χρησιμοποιώντας ένα στατιστικό πρόγραμμα (εδώ SPSS) μπορούμε με διάφορες εντολές (π.χ. AGGREGATE) να δημιουργήσουμε της νέες μεταβλητές. Παράδειγμα εδώ αποτελεί η χωρική βάση με τα ποσοστά και τους πληθυσμούς των φύλων (αρχείο «bd_spatial_sex»), η οποία περιέχει την χωρική μεταβλητή ota_c (δήμοι) και τις νέες μεταβλητές sex_fin_1 (ποσοστό αντρών στο δήμο x), sex_fin_2 (ποσοστό γυναικών στο δήμο x), pop_t (συνολικό πληθυσμό στο δήμο x), pop_sex_1 (πληθυσμό αντρών στο δήμο x) και pop_sex_2 (πληθυσμό γυναικών στο δήμο x).

Πίνακας 3.6: Τμήμα χωρικής βάσης ποσοστών και πληθυσμών ανά φύλο

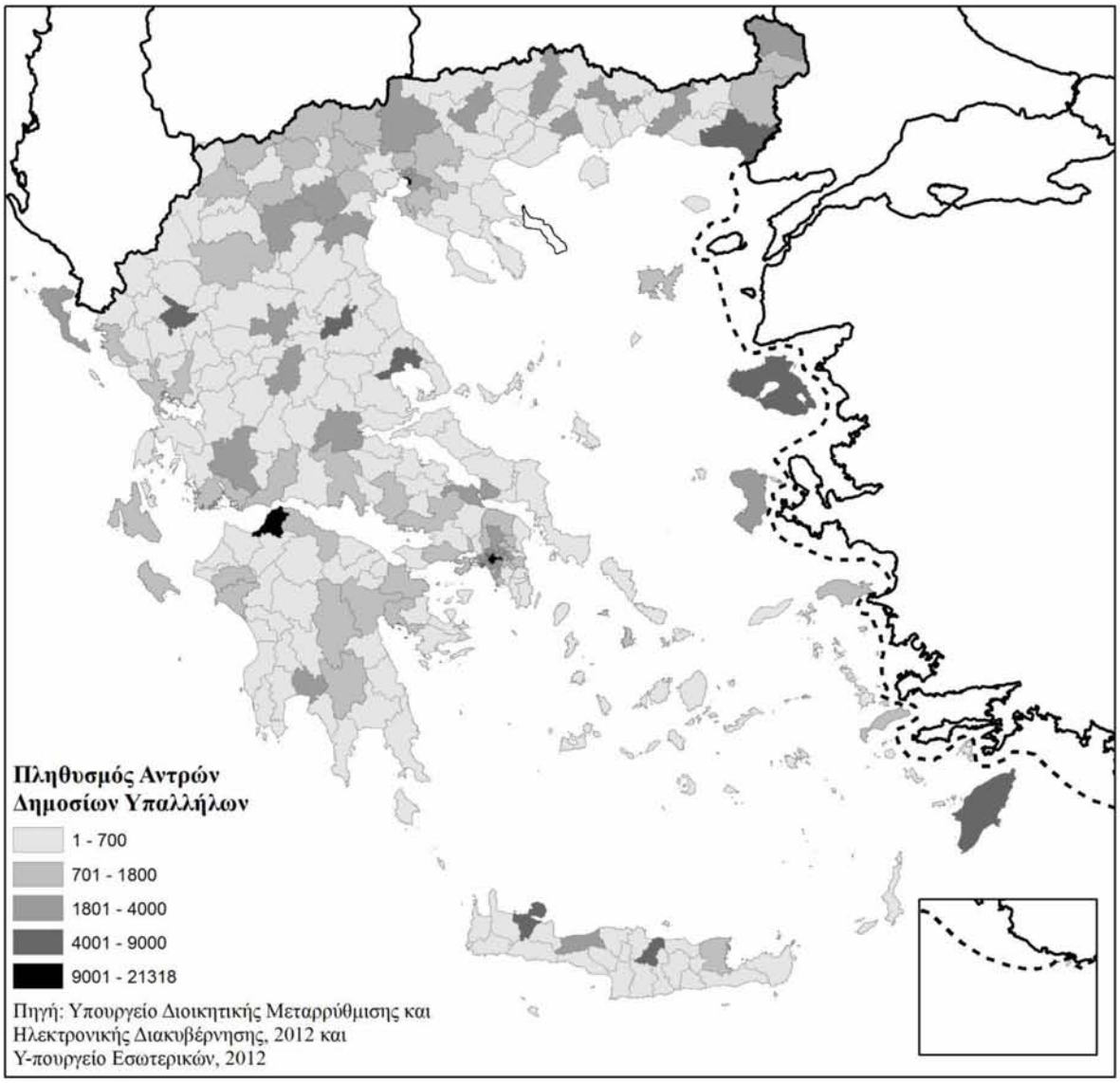
	ota_c	sex_fin1	sex_fin2	pop_t	pop_sex_1	pop_sex_2
1	9001	0,625	0,375	557	348	209
2	9002	0,531	0,469	4305	2287	2018
3	9003	0,651	0,349	195	127	68
4	9004	0,642	0,358	123	79	44
5	9005	0,621	0,379	457	284	173
6	9006	0,608	0,392	8101	4927	3174
7	9007	0,714	0,286	1723	1231	492
8	9008	0,741	0,259	3395	2515	880
9	9009	0,717	0,283	244	175	69
10	9010	0,748	0,252	1083	810	273
11	9011	0,618	0,382	427	264	163
12	9012	0,481	0,519	4554	2190	2364
13	9013	0,591	0,409	785	464	321
14	9014	0,602	0,398	1077	648	429
15	9015	0,649	0,351	555	360	195
16	9016	0,901	0,099	91	82	9
17	9017	0,564	0,436	5196	2931	2265
18	9018	0,793	0,207	241	191	50
19	9019	0,930	0,070	86	80	6
20	9020	0,730	0,270	189	138	51
21	9021	0,607	0,393	4901	2974	1927
22	9022	0,756	0,244	409	309	100
23	9023	0,601	0,399	1357	816	541
24	9024	0,538	0,462	4262	2292	1970
25	9025	0,516	0,484	1327	685	642
26	9026	0,529	0,471	1921	1017	904
27	9027	0,650	0,350	685	445	240
28	9028	0,540	0,460	1215	656	559
29	9029	0,517	0,483	2639	1364	1275
30	9030	0,492	0,508	3367	1656	1711

Πηγή: Υπουργείο Διοικητικής Μεταρρύθμισης και Ηλεκτρονικής Διακυβέρνησης, 2012 και Υπουργείο Εσωτερικών, 2012

Ουσιαστικά οι μεταβλητές pop_sex_1 και pop_sex_2 προέρχονται από τον πολλαπλασιασμό των sex_fin_1 και sex_fin_2 με την pop_t.

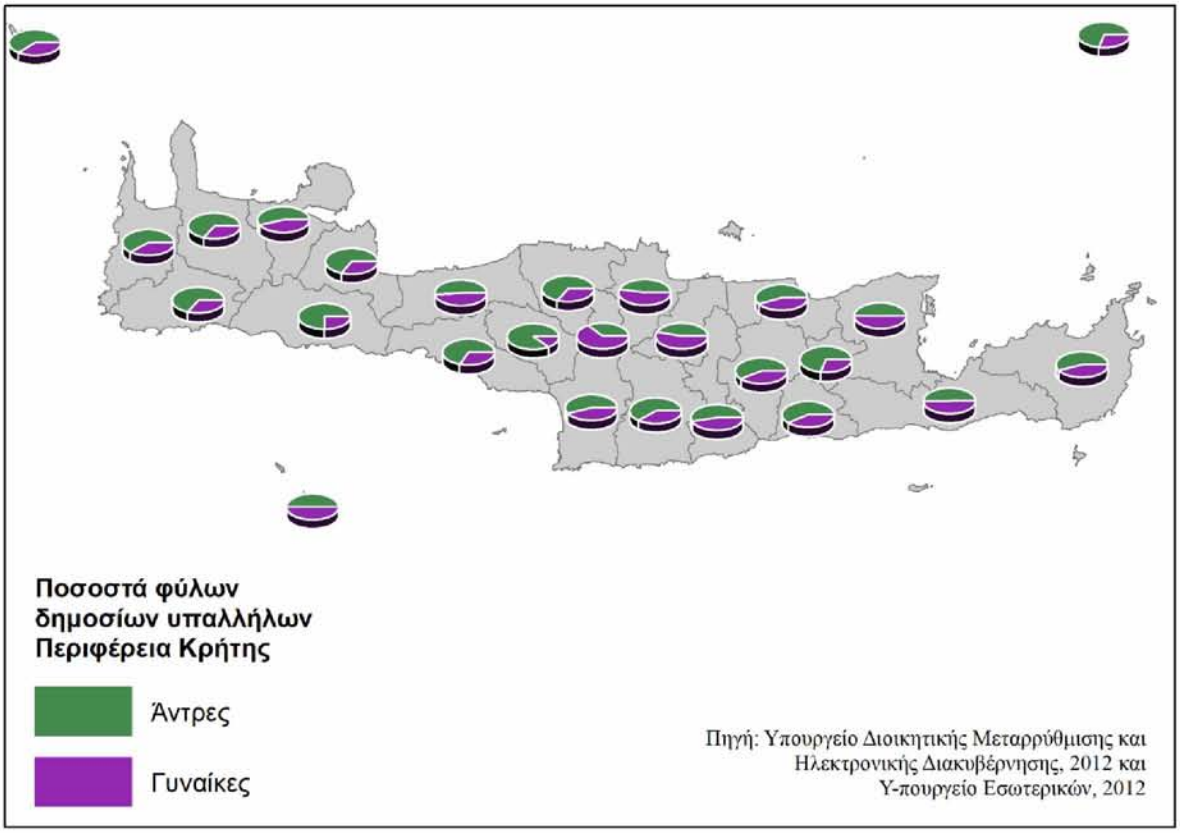
Μέσω αυτής της βάσης έχει υπολογιστεί όπως είναι φανερό, το πλήθος κάθε φύλου σε κάθε δήμο, οπότε και μπορεί και να παρουσιαστεί και σε χάρτη.

Χάρτης 3.1: Δείγμα πλήθους αντρών υπαλλήλων δημοσίου Ελλάδας, 11-7-2012



Ομοίως μπορούμε να δημιουργήσουμε τον χάρτη με το πλήθος των γυναικών ή τους αντίστοιχους χάρτες των ποσοστών κάθε φίλου.

Χάρτης 3.2: Ποσοστά φύλων υπαλλήλων δημοσίου Περιφέρειας Κρήτης, 11-7-2012



Εφαρμόζοντας την ίδια μεθοδολογία συστηματικά μπορούμε να παράγουμε χάρτες για την ανάλυση και των υπολοίπων μεταβλητών (όπως εργασιακή σχέση, οικογενειακή κατάσταση κ.α.). Όμως το πραγματικό ενδιαφέρων παρουσιάζεται στην συνέχεια, όπου εξετάζονται οι σύνθετες μεταβλητές.

- **2η Κατηγορία: Χωρική βάση δεδομένων με σύνθετες «δισδιάστατες» μεταβλητές**

Όπως και στις απλές μεταβλητές έτσι και εδώ, με την χρήση των εντολών του στατιστικού προγράμματος μπορούμε να δημιουργήσουμε νέες χωρικές βάσεις, αλλά αυτή την φορά με σύνθετες μεταβλητές.

  - **1η περίπτωση: ποσοστά και πλήθη ανά φύλο και ανά κατηγορία εκπαίδευσης (**αρχείο «bd_spatial_sex_edu»**)**

Ουσιαστικά κατασκευάζουμε τόσες μεταβλητές όσες και το γινόμενο των τιμών των αρχικών μεταβλητών. Δηλαδή, 2 από την μεταβλητή sex επί 7 από την μεταβλητή edu (6 κανονικές και 1 τα missing data). Οι νέες αυτές μεταβλητές δείχνουν μεγαλύτερη πληροφορία σε όσο το δυνατόν πιο συμπιεσμένη μορφή. Παραδείγματα από τις 14 νέες μεταβλητές είναι οι ακόλουθες:

i. sex_edu_fin_sex1_edu1 (το ποσοστό των ατόμων που ανήκουν στην κατηγορία 1 της μεταβλητής sex και στην κατηγορία 1 της μεταβλητής edu, δηλαδή άντρες με εκπαίδευση ΔΕ)

ii. sex_edu_fin_sex1_edu2 (το ποσοστό των ατόμων που ανήκουν στην κατηγορία 1 της μεταβλητής sex και στην κατηγορία 2 της μεταβλητής edu, δηλαδή άντρες με εκπαίδευση ΕΕΠ)

iii. sex_edu_fin_sex1_edu3 (το ποσοστό των ατόμων που ανήκουν στην κατηγορία 1 της μεταβλητής sex και στην κατηγορία 3 της μεταβλητής edu, δηλαδή άντρες με εκπαίδευση Ειδικών Θέσεων)

iv. sex_edu_fin_sex1_edu4 (το ποσοστό των ατόμων που ανήκουν στην κατηγορία 1 της μεταβλητής sex και στην κατηγορία 4 της μεταβλητής edu, δηλαδή άντρες με εκπαίδευση ΠΕ), ομοίως οι υπόλοιπες για το sex=1

v. sex_edu_fin_sex2_edu1 (το ποσοστό των ατόμων που ανήκουν στην κατηγορία 2 της μεταβλητής sex και στην κατηγορία 1 της μεταβλητής edu, δηλαδή γυναίκες με εκπαίδευση ΔΕ)

vi. sex_edu_fin_sex2_edu2 (το ποσοστό των ατόμων που ανήκουν στην κατηγορία 2 της μεταβλητής sex και στην κατηγορία 2 της μεταβλητής edu, δηλαδή γυναίκες με εκπαίδευση ΕΕΠ)

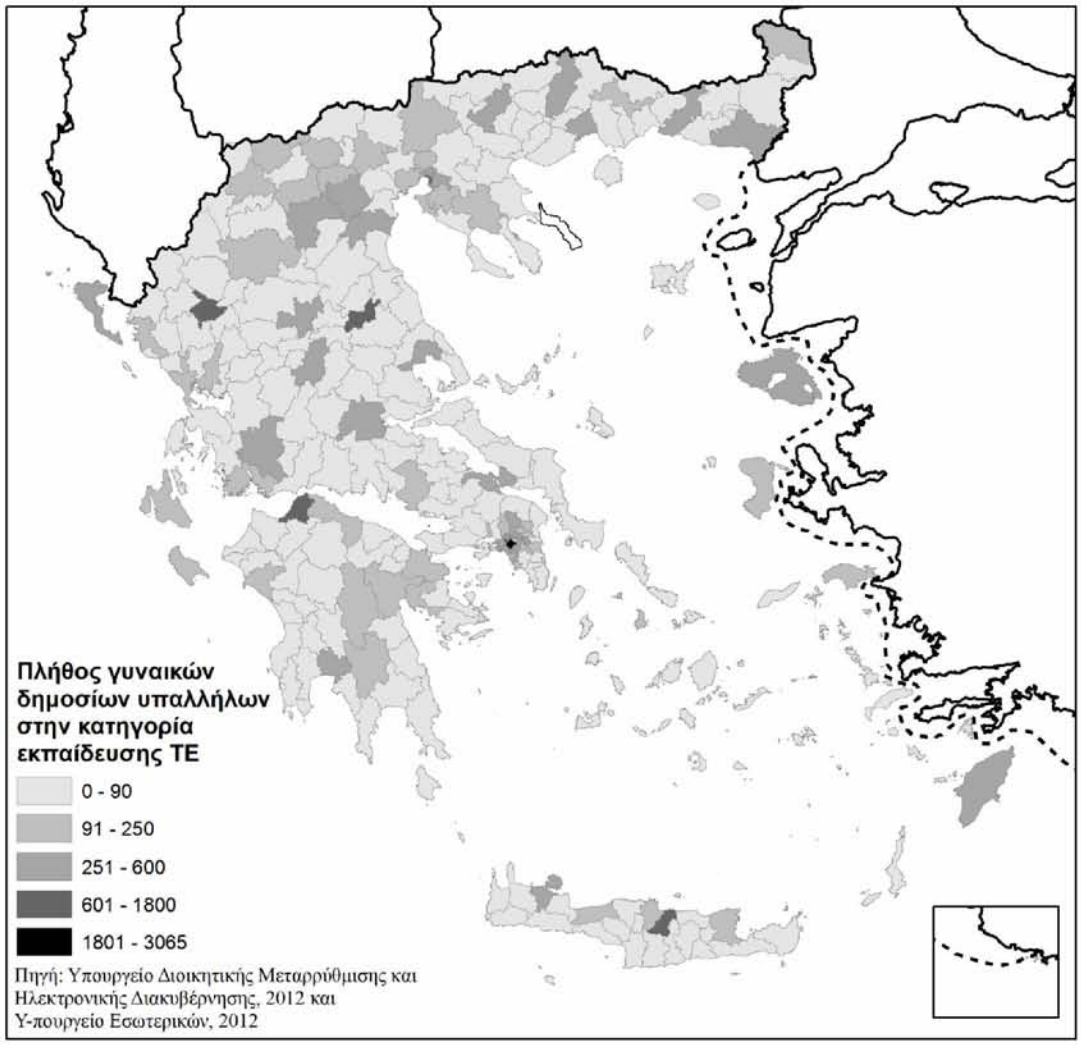
vii. sex_edu_fin_sex2_edu3 (το ποσοστό των ατόμων που ανήκουν στην κατηγορία 2 της μεταβλητής sex και στην κατηγορία 3 της μεταβλητής edu, δηλαδή γυναίκες με εκπαίδευση Ειδικών Θέσεων)

viii. sex_edu_fin_sex2_edu4 (το ποσοστό των ατόμων που ανήκουν στην κατηγορία 2 της μεταβλητής sex και στην κατηγορία 4 της μεταβλητής edu, δηλαδή γυναίκες με εκπαίδευση ΠΕ), ομοίως οι υπόλοιπες για το sex=2

Επίσης, έχοντας τον πληθυσμό ανά δήμο (pop_t) και πολλαπλασιάζοντάς το με τα ποσοστά των παραπάνω μεταβλητών δημιουργούμε και τις επόμενες 14 μεταβλητές, οι οποίες περιέχουν τα ίδια στοιχεία αλλά σε καθαρούς πλέον πληθυσμούς.

Οποιοδήποτε από τα παραπάνω στοιχεία μπορεί να χρησιμοποιηθεί στα πλαίσια μελέτης του χώρου για πιο εξειδικευμένες περιπτώσεις. Για παράδειγμα η μελέτη της κατανομής στον ελληνικό χώρο των γυναικών του δημοσίου με επίπεδο εκπαίδευσης ΤΕ, μπορεί να πραγματοποιηθεί μέσω του παρακάτω χάρτη.

Χάρτης 3.3: Δείγμα πλήθους γυναικών υπαλλήλων δημοσίου Ελλάδας με κατηγορία εκπαίδευσης ΤΕ, 11-7-2012



- **2η περίπτωση: ποσοστά και πλήθη ανά φύλο και ανά κατηγορία ηλικίας (**αρχείο «bd_spatial_sex_age»)

Ομοίως εφαρμόζοντας συστηματικά την παραπάνω μεθοδολογία δημιουργείτε η βάση με τις 24 μεταβλητές (2 τιμές φύλου, 11 τιμές ηλικιών + 1 τιμή για τα missing data στην ηλικία).

i. sex_age_fin_sex1_age1 (το ποσοστό των ατόμων που ανήκουν στην κατηγορία 1 της μεταβλητής sex και στην κατηγορία 1 της μεταβλητής age, δηλαδή άντρες με ηλικία 20-24)

ii. sex_age_fin_sex1_age2 (το ποσοστό των ατόμων που ανήκουν στην κατηγορία 1 της μεταβλητής sex και στην κατηγορία 2 της μεταβλητής age, δηλαδή άντρες με ηλικία 25-29)

iii. sex_age_fin_sex1_age3 (το ποσοστό των ατόμων που ανήκουν στην κατηγορία 1 της μεταβλητής sex και στην κατηγορία 3 της μεταβλητής age, δηλαδή άντρες με ηλικία 30-34), ομοίως τα υπόλοιπα έως sex_age_fin_sex1_ege99.

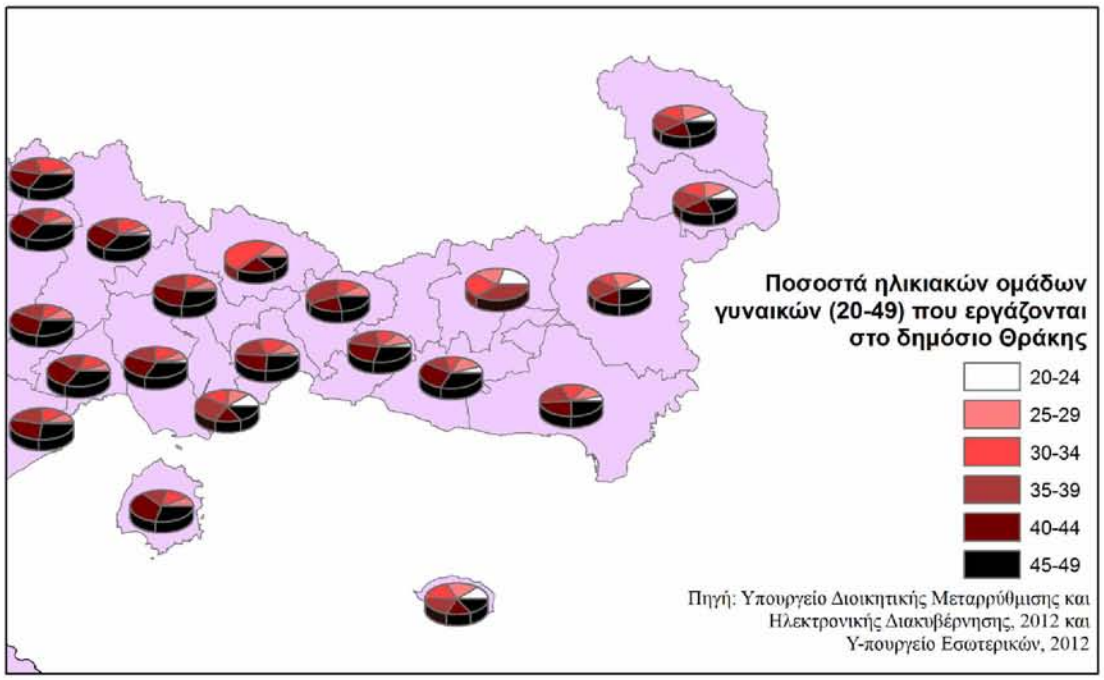
iv. sex_age_fin_sex2_age1 (το ποσοστό των ατόμων που ανήκουν στην κατηγορία 2 της μεταβλητής sex και στην κατηγορία 1 της μεταβλητής age, δηλαδή γυναίκες με ηλικία 20-24)

v. sex_age_fin_sex2_age2 (το ποσοστό των ατόμων που ανήκουν στην κατηγορία 2 της μεταβλητής sex και στην κατηγορία 2 της μεταβλητής age, δηλαδή γυναίκες με ηλικία 25-29)

vi. sex_age_fin_sex2_age3 (το ποσοστό των ατόμων που ανήκουν στην κατηγορία 2 της μεταβλητής sex και στην κατηγορία 3 της μεταβλητής age, δηλαδή γυναίκες με ηλικία 30-34), ομοίως τα υπόλοιπα έως sex_age_fin_sex2_ege99.

Ακολουθεί παράδειγμα μελέτης για την κατανομή στον χώρο της Θράκης των γυναικών του δημοσίου, οι οποίες ανήκουν στις ηλικιακές ομάδες 20-24, 25-29, 30-34, 35-39, 40-44 και 45-49 μέσω του παρακάτω χάρτη.

Χάρτης 3.4: Ποσοστά γυναικών υπαλλήλων δημοσίου Θράκης, ηλικιακές ομάδες 24-49, 11-7-2012



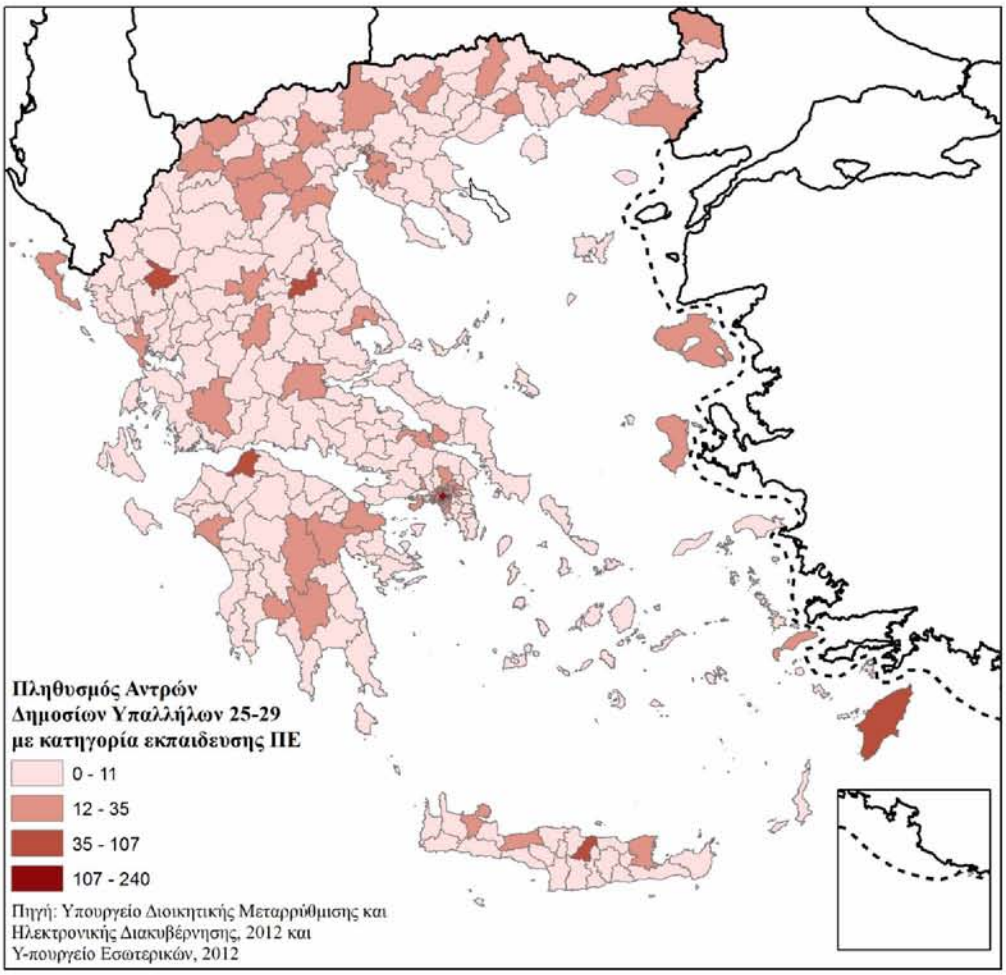
Ενώ τέλος, περνάμε στις πιο σύνθετες ακόμα μεταβλητές «3 διαστάσεων», όπου και ολοκληρώνεται η ανάπτυξη των παραδειγμάτων αυτής της εργασίας.

- **3η Κατηγορία: Χωρική βάση δεδομένων με σύνθετες «τρισδιάστατες» μεταβλητές (ποσοστά και πλήθη ανά φύλο, κατηγορία εκπαίδευσης και κατηγορία ηλικίας)** (αρχείο «bd_spatial_sex_age_edu»)

Με εφαρμογή συστηματικά της ίδιας μεθοδολογίας μπορούμε να φτάσουμε να έχουμε τις νέες 168 μεταβλητές (μόνο για τα ποσοστά, κατά αντιστοιχία θα έχουμε τις διπλές αν αποφασίσουμε να κάνουμε και τις μεταβλητές με τους κανονικούς πληθυσμούς). Προέρχονται από το γινόμενο των 2 φύλων, των 7 τιμών στην εκπαίδευση και των 12 τιμών στις ηλικιακές ομάδες. Η πρώτη μεταβλητή είναι η sex_age_edu_fin10101, δηλαδή το πλήθος των αντρών (το πρώτο 1) που έχουν ηλικία 20-24 (το 01) και ανήκουν στην κατηγορία εκπαίδευσης ΔΕ (το τελευταίο 01). Συνεχίζοντας αντίστοιχα με την ίδια λογική καταλήγουμε στην τελευταία μεταβλητή sex_age_edu_fin29999, η οποία παρουσιάζει το πλήθος των γυναικών (2) η οποίες δεν δήλωσαν ούτε ηλικία (99), ούτε κατηγορία εκπαίδευσης (99).

Ο παρακάτω χάρτης μπορεί να χρησιμοποιηθεί για την μελέτη των αντρών ηλικίας 25-29 με κατηγορία εκπαίδευσης ΠΕ για όλο τον ελληνικό χώρο.

Χάρτης 3.5: Πλήθος αντρών 25-29, επιπέδου εκπαίδευσης ΠΕ, εργαζομένων στο ελληνικό δημόσιο 11-7-2012



3.3.4 ΠΑΡΑΔΕΙΓΜΑ ΜΕΛΕΤΗΣ (ΕΚΠΑΙΔΕΥΣΗ ΔΗΜΟΣΙΩΝ ΥΠΑΛΛΗΛΩΝ ΕΛΛΑΔΟΣ)

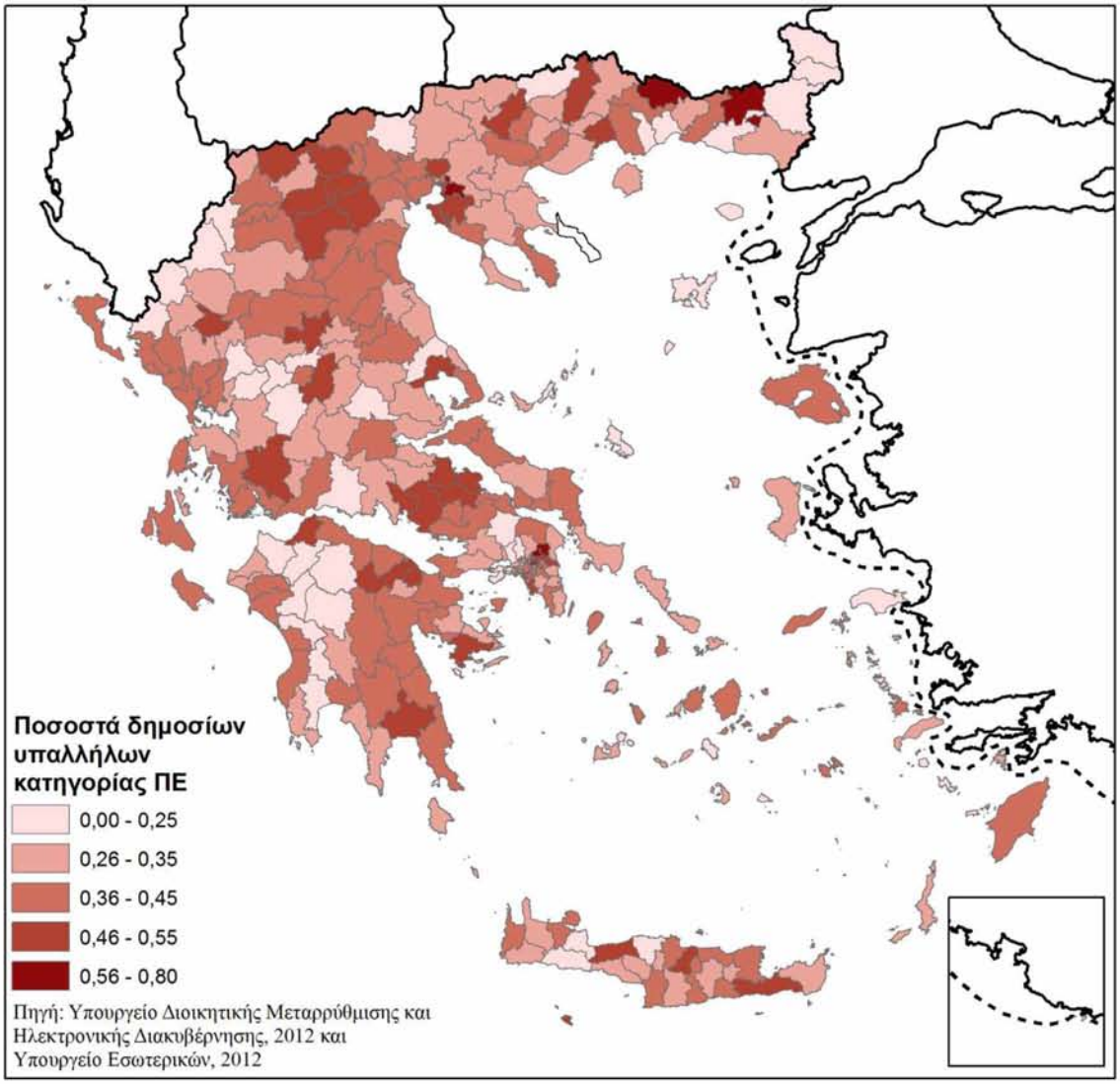
Στην συνέχεια της μελέτης περίπτωσης πραγματοποιείται η αρχική ανάλυση/παρουσίαση του μορφωτικού επιπέδου των ΔΥ και των χωρικών του χαρακτηριστικών. Αποτελεί παράδειγμα της δυνατότητας εξαγωγής συσσωρευμένης πληροφορίας και συμπερασμάτων από την αντίστοιχη βάση δεδομένων.

Αρχικά ορίζονται κάποιες βασικές ερωτήσεις, η οποίες θα απαντηθούν μέσω της χρήσης των νέων βάσεων. Από της απαντήσεις τους θα παραχθούν τα συμπεράσματα για την χωρική μορφή των ΔΥ και της εκπαίδευσής τους.

Οι βασικές ερωτήσεις μπορούν να ποικίλουν από μελέτη σε μελέτη, εδώ τέθηκαν οι ακόλουθες: α) Σε ποιες περιοχές απασχολείται το μορφωμένο δυναμικό (ΠΕ, ΤΕ, ΕΕΠ και Ειδικών θέσεων); Και β) Αντίστοιχα σε ποιες το δυναμικό χαμηλού επιπέδου (ΔΕ και ΥΕ);

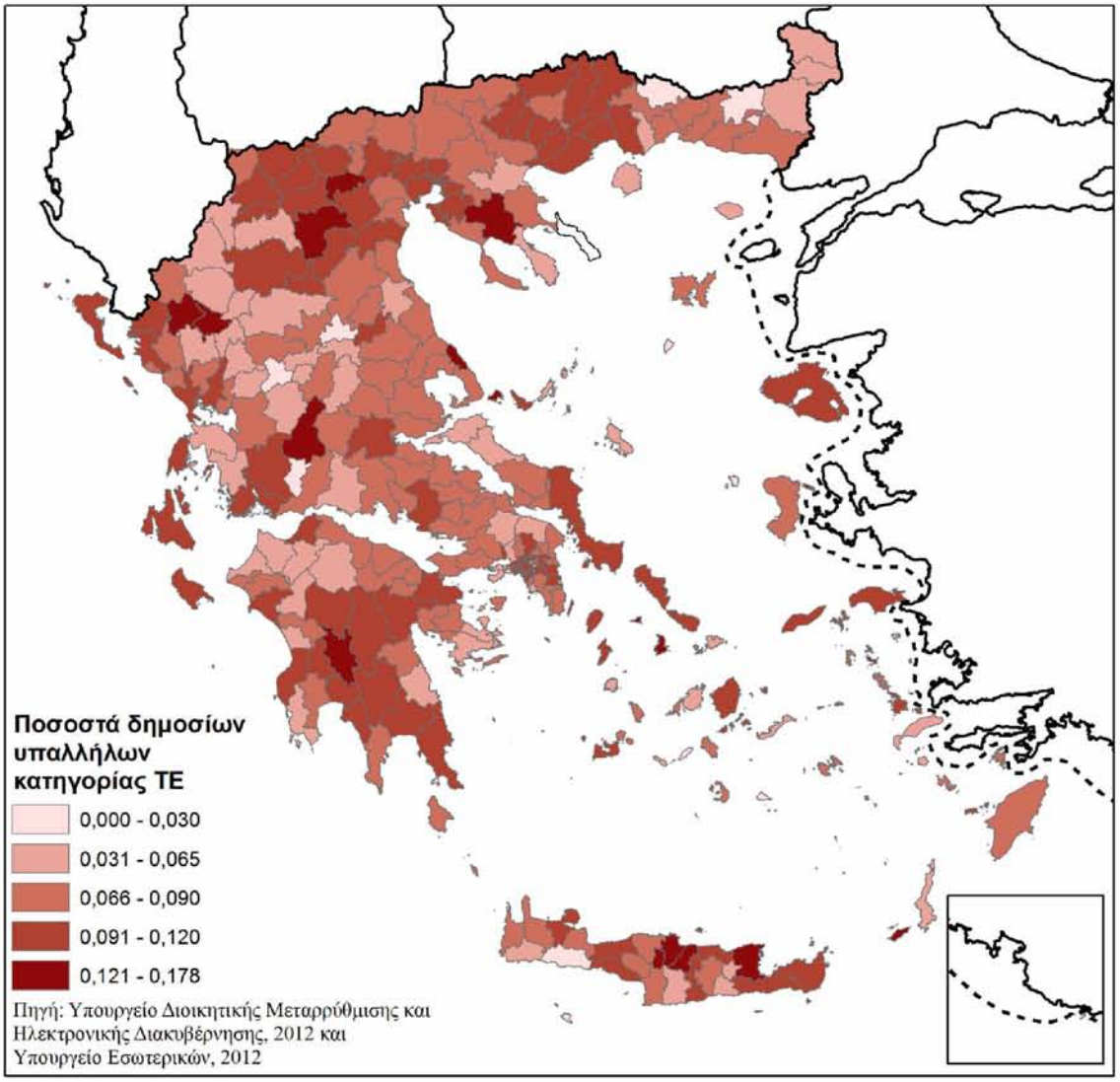
Για την πρώτη ερώτηση δημιουργούνται οι τέσσερεις επόμενοι χάρτες, ένας για κάθε κατηγορία εκπαίδευσης.

Χάρτης 3.6: Ποσοστά ΔΥ επιπέδου εκπαίδευσης ΠΕ, 11-7-2012



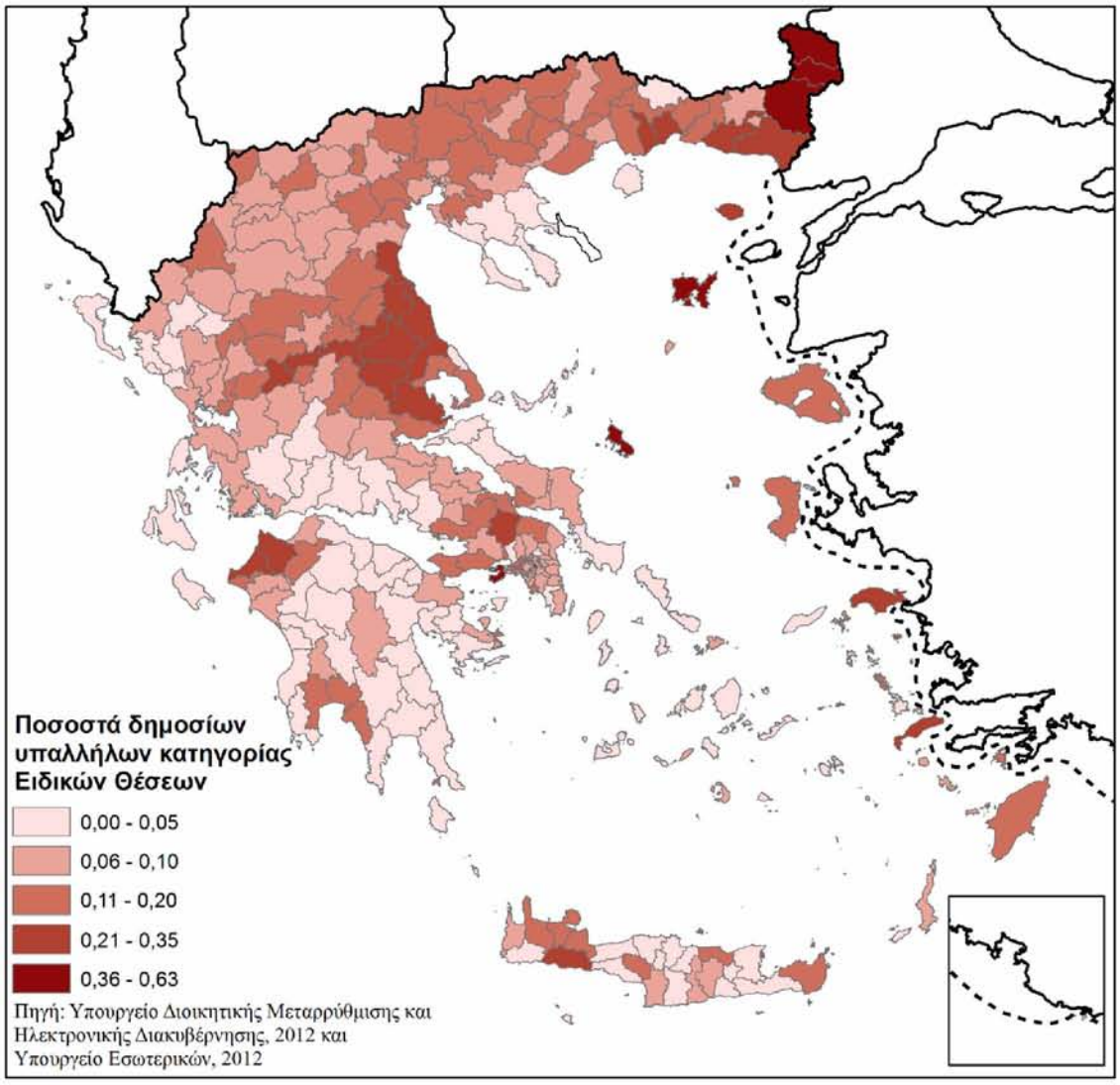
Όπως παρατηρείται στον παραπάνω χάρτη για των υπαλλήλων πανεπιστημιακής εκπαίδευσης, οι μεγαλύτερες συγκεντρώσεις σημειώνονται στα μεγάλα αστικά κέντρα και στις πρωτεύουσες των νομών (Καποδίστριας) και των νέων περιφερειών (Καλλικράτης). Αυτό λογικά συμβαίνει λόγω των απαραίτητων υπηρεσιών που πρέπει να παρέχουν οι δήμοι/περιφέρειες/νομοί, όπως πολεοδομίες, της παλιές νομαρχίες, υπουργεία κ.α. Επίσης σημειώνεται διάχυση της μεταβλητής στις γειτονικές περιοχές των προηγουμένων περιοχών ενώ μειωμένα ποσοστά εμφανίζουν οι ορεινές περιοχές και η παραμεθόριος.

Χάρτης 3.7: Ποσοστά ΔΥ επιπέδου εκπαίδευσης ΤΕ, 11-7-2012



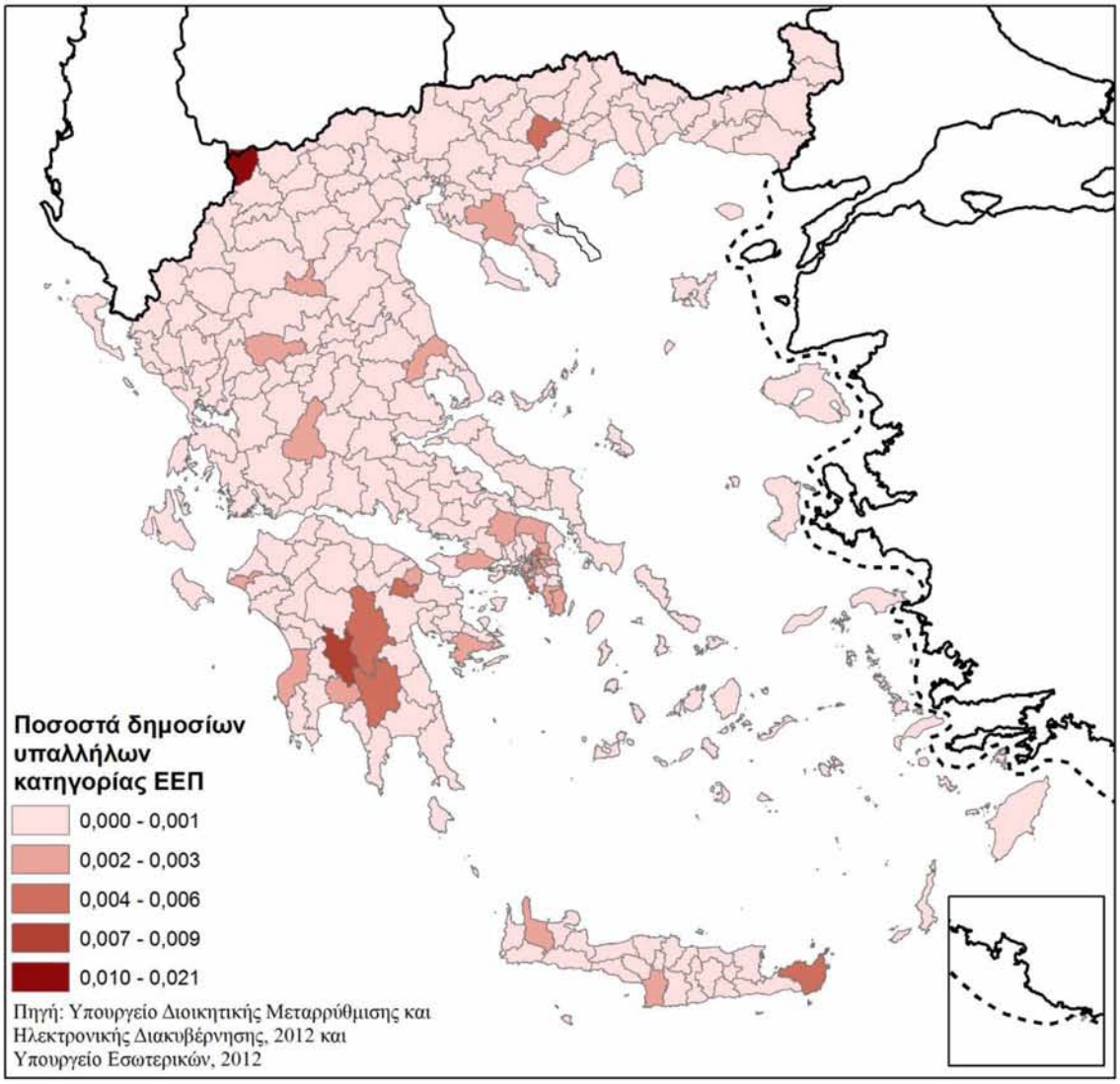
Όσον αφορά τους ΔΥ τεχνολογικής εκπαίδευσης, ομοίως με τους ΠΕ, συγκεντρώνονται στα μεγάλα αστικά κέντρα παρά τις ορεινές και παραμεθόριες περιοχές. Όμως εδώ, η διάχυση είναι πολύ μεγαλύτερη. Επίσης, παρατηρείται σημαντικός αριθμός μεγάλων συγκεντρώσεων και σε δευτερεύουσες περιοχές, πιθανόν λόγω βιομηχανικών και τεχνολογικών χρήσεων στις ευρύτερες περιοχές των αστικών περιοχών.

Χάρτης 3.8: Ποσοστά ΔΥ επιπέδου εκπαίδευσης Ειδικών θέσεων, 11-7-2012



Στην παρούσα κατηγορία ΔΥ (Ειδικών θέσεων), δεν παρατηρείται συγκεκριμένο «μοτίβο» χωρικής κατανομής, πιθανόν λόγω της φύσης της συγκεκριμένης κατηγορίας. Ίσως, η πρόσληψη των ΔΥ Ειδικών θέσεων να εξαρτάται από τις ιδιάζουσες περιπτώσεις κάθε δήμου.

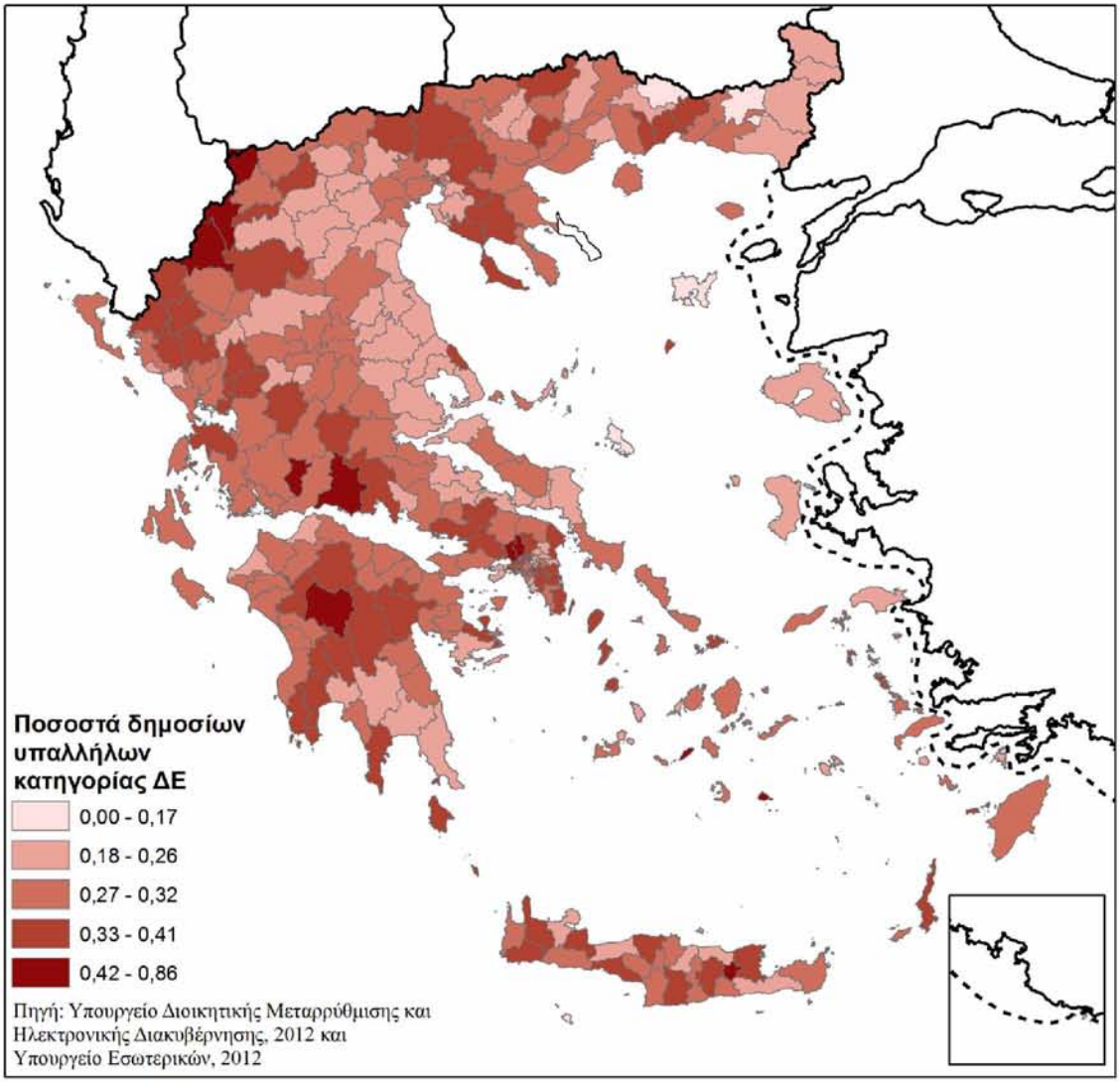
Χάρτης 3.9: Ποσοστά ΔΥ επιπέδου εκπαίδευσης ΕΕΠ, 11-7-2012



Ομοίως με την προηγούμενη κατηγορία (Ειδικών θέσεων) και εδώ οι ΔΥ Ειδικού Επιστημονικού Προσωπικού, δεν παρουσιάζουν συγκεκριμένη κατανομή στον ελληνικό χώρο. Βέβαια εδώ δεν είναι μόνο λόγω της φύσης του επαγγέλματος, αλλά και λόγω του χαμηλού πληθυσμού που σημειώνει η συγκεκριμένη μεταβλητή. Επίσης σημειώνει και τα χαμηλότερα ποσοστά σε σχέση με όλες τις υπόλοιπες.

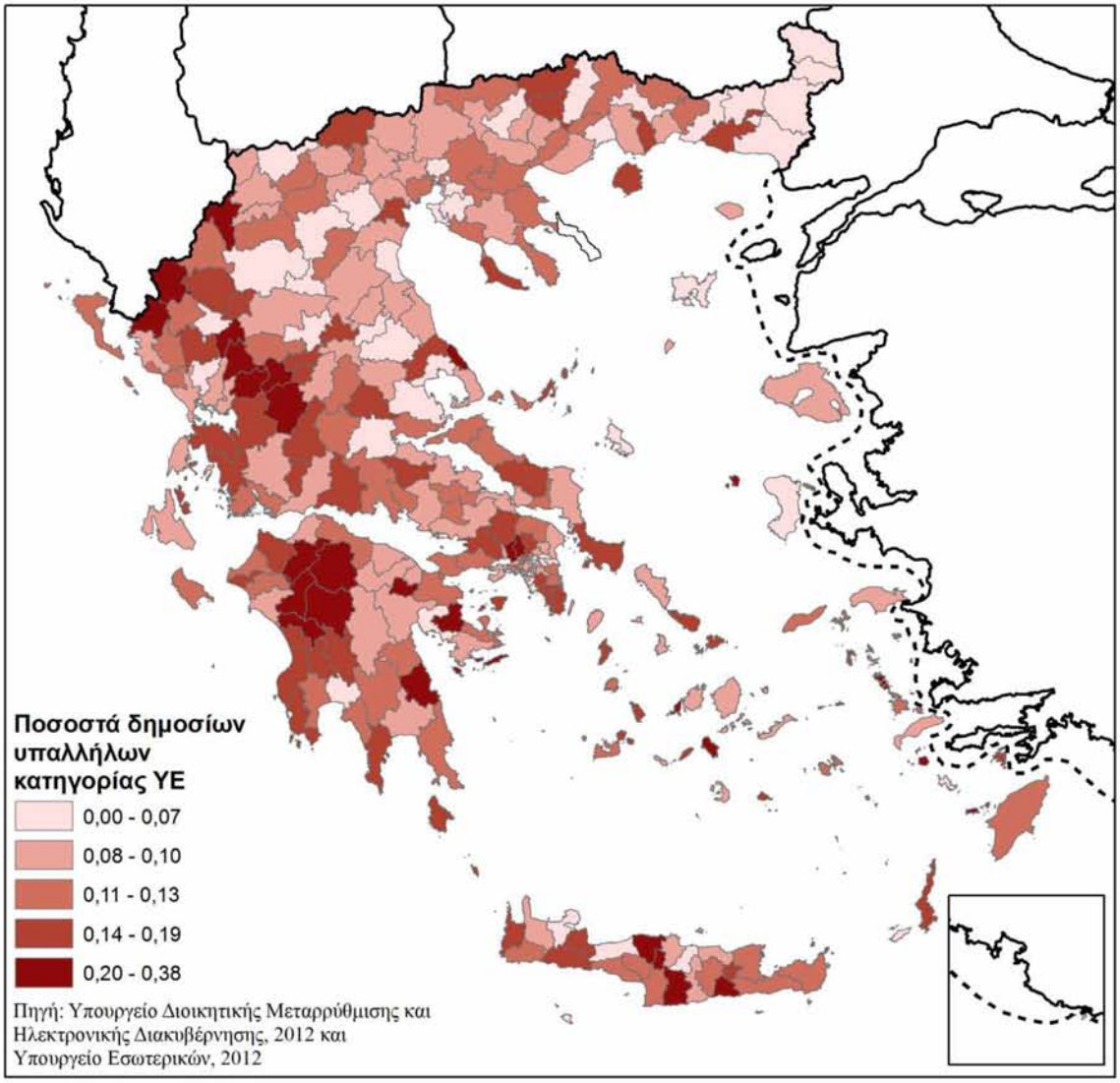
Περνώντας στην επόμενη ερώτηση (για την ομάδα των ΔΥ με χαμηλότερη εκπαίδευση), παρουσιάζονται δύο χάρτες, ένας για ΔΕ (Δευτεροβάθμιας Εκπαίδευσης) και ένας για ΥΕ (Υποχρεωτικής Εκπαίδευσης).

Χάρτης 3.10: Ποσοστά ΔΥ επιπέδου εκπαίδευσης ΔΕ, 11-7-2012



Σε αυτό το σημείο αρχίζουμε και παρατηρούμε την αντίστροφη σχέση με της κατηγορίες του υψηλότερου μορφωτικού επιπέδου. Οι ΔΥ παρουσιάζουν συγκεντρώσεις σε ορεινές και «απομακρυσμένες» περιοχές ή περιοχές συνόρων. Βέβαια και εδώ παρουσιάζεται σχετικά έντονη διάχυση, πιθανόν λόγω της ανάγκης για κάλυψη αρκετών χαμηλών θέσεων/αρμοδιοτήτων διαδημοτικά.

Χάρτης 3.11: Ποσοστά ΔΥ επιπέδου εκπαίδευσης ΥΕ, 11-7-2012



Τέλος, στην κατηγορία ΥΕ γίνεται πλέον ξεκάθαρη η αντίθεση μεταξύ των κατηγοριών υψηλής και χαμηλής εκπαίδευσης. Οι θέσεις με χαμηλότερη εκπαίδευση παρουσιάζονται στις ορεινές και «ακριτικές» περιοχές, δηλαδή σε περιοχές με «χαμηλές απαιτήσεις» και λιγότερο προτιμητέες από τον μορφωμένο πληθυσμό.

3.3.5 ΣΥΜΠΕΡΑΣΜΑΤΑ ΜΕΛΈΤΗΣ

Οπότε συγκεντρωτικά από τα παραπάνω διεξάγονται τα ακόλουθα συμπεράσματα:

- Οι περιοχές με μεγάλα αστικά κέντρα και πρωτεύουσες συγκεντρώνουν μεγαλύτερους πληθυσμούς υψηλά εκπαιδευμένου προσωπικού, πιθανόν λόγω μεγαλύτερων απαιτήσεων σε μορφωμένο πληθυσμό και προτίμησής τους για το καλύτερο βιοτικό επίπεδο.
- Σε αντίθεση οι ορεινές και απομακρυσμένες ή με δύσκολη πρόσβαση περιοχές παρουσιάζουν μικρές συγκεντρώσεις σε πληθυσμό τέτοιου μορφωτικού επιπέδου. Λόγοι για αυτό το φαινόμενο είναι το χαμηλότερο βιοτικό επίπεδο και οι χαμηλές απαιτήσεις σε υπηρεσίες του δημοσίου.

- Οι κατηγορίες ΕΕΠ και Ειδικών Θέσεων δεν παρουσιάζουν «συστηματικές εμφανίσεις» στον χώρο, αφού ο ιδιάζον χαρακτήρας τους και ο χαμηλός αριθμός τέτοιων θέσεων δεν τους το επιτρέπουν.
- Τέλος, οι «μεσαίες» κατηγορίες, δηλαδή ΔΕ και ΤΕ, σημειώνουν διάχυση στον ελληνικό χώρο. Αιτία αυτού αποτελεί πιθανόν η «ευελιξία» των αντίστοιχων θέσεων και απαιτήσεων τους. Μόνο για τους υπαλλήλους της ΤΕ μπορεί να ειπωθεί πως παρουσιάζουν ελαφρώς μεγαλύτερες συγκεντρώσεις σε περιοχές τεχνολογικής και βιομηχανικής χρήσης.

3.4 ΣΥΜΠΕΡΑΣΜΑΤΙΚΑ

Η παραπάνω μεθοδολογία μπορεί να χρησιμοποιηθεί για την παραγωγή πλήθους βάσεων σύνθετων χωρικών μεταβλητών με σκοπό την «αποστράγγιση» της απαραίτητης πληροφορίας. Θέτοντας το βασικό «ερώτημα» της μελέτης, μπορούμε να δημιουργήσουμε την αντίστοιχη χωρική βάση με τις σωστές απλές ή σύνθετες χωρικές μεταβλητές, μέσω των οποίων ανέρχεται η «απάντηση».

Πέρα από την μεταβλητή της εκπαίδευση (τμήμα της οποίας αναλύθηκε παραπάνω), θα ήταν δυνατόν με την παρούσα βάση να «εξαντλήσουμε» και άλλες πληροφορίες που θα χρησίμευαν σε ανάλογη έρευνα (π.χ. η σχέση μεταξύ χώρου και οικογενειακής κατάστασης/φύλου/ηλικίας, αλλά και σύνθεσης μεταξύ αυτών όπως παρουσιάστηκε στα παραδείγματα του 3.3.3)

Κλείνοντας, συμπεραίνουμε από το σύνολο της μελέτης περίπτωσης πως οι βάσεις χωρικών δεδομένων αν και χρειάζονται αρκετό χρόνο για την προετοιμασία τους, αποδίδουν πλήθος στοιχείων και αξιόπιστα συμπεράσματα ώστε να δημιουργήσουμε τους πυλώνες των εκάστοτε αναλύσεων και ερευνών.

4. ΕΠΙΛΟΓΟΣ

Όπως παρατηρήθηκε στην μελέτη περίπτωσης, με την εφαρμογή του θεωρητικού υποβάθρου και την ανάπτυξη εμπειρικών μεθόδων, πλέον είναι εφικτό να παραχθούν χωρικές βάσεις (όποτε παρέχονται χωρικά δεδομένα) από διαφόρων ειδών βάσεις όπως στατιστικές, κοινωνικοοικονομικές κ.α., για την ευκολότερη χρήση και εξαγωγή πληροφοριών από αναλυτές του χώρου.

Αν και παρουσιάζει μειονέκτημα λόγω εκτεταμένου χρόνου επεξεργασίας, με αυτόν τον τρόπο παράγεται συνεχόμενα νέα χωρική πληροφορία η οποία μπορεί να χρησιμοποιηθεί είτε για αναγνώριση φαινομένων, είτε για την εξαγωγή συμπερασμάτων, είτε ακόμα και για την χρήση της στα πλαίσια λήψης αξιόπιστων αποφάσεων για των σχεδιασμό και την οργάνωση του χώρου.

ΠΗΓΕΣ ΤΕΚΜΗΡΙΩΣΗΣ

ΒΙΒΛΙΟΓΡΑΦΙΚΕΣ ΑΝΑΦΟΡΕΣ

ΕΛΛΗΝΙΚΕΣ:

- Κιντής Α (1999) "Στατιστικές και Οικονομετρικές Μέθοδοι", Gutenberg, Αθήνα, 27-35, 39-53
- Κιντής Α. (2010) "Σύγχρονη Οικονομετρική Ανάλυση", τόμος 1, Gutenberg, Αθήνα, 33-37
- Παπαδήμας Ο. και Κοίλιας Χ. (1998) "Εφαρμοσμένη Στατιστική", Εκδόσεις Νέων Τεχνολογιών, Αθήνα, 21-102
- Τραχανάς Κ. (2003) "Οικονομετρικά υποδείγματα και οικονομικές – διοικητικές αποφάσεις", ΣΠΟΥΔΑΙ, Τόμος 53, Τεύχος 2ο, Πανεπιστήμιο Πειραιά, Αθήνα, 90-92

ΞΕΝΟΓΛΩΣΣΕΣ:

- Cohen J. & Cohen P. (1983) "Applied multiple regression correlation analysis for the behavioral sciences" (2nd ed.), Hillsdale, NJ: Lawrence Erlbaum Associates
- Dempster A. P., Laird N. M. & Rubin D. B. (1997) "Maximum Likelyhood estimators from incomplete data via the EM algorithm", Journal of the Royal Statistical Society, Series B, 39, 1-38
- Froeschl K.A. (1997) "Metadata Management in Statistical Information Processing", Wien: Springer, ISBN 3-211-82987-3
- Froeschl K.A. (1999) "Metadata Management in Official Statistics – An IT-Based Methodology Approach", Austrian Journal of Statistics, Vol 28 No2, 49–79
- Froeschl K.A., Yamada T. και Kudrna R. (2002) "Industrial Statistics Revisited: From Footnotes to Meta-Information Management", Austrian Journal of Statistics, 31(1), 9–34
- Graham J. W. & Schafer J. L. (1999) "On the performance of multiple imputation for multivariate data with small sample size. In Statistical Strategies for smallsample research", 1-29

- Graham J. W. & Schafer J. L. (1999) "Missing Data: Our View of the State of the Art", American Psychological Association, 7 (2), 147-177
- Grossmann W. και Papageorgiou H. (1997) "Data and Metadata Representation of Highly Aggregated Economic Time-Series", 51st Session of the International Statistical Institute, Contributed Papers, 2, 485-486
- Hand D.J. (1993) "Data, metadata, and information", Statistical Journal of the United Nations, 10, 143–151
- Hartley H. O. & Hocking R. R. (1971) "The analysis of incomplete data", Biometrics, 27, 783-823
- Kim J. O. & Curry J. (1977) "The treatment of missing data in multivariate analysis", Sociological Methods and Research, 6 (2), 215-240
- Little R. J. & Rubin D. B. (1987) "Statistical analysis with missing data", New York: Wiley
- Malvestuto F.M. (1993) "A Universal-Scheme Approach to Statistical Databases Containing Homogeneous Summary Tables", ACM Transactions on Database Systems, 18, 678-708
- Papageorgiou H., Vardaki, M. και Pentaris F. (2000) "Data and Metadata Transformations", Research in Official Statistics (ROS), 3(2), 27-43
- Rubin D. B. (1996) "Multiple imputation after 18+ years", Journal of American Statistical Association, 91, 473-489
- Schafer J. L. (1997) "Analysis of incomplete multivariate data", New York: Chapman & Hall
- Sundgren B. (1996) "Making Statistical Data More Available", International Statistical Review, 64, 23-38
- Sundgren B. (1999) "Information Systems Architecture for National and International Statistical Offices Guidelines and Recommendations", Conference of European Statisticians, UNECE, Statistical Standards and Studies, paper No. 51, UN, Geneva
- Sundgren B. (2000) "The Swedish Statistical Metadata System", ECE Workshop on Statistical Metadata, Working Paper 1.6, Luxembourg

- Sundgren B (2004) "Documentation templates and metadata models at Statistics Sweden", Metadata Working Group 2004, Luxembourg

ΑΝΕΚΔΟΤΕΣ ΜΕΛΕΤΕΣ

- Ζημεράς Σ. (2003) "Στατιστικά Πακέτα 1", Πανεπιστήμιο Αιγαίου, Σημειώσεις Μαθήματος, Τμήμα Στατιστικής και Αναλογιστικής Επιστήμης, Σάμος, 3-18
- Ντάλα Μ. (2009) "Εφαρμογή αλγορίθμων επαγωγικού λογικού προγραμματισμού στην σχεσιακή εξόρυξη δεδομένων", Μεταπτυχιακή Εργασία, Πανεπιστήμιο Πάτρας, Σχολή θετικών επιστημών, Τμήμα Μαθηματικών, Πάτρα, 14-41
- Παπαδάκη Σ. (2009) "Αξιολόγηση Στατιστικών Μεθοδολογιών για τη Διαχείριση Δεδομένων με Ελλιπή Στοιχεία", Μεταπτυχιακή Διπλωματική Εργασία, Πολυτεχνείο Κρήτης, Τμήμα Μηχανικών Παραγωγής και Διοίκησης, Χανιά, 13-102

ΔΙΑΔΙΚΤΥΑΚΟΙ ΤΟΠΟΙ

- Ρέππας Δ. 2012, Συνέντευξη Τύπου Υπουργού Διοικητικής Μεταρρύθμισης και Ηλεκτρονικής Διακυβέρνησης, και Υφυπουργών Ντίνου Ρόβλια και Παντελή Τζωρτζάκη, διαθέσιμο στο site του Υπουργείου Διοικητικής Μεταρρύθμισης και Ηλεκτρονικής Διακυβέρνησης: http://www.ydmed.gov.gr/?p=1958, προσπελάστηκε στις 5/10/2012
- Υπουργείο Εσωτερικών 2012, Αρχείο κωδικών δήμων Καλλικράτη, διαθέσιμο στο site του Υπουργείου Εσωτερικών: http://www.google.gr/url?sa=t&rct=j&q=%CE%BA%CF%89%CE%B4%CE%B9%CE%BA%CE%BF%CE%B9+%CE%B4%CE%B7%CE%BC%CF%89%CE%BD+xls&source=web&cd=3&ved=0CDQQFjAC&url=http%3A%2F%2Fwww.ypes.gr%2FUserFiles%2Ff0ff9297-f516-40ff-a70e-eca84e2ec9b9%2Fkallikraths_kwdikologio1_31_5_11.xls&ei=rgi-UJIOqdPhBLXWgYAN&usg=AFQjCNHh0QEt4pAKUL8PPXdJpOLpBBLHjw προσπελάστηκε στις 5/10/2012

ΒΑΣΗ ΔΕΔΟΜΕΝΩΝ

- Υπουργείο Διοικητικής Μεταρρύθμισης και Ηλεκτρονικής Διακυβέρνησης, Γενική Διεύθυνση Κατάστασης Προσωπικού Ομάδα Απογραφής 2012, μορφή xls, ημερομηνία εξαγωγής δεδομένων 11-7-2012