ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ
ΣΧΟΛΗ ΣΤΑΤΙΣΤΙΚΗΣ ΚΑΙ ΑΣΦΑΛΙΣΤΙΚΗΣ ΕΠΙΣΤΗΜΗΣ
ΤΜΗΜΑ ΕΦΑΡΜΟΣΜΕΝΗΣ ΣΤΑΤΙΣΤΙΚΗΣ

ΔΕΙΓΜΑΤΟΛΗΨΙΑ ΧΡΟΝΟΣΗΜΑΣΜΕΝΩΝ, ΑΚΟΛΟΥΘΙΑΚΩΝ, ΣΥΝΘΕΤΩΝ ΤΥΠΩΝ ΔΕΔΟΜΕΝΩΝ

Διπλωματική Εργασία

ΚΟΥΦΟΠΟΥΛΟΥ ΕΥΘΥΜΙΑ

Πανεπιστήμιο Πειραιώς

Πανεπιστήμιο Πειραιώς

Η Τριμελής Επιτροπή

Ι. Θεοδωρίδης	**Ελ. Κοφίδης**	**Ν. Πελέκης**
Αναπλ. Καθηγητής	**Επίκ. Καθηγητής**	**Λέκτορας**

Περιεχόμενα

Πανεπιστήμιο Πειραιώς

Κατάλογος Σχημάτων

Πανεπιστήμιο Πειραιώς

Κατάλογος Πινάκων

Πανεπιστήμιο Πειραιώς

Περίληψη

Στόχος της παρούσας διπλωματικής εργασίας είναι η μελέτη και η προσαρμογή των διαφόρων τεχνικών δειγματοληψίας σε σύνθετους τύπους δεδομένων, των οποίων το κοινό χαρακτηριστικό είναι ότι πρόκειται για χρονοσημασμένα, ακολουθιακά δεδομένα. Αρχικά θα να γίνει ορθή συλλογή και καταγραφή χρονοσημασμένων δεδομένων χρησιμοποιώντας τις διάφορες μεθόδους δειγματοληψίας. Στη συνέχεια, τα δεδομένα θα τύχουν επεξεργασίας και θα παραχθούν τα σωστά αποτελέσματα και συμπεράσματα.

Η συλλογή του τεράστιου όγκου δεδομένων μπορεί να γίνει με διάφορες μεθόδους δειγματοληψίας. Πιο συγκεκριμένα, η μέθοδος της δειγματοληψίας αφορά τη λήψη ενός τμήματος στοιχείων από κάποιο ευρύτερο σύνολο στοιχείων και κατηγοριοποιείται σε δύο υπό-ομάδες, τη δειγματοληψία βάση πιθανοτήτων και τη δειγματοληψία χωρίς πιθανότητα.

Στην δεύτερη μας ενότητα θα παρουσιάσουμε τις διάφορες μεθόδους της δειγματοληψίας πιθανοτήτων που είναι η Απλή τυχαία δειγματοληψία, η Συστηματική Δειγματοληψία, η Δειγματοληψία με διαστρωμάτωση, η Δειγματοληψία κατά συστάδες και η Πολυσταδιακή Δειγματοληψία όπως επίσης και τις διάφορες μεθόδους της δειγματοληψίας χωρίς πιθανότητα που είναι η Δειγματοληψία ευκολίας ή συμπτωματική δειγματοληψία και η Δειγματοληψία αναλογίας ή ποσοστιαία δειγματοληψία. Στην Δειγματοληψία βάση πιθανοτήτων οι παρατηρήσεις του δείγματος επιλέγονται ανεξάρτητα και με ίσες πιθανότητες ενώ στην Δειγματοληψία χωρίς πιθανότητα η επιλογή των επιμέρους παρατηρήσεων που αποτελούν το δείγμα γίνεται με ένα σταθερό και προκαθορισμένο (συστηματικό) τρόπο.

Στη συνέχεια θα αναλύσουμε τις έννοιες χρονοσειρά, τροχιά και webclicks, έννοιες που αποτελούν χαρακτηριστικά παραδείγματα ακολουθιακών δεδομένων. Επίσης, θα αναφερθούμε σε μεθόδους data mining που στη συνέχεια θα χρησιμοποιηθούν για την αξιολόγηση των αποτελεσμάτων μας.

Θα γίνει μία αναφορά στις χρονοσειρές (Timeseries) και παρουσίαση κάποιων αντιπροσωπευτικών παραδειγμάτων, έτσι ώστε να γίνει πιο κατανοητή η έννοια των ακολουθιακών δεδομένων. Συγκεκριμένα, μία χρονοσειρά (Timeseries) είναι μια ακολουθία από σημεία δεδομένων, η οποία μετράται συνήθως σε διαδοχικές χρονικές στιγμές που απέχουν κατά ομοιόμορφα διαστήματα χρόνου μεταξύ τους. Οι παρατηρήσεις που αποτελούν μία χρονοσειρά παίρνονται σε ορισμένες χρονικές στιγμές ή περιόδους που ισαπέχουν μεταξύ τους και συλλέγονται μέσω κάποιας μεθόδου δειγματοληψίας.

Έπειτα, θα γίνει μία αναφορά στις τροχιές (Trajectories) και η παρουσίαση κάποιων αντιπροσωπευτικών παραδειγμάτων, έτσι ώστε να γίνει πιο κατανοητή η έννοια των καταγεγραμμένων διαδρομών ενός κινούμενου αντικειμένου (π.χ. οι καθημερινές διαδρομές που ακολουθούν τα εμπορικά φορτηγά στο κέντρο της Αθήνας).

Η δεύτερη ενότητα κλείνει με την ανάλυση των Webclicks που αποτελούν την καταγραφή του πλήθους των κλικ που γίνονται σε κάποια συγκεκριμένη ιστοσελίδα από τους διάφορους χρήστες. Είναι μία μέθοδος δειγματοληψίας με απώτερο σκοπό την διεξαγωγή μελέτης και την εξαγωγή συμπερασμάτων έτσι ώστε να ωφεληθούν οι ιδιοκτήτες των διαφόρων ιστοσελίδων.

Στην τρίτη ενότητα θα παρουσιάσουμε σχετικές μελέτες που έγιναν βασισμένες στις έννοιες της δειγματοληψίας τροχιών και χρονοσειρών. Επίσης, θα γίνει εφαρμογή των μεθόδων δειγματοληψίας σε

Πανεπιστήμιο Πειραιώς

ήδη συλλεγμένα δεδομένα έτσι ώστε να εξαχθούν κάποια συμπεράσματα και να αξιολογηθούν οι μέθοδοι που χρησιμοποιήθηκαν.

Τα άρθρα που έχουν επιλεγεί και θα συμπεριληφθούν στην εργασία μας υπό μορφή περίληψης είναι τα: «Segmentation and Sampling of Moving Object Trajectories Based on Representativeness», «Trajectory Sampling for Direct Traffic Observation» και «Unsupervised Trajectory Sampling». Τα τρία αυτά άρθρα αποτελούν σχετικές μελέτες βασισμένες στις έννοιες της τροχιάς και της δειγματοληψίας.

Τα δεδομένα που έχουμε προαναφέρει αφορούν τροχιές πλοίων, που καταγράφηκαν κατά τη διάρκεια τριών ημερών, και χρονοσειρές τιμών κλεισίματος μετοχών κατά τη διάρκεια τεσσάρων χρόνων. Αυτά τα δύο ήδη δεδομένων θα αναλυθούν και στη συνέχεια θα εφαρμοστούν μέθοδοι δειγματοληψίας σε αυτά Τα δεδομένα έχουν βρεθεί διαδικτυακά και έχουν τύχει επεξεργασίας ούτως ώστε να μπορούν να χρησιμοποιηθούν.

Τέλος, στην τελευταία μας ενότητα θα καταγραφούν τα γενικά αποτελέσματα και συμπεράσματα που έχουν εξαχθεί από τη μελέτη των δυο προηγούμενων κεφαλαίων.

Πανεπιστήμιο Πειραιώς

Abstract

The aim of this thesis is the design and adaptation of sampling techniques of complex data types, whose common feature is that they are time-stamped and sequential data. Firstly, a proper collection and recording of these time-stamped data through various methods of sampling is needed. Then, data will be processed and produce some results and conclusions.

The collection of huge amount of data can be done by various methods of sampling. More specifically, the method of sampling is the obtaining of a portion of data from a broader set of data and the categorization of them into two sub-groups, based on probability sampling and non-probability sampling.

In the second section, we will present the various methods of probability sampling, including the Simple random Sampling, the Systematic Sampling, the Stratified Sampling, the Cluster Sampling and the Multistage Sampling, as well as the various methods of sampling without probability, which are the Convenience Sampling or Random Sampling and the Ratio Sampling or Percentage Sampling. In probability sampling the observations of the sample are chosen independently and with equal chances, while in sampling without probability the selection of the individual observations which form the sample is made in a fixed and predetermined (systematic) way.

Next, we will analyze the terms of Timeseries, Trajectories and Webclicks, terms that constitute typical examples of sequential data.. In addition, we will refer to methods of data mining, which will then be used to evaluate our results.

There will be a reference to the timeseries and a presentation of some representative examples, in order to make the term of sequential data more understandable. In particular, a Timeseries is a sequence of data points, commonly measured in successive time points separated by equal intervals of time. The observations that represent a Timeseries are obtained at certain time points or periods of time, which equidistant from one another, and are collected through a sampling method.

Onwards, there will be a reference to the Trajectories and a presentation of some representative examples, in order to make the term of the recorded tracks of a moving object more understandable (e.g. daily routes followed by commercial trucks in the center of Athens).

The second section concludes with the analysis of the Webclicks, which constitute the recording of the number of clicks made at a specific site by the various users. It is a method of sampling with final aim the conduct of study and the drawing of conclusions, in order to benefit the owners of the various websites.

In the third section, we will present relevant studies based on the terms of sampling Trajectories and Timeseries. There will also be an application of the sampling methods in already collected data, in order to draw some conclusions and get in contact with the methods.

The selected articles which will be included in our project in a summary form are: «Segmentation and Sampling of Moving Object Trajectories Based on Representativeness», «Trajectory Sampling for Direct Traffic Observation» and «Unsupervised Trajectory Sampling». These three articles represent relevant studies based on the terms of trajectory and sampling. Two types of data will be used, analyzed and go through sampling methods. The first type includes the coordinates of ships recorded during three continuous days,

Πανεπιστήμιο Πειραιώς

and the second type consists of closing share prices for the last four years. The data to be used are found online and have been treated so that they can be used.

Finally, in the last section we will record the overall results and conclusions drawn from the study of the two previous chapters.

Πανεπιστήμιο Πειραιώς

Ευχαριστίες

Πρώτα απ’ όλα θα ήθελα να ευχαριστήσω τον επιβλέποντα της διπλωματικής μου εργασίας κο Νικόλαο Πελέκη, για την πολύτιμη βοήθεια του και καθοδήγηση του σε όλη τη διάρκεια της εργασίας μου. Επίσης θα ήθελα να ευχαριστήσω τα υπόλοιπα μέλη της επιτροπής κο Θεοδωρίδη και κο Κοφίδη. Ευχαριστώ τους φίλους μου για όλη τη στήριξη και την κατανόηση. Πάνω απ όλα ευχαριστώ την οικογενεία μου για ότιδηποτε έχουν προσφέρει ως σήμερα και για όλη τη στήριξη στις σπουδές μου.

Ευθυμία Κουφοπούλου

Πανεπιστήμιο Πειραιώς

ΕΝΟΤΗΤΑ 1

Εισαγωγή

1.1. Αντικείμενο της Διπλωματικής εργασίας.

Στην ενότητα αυτή θα εξηγήσουμε το αντικείμενο της παρούσας διπλωματικής εργασίας και θα μιλήσουμε γενικά για το τι πρόκειται να παρουσιαστεί στα επόμενα κεφάλαια. Στόχος αυτής της ενότητας είναι η επεξήγηση του σκοπού και του θέματος της εργασίας μας όπως επίσης και η αναφορά της δομής που θα ακολουθηθεί.

1.1. Αντικείμενο της Διπλωματικής εργασίας

Στόχος της παρούσας διπλωματικής εργασίας είναι η μελέτη και η προσαρμογή των διάφορων τεχνικών δειγματοληψίας σε σύνθετους τύπους δεδομένων. Οι τιμές που θα χρησιμοποιηθούν αποτελούν χρονοσημασμένα, ακολουθιακά δεδομένα, θα ληφθεί υπόψη δηλαδή το πότε (ημερομηνία και ώρα) εμφανίστηκαν και καταγράφηκαν.

Οι διάφορες τεχνικές δειγματοληψίας αναφέρονται στον τρόπο με τον οποίο μπορούν να συλλέγουν τα δεδομένα για κάποιον προκαθορισμένο σκοπό μελέτης. Η όποια μελέτη διεξάγεται αφορά δεδομένα με κοινά χαρακτηριστικά, όπως για παράδειγμα η καθημερινή πορεία που ακολουθούν τα πλοία κάποιας συγκεκριμένης εταιρείας μέσα στην περιοχή της Μεσογείου. Τα δεδομένα που συλλέγονται συνήθως είναι πολύ μεγάλου όγκου και για αυτό καταγράφονται και αποθηκεύονται σε βάσεις δεδομένων έτσι ώστε στη συνέχεια να μπορέσουν να τύχουν επεξεργασίας. Η δειγματοληψία αποτελεί το πρώτο βήμα για μία έρευνα ή μία δημοσκόπηση και χρησιμοποιείται για τους σκοπούς των τεχνικών εξόρυξης δεδομένων.

Αρχικά θα μιλήσουμε για τις χρονοσειρές οι οποίες είναι ακολουθίες από σημεία δεδομένων και συνήθως μετρούνται σε διαδοχικές χρονικές στιγμές που απέχουν κατά ομοιόμορφα διαστήματα χρόνου μεταξύ τους. Οι παρατηρήσεις που αποτελούν μία χρονοσειρά παίρνονται σε ορισμένες χρονικές στιγμές ή περιόδους που ισαπέχουν μεταξύ τους. Για το συγκεκριμένο είδος ακολουθιών θα χρησιμοποιηθούν οι τιμές κλεισίματος μετοχών του ελληνικού χρηματιστηρίου που λήφθηκαν μεταξύ το 2011 και του 2014. Οι τιμές αυτές αποτελούν ακολουθιακά δεδομένα στα οποία θα εφαρμοστούν τεχνικές δειγματοληψίας ούτως ώστε να παραχθούν αποτελέσματα που στη συνέχεια θα μπορέσουν να γενικευτούν για το σύνολο των μετοχών.

Στη συνέχεια θα μιλήσουμε για τις τροχιές που αποτελούν τη διαδρομή που ακολουθεί ένα κινούμενο αντικείμενο μέσα στο χώρο ως μία συνάρτηση του χρόνου. Το αντικείμενο θα μπορούσε να είναι ένα πλοίο που κάνει καθημερινές διαδρομές μέσα στη θάλασσα της Μεσογείου. Αξίζει να σημειωθεί ότι όταν μιλάμε για τροχιές μιλάμε για «αντικείμενα» που η κίνηση τους μπορεί να εντοπιστεί μέσω GPS ή έχουν πάνω τους ένα μηχανισμό GPS μέσω του οποίου καταγράφεται η κίνηση τους.

Πανεπιστήμιο Πειραιώς

Ακόμη θα αναφερθούμε σε μία σχετικά πρόσφατη καινοτόμα τεχνική διαχείρισης των κινήσεων των χρηστών του διαδικτύου που ονομάζεται Webclicks. Σε γενικές γραμμές, η μέθοδος αυτή αφορά την καταγραφή του πλήθους των κλικ που γίνονται σε κάποια συγκεκριμένη ιστοσελίδα από τους διάφορους χρήστες και αποτελεί εργαλείο για τη μέτρηση της κυκλοφορίας του Ιστού, ενώ μπορεί επίσης να χρησιμοποιηθεί και ως εργαλείο για τις επιχειρήσεις και την έρευνα της αγοράς, για να αξιολογηθεί και να βελτιωθεί η επισκεψιμότητα μιας ιστοσελίδας.

Όλα τα πιο πάνω είδη ακολουθιών και δεδομένων αποτελούν δεδομένα στα οποία μπορούν να συλλέγουν μέσω μεθόδων δειγματοληψίας και στη συνέχεια να αξιολογηθούν μέσω τεχνικών εξόρυξης δεδομένων όπως είναι για παράδειγμα η τεχνική Ομαδοποίησης (Clustering). Πιο συγκεκριμένα, εφόσον έχουμε τα αρχικά μας δεδομένα – είτε αυτά είναι τροχιές πλοίων είτε οτιδήποτε άλλο – εφαρμόζουμε μία μέθοδο mining (π.χ. Clustering) ούτως ώστε να απομείνει μόνο το δείγμα που μας ενδιαφέρει. Έπειτα εφαρμόζουμε ξανά την ίδια μέθοδο mining στο δείγμα μας και αναμένουμε να βρούμε όσο το δυνατόν ίδια μοτίβα έτσι ώστε να εξάγουμε γενικά συμπεράσματα.

Όπως προαναφέραμε, η ανάλυση των δεδομένων και η εξόρυξη δεδομένων αποτελεί ένα πεδίο πολύ ενδιαφέρον που βρίσκεται ακόμα σε εξέλιξη. Για το λόγο αυτό θα συμπεριληφθούν σχετικές μελέτες που αφορούν στη δειγματοληψία των τροχιών. Μέσω αυτών των μελετών θα δούμε πως η δειγματοληψία οδηγεί στην ανακάλυψη προτύπων συμπεριφοράς κινούμενων αντικειμένων και πως η διαδικασία αυτή μπορεί να εκμεταλλευτεί σε διάφορα πεδία.

Τέλος, θα κλείσουμε με το πρακτικό κομμάτι που θα αποτελείται από την δειγματοληψία δύο διαφορετικών ειδών δεδομένα. Και στις δύο περιπτώσεις θα ασχοληθούμε με χρονοσειρές, μέσω των οποίων θα προσπαθήσουμε να αξιολογήσουμε σχετικές τεχνικές σε τέτοιους τύπους δεδομένων. Τα δεδομένα μας αφορούν τροχιές πλοίων (πληροφορίες για τα σημεία στον χωροχρόνο κατά τη διάρκεια 3 ημερών μεταξύ των ημερομηνιών 31 Δεκεμβρίου 2008 και 02 Ιανουαρίου 2009) και ακολουθιακές τιμές κλεισίματος διαφόρων μετοχών από το 2011 μέχρι και το 2014.

Η μέθοδος που θα χρησιμοποιήσουμε για την αξιολόγηση των τεχνικών δειγματοληψίας είναι να χρησιμοποιήσουμε εξόρυξη δεδομένων (data mining) χρησιμοποιώντας τη μέθοδο της Ομαδοποίησης (Clustering) στα αρχικά δεδομένα και έπειτα θα κάνουμε το ίδιο mining στο δείγμα. Μετά από τη διπλή εφαρμογή της μεθόδου ομαδοποίησης περιμένουμε να βρούμε όσο το δυνατόν ίδια πρότυπα (patterns). Η δειγματοληψία που θα ακολουθήσουμε είναι σε επίπεδο ολόκληρης της ακολουθίας, διαλέγω δηλαδή κάποιες χρονοσειρές, ή τροχιές από το σύνολο και εφαρμόζω τις διάφορες μεθόδους.

Τα συμπεράσματα θα είναι μία σύνοψη των όσων θα αναφερθούν σε όλα τα προηγούμενα κεφάλαια και θα αποτελέσουν το κλείσιμο της παρούσας διπλωματικής εργασίας.

Πανεπιστήμιο Πειραιώς

ΕΝΟΤΗΤΑ 2

Βασικές Έννοιες

Στην ενότητα αυτή θα μιλήσουμε για τις πιο βασικές έννοιες που αφορούν το θέμα της διπλωματικής μας εργασίας, έτσι ώστε να γίνει μια εισαγωγή στο θέμα και να κατανοηθούν οι ορισμοί που θα χρησιμοποιηθούν στις επόμενες ενότητες. Αρχικά θα μιλήσουμε για τις διάφορες μεθόδους δειγματοληψίας, έπειτα για τις ακολουθίες που ονομάζονται χρονοσειρές, για τις τροχιές, τα «web-clicks» όπως επίσης και για τη μέθοδο «Clustering» - η οποία στη συνέχεια θα χρησιμοποιηθεί για την αξιολόγηση των μεθόδων δειγματοληψίας.

2.1. Δειγματοληψία

Εισαγωγή στη δειγματοληψία

Οι διάφορες μέθοδοι για τη συγκέντρωση στατιστικών στοιχείων μπορούν να χωριστούν σε δύο μεγάλες ομάδες, τις απογραφές και τις δειγματοληπτικές έρευνες.

Η δειγματοληψία ως διαδικασία αφορά τη λήψη ενός τμήματος στοιχείων από κάποιο ευρύτερο σύνολο. Το τμήμα του πληθυσμού που απογράφεται ονομάζεται δείγμα. Πιο συγκεκριμένα, για την καταγραφή χρονοσημασμένων δεδομένων χρειάζεται να γίνει η ορθή συλλογή τους μέσω κάποια δειγματοληψίας και στη συνέχεια να παραχθούν τα σωστά αποτελέσματα και συμπεράσματα μέσω της χρήσης τους. Για να μπορέσει η δειγματοληψία να θεωρηθεί επιτυχής πρέπει η επιλογή του δείγματος να παράγει αποτελέσματα τα οποία να μπορούν να είναι γενικεύσιμα για το σύνολο του πληθυσμού.

Σκοπός των δειγματοληπτικών ερευνών είναι ο όσο το δυνατόν πιο ακριβής προσδιορισμός των ιδιοτήτων του πληθυσμού, μελετώντας τα στοιχεία του δείγματος. Συνήθως οι εκτιμήσεις για τις ιδιότητες του πληθυσμού αποτελούν προσεγγίσεις που περιέχουν κάποιο σφάλμα. Η γενίκευση των πληροφοριών του δείγματος σε ολόκληρο τον πληθυσμό συνεπάγεται αβεβαιότητα η οποία και μπορεί να μετρηθεί υπό την προϋπόθεση ότι το δείγμα είναι τυχαίο. [3]

Τα δεδομένα της δειγματοληψίας θα πρέπει να προέρχονται από το σωστό σύνολο, να μαζευτούν προσεγμένα, να εξασφαλιστεί η ορθότητα τους και να αντιπροσωπεύουν τον πληθυσμό από τον οποίο επιλέχθηκαν. Ένα δείγμα πρέπει να είναι αντιπροσωπευτικό, δηλαδή πρέπει να εκφράζει τις

Πανεπιστήμιο Πειραιώς

διαφοροποιήσεις του πληθυσμού όσο γίνεται πιο πιστά έτσι ώστε να αποτελεί αξιόπιστο αντικαταστάτη του πληθυσμού. Για να επιτευχθεί αυτό, είναι σημαντικό να αποφεύγεται η μεροληπτικότητα (bias) που είναι αποτέλεσμα της επιλογής για δειγματοληψία μη αντιπροσωπευτικών τμημάτων του πληθυσμού.

Επιπρόσθετα, για την επιλογή του κατάλληλου δειγματοληπτικού υποβάθρου – του στατιστικού πληθυσμού από όπου λαμβάνεται το δείγμα – απαιτείται κρίση εφόσον τα αποτελέσματα μιας δειγματοληψίας αναφέρονται αποκλειστικά και μόνο στον πληθυσμό από τον οποίο έχει ληφθεί το δείγμα. Η επιλογή ενός δειγματοληπτικού υποβάθρου το οποίο περιέχει ολόκληρο το φάσμα των μετρήσεων και όλα τα είδη των απαριθμήσεων του υπό εξέταση πληθυσμού, είναι το πρώτο και βασικό βήμα σε μια δειγματοληψία.

Η μέθοδος αυτή αναζητά πληροφόρηση από ανθρώπους, π.χ. προεκλογική σφυγμομέτρηση, πορείες κινούμενων αντικειμένων στο χωρόχρονο, πλοήγηση χρηστών σε ιστοσελίδες του διαδικτύου και ακολουθίες αγοραπωλησιών. Επίσης, σε μια δειγματοληψία το Ποσοστό Ανταπόκρισης αναφέρεται στην αναλογία των ανθρώπων που συμμετέχει στην σφυγμομέτρηση και είναι πολύ βασική παράμετρος. Οι σφυγμομετρήσεις μπορούν να γίνουν με ποικίλους τρόπους όπως για παράδειγμα μέσω μίας προσωπικής/τηλεφωνικής συνέντευξης ή μέσω ερωτηματολογίων.

Βασικοί Όροι Δειγματοληψίας

Κατά τη διαδικασία της δειγματοληψίας χρησιμοποιούνται οι ακόλουθες ορολογίες:

Απλό στοιχείο (elementary): ορίζεται κάθε μονάδα του συνόλου στην οποία γίνεται μια διαδικασία μέτρησης ή παρατήρησης κάποιας ιδιότητας και η καταγραφή των αποτελεσμάτων.

Πληθυσμός (population): ορίζεται ένα βασικό σύνολο στοιχείων που πρόκειται να μελετηθεί ως προς μια ή περισσότερες χαρακτηριστικές ιδιότητες. Είναι το σύνολο δηλαδή των απλών στοιχείων για τα οποία μας ενδιαφέρει να βγάλουμε συμπεράσματα.

Δειγματοληπτική μονάδα (sampling unit): ορίζεται το στοιχείο ή η συλλογή στοιχείων που μπορεί να επιλεγεί σε κάποιο στάδιο της δειγματοληψίας. Αν χρησιμοποιήσουμε απλή δειγματοληψία, η δειγματοληπτική μονάδα είναι το στοιχείο με την έννοια που ορίστηκε παραπάνω. Αν όμως η δειγματοληψία εκτελείται σε στάδια και επιλέγουμε πρώτα μια ομάδα στοιχείων και κατόπιν κάποια στοιχεία μέσα από την ομάδα, τότε η ομάδα είναι η δειγματοληπτική μονάδα.

Δειγματοληπτικό πλαίσιο: Στην απλούστερη του μορφή είναι μια λίστα με στοιχεία που καλύπτουν τον ερευνώμενο πληθυσμό. Μπορεί να αποτελεί μια φυσική λίστα όπως ένας κατάλογος, μπορεί όμως να είναι και ένα οικοδομικό σχέδιο μιας πόλης ή ακόμη και μια εννοιολογική λίστα π.χ. πόσα αυτοκίνητα πάρκαραν σε ένα σημείο για κάποιο συγκεκριμένο χρονικό διάστημα. Ο προσδιορισμός του πλαισίου παίζει σημαντικό ρόλο στο σχεδιασμό μιας δειγματοληπτικής έρευνας.

Μονάδα παρατήρησης: Έστω ότι κάνουμε μια δειγματοληπτική έρευνα σε νοικοκυριά και συλλέγουμε στοιχεία για το σύνολο κάθε νοικοκυριού. Τότε το νοικοκυριό είναι η μονάδα παρατήρησης.

Μέγεθος δείγματος: ορίζεται το πλήθος των στοιχείων που διαμορφώνουν το δείγμα και συμβολίζεται με το γράμμα n.

Δείγμα: ορίζεται μια συλλογή δειγματοληπτικών μονάδων από το πλαίσιο. Τόσο το μέγεθος όσο και ο τρόπος επιλογής του αποτελούν βασική προϋπόθεση για την επιτυχία μιας δειγματοληπτικής έρευνας.

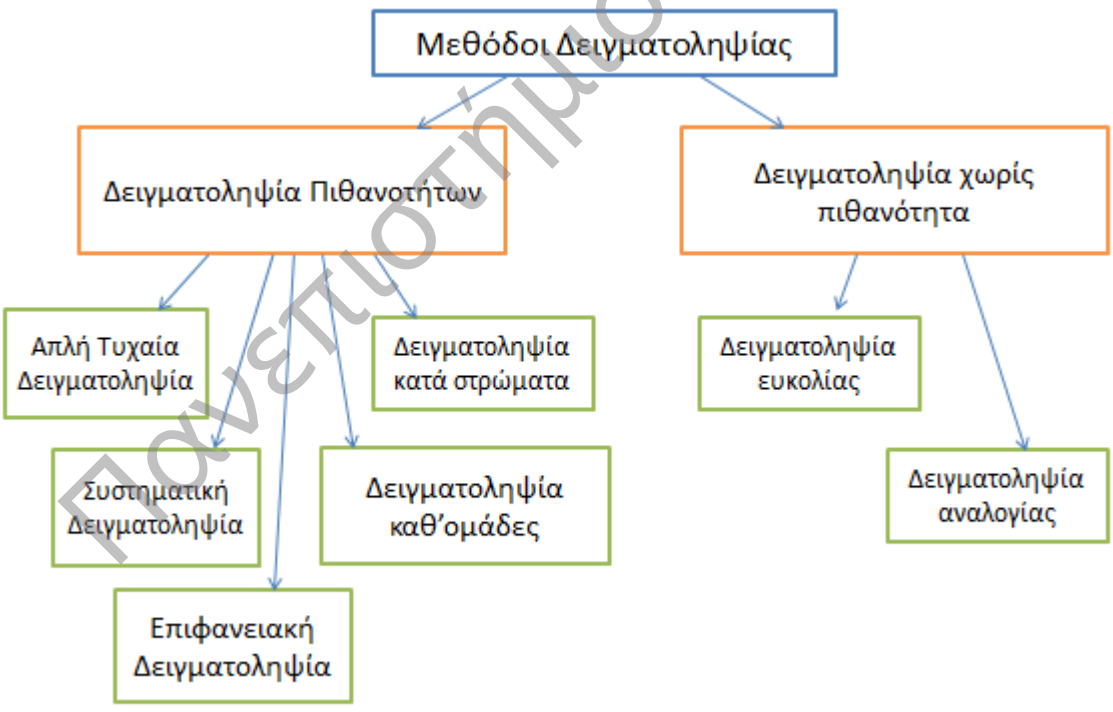
Πανεπιστήμιο Πειραιώς

Δειγματοληπτική Διαδικασία

Για την κατάλληλη επιλογή του δειγματοληπτικού υπόβαθρου χρειάζεται η διαδικασία επιλογής των παρατηρήσεων που θα αποτελέσουν το δείγμα. Αυτή η διαδικασία επιλογής διακρίνεται σε 2 βασικά είδη δειγματοληψίας: την τυχαία δειγματοληψία ή αλλιώς δειγματοληψία με πιθανότητα (Probability Sampling) και τη συστηματική δειγματοληψία ή αλλιώς δειγματοληψία χωρίς πιθανότητα (Nonprobability sampling).

Από τη μία, στην τυχαία δειγματοληψία οι παρατηρήσεις του δείγματος επιλέγονται ανεξάρτητα και με ίσες πιθανότητες εφόσον πρώτο τα μέλη ενός πληθυσμού έχουν ίσες πιθανότητες να επιλεγούν για το δείγμα ενώ η επιλογή ενός μέλους του πληθυσμού για το δείγμα με κανένα τρόπο δεν επηρεάζει την επιλογή ενός οποιουδήποτε άλλου. Αυτού του είδους η δειγματοληψία βασίζεται στους νόμους των πιθανοτήτων και είναι ελεγχόμενη ως προς τις παραμέτρους της και δίνει τη δυνατότητα να γενικευτούν τα συμπεράσματα που εξάγονται από ένα δείγμα έτσι ώστε να υπολογίσουμε το σφάλμα εκτίμησης (της γενίκευσης).

Από την άλλη, στη συστηματική δειγματοληψία η επιλογή των επιμέρους παρατηρήσεων που αποτελούν το δείγμα γίνεται με ένα σταθερό και προκαθορισμένο (συστηματικό) τρόπο. Συγκεκριμένα, σε αυτή τη διαδικασία το διάστημα μεταξύ των παρατηρήσεων είναι σταθερό και καθορισμένο και δεν γίνεται χρήση τυχαίων αριθμών. Αυτού του είδους η δειγματοληψία εφαρμόζεται σε περιπτώσεις που δεν είναι εφικτή η δειγματοληψία με πιθανότητα ή όταν θέλουμε να γίνει γρήγορα μια εφαρμογή της έρευνας. Τα αποτελέσματα μιας έρευνας που γίνονται με αυτή τη μέθοδο δεν είναι γενικεύσιμα και δεν μπορούν να μας δώσουν πληροφορίες για το σύνολο του πληθυσμού.



Εικόνα 1 - Είδη Δειγματοληψίας

Πανεπιστήμιο Πειραιώς

Δειγματοληψία με πιθανότητα

Απλή Τυχαία ή Απεριόριστη Δειγματοληψία (Simple Random Sampling) [4]

Η απλή τυχαία δειγματοληψία ή αλλιώς Απεριόριστη Δειγματοληψία χρησιμοποιείται εκτεταμένα λόγω της απλότητάς στη χρήση και στην παραγωγή αποτελεσμάτων και συμπερασμάτων. Επίσης, πέρα από την αυτοτελή χρήση του, το σχήμα αυτό χρησιμεύει και ως βάση για συνθετότερα δειγματοληπτικά σχήματα, όπως, για παράδειγμα, η στρωματοποιημένη απλή τυχαία δειγματοληψία (stratified simple random sampling) και η δειγματοληψία κατά ομάδες (cluster sampling).

Σε αυτή τη μεθοδολογία όλα τα στοιχεία του πληθυσμού θα πρέπει να είναι καταγραμμένα σε έναν κατάλογο, ο οποίος θα χρησιμοποιηθεί ως το δειγματοληπτικό πλαίσιο. Το δείγμα λαμβάνεται από αυτό το ενιαίο πλαίσιο, με την αντιστοίχιση των αριθμών στα μέλη του καταλόγου (εφόσον δεν υπάρχει ήδη). Στη συνέχεια επιλέγονται τυχαία από τον κατάλογο τόσα μέλη ώστε να σχηματιστεί πλήθος ίσο με το μέγεθος του δείγματος που χρειαζόμαστε. Όλα τα μέλη της λίστας έχουν την ίδια πιθανότητα να εκλεγούν εφόσον κάθε μη εκλεγείσα μονάδα έχει ίση πιθανότητα εκλογής με τις υπόλοιπες μονάδες. Πιο αναλυτικά, εάν μία μονάδα δεν επιλεγεί σε κάποια από τις διαδοχικές επιλογές, εξακολουθεί να έχει ίσες πιθανότητες εκλογής με τις υπόλοιπες μονάδες στην επόμενη επιλογή.

Περιγραφή της Διαδικασίας

Οι τρόποι λήψης του δείγματος μπορεί να γίνει με δύο διαφορετικούς τρόπους: με τη μέθοδο των λαχνών και με τη μέθοδο των τυχαίων αριθμών.

Με βάση την πρώτη μέθοδο αριθμούμε τις μονάδες του πλαισίου και δημιουργούμε λαχνούς με κάθε ένα από τους αριθμούς αυτούς. Στη συνέχεια, τοποθετούμε όλους τους λαχνούς σε μια κληρωτίδα τους ανακατεύουμε κι έπειτα επιλέγουμε διαδοχικά όσους λαχνούς χρειαζόμαστε, ανάλογα με το μέγεθος του δείγματος. Έτσι, το τελικό δείγμα αποτελείται από όλες τις μονάδες που έχουν επιλεγεί τυχαία κατά την κλήρωση.

Στη δεύτερη μέθοδο χρησιμοποιούνται οι πίνακες των τυχαίων αριθμών, όπου οι αριθμοί έχουν επιλεγεί με κάποιο τρόπο κλήρωσης και μπορούμε να τους διαβάσουμε με οποιοδήποτε τυχαίο τρόπο. Ο τρόπος επιλογής των αριθμών επιλέγεται εκ των προτέρων (για παράδειγμα μπορεί να επιλέγονται οι αριθμοί της πρώτης στήλης για κάθε δεύτερη σελίδα). Στη συνέχεια επιλέγουμε τόσους αριθμούς όσο είναι και το μέγεθος του δείγματος. Επίσης, οι αριθμοί επιλέγονται με τόσα ψηφία, όσα έχει και ο τελευταίος αριθμός στο πλαίσιο (π.χ. αν ο τελευταίος αριθμός του καταλόγου είναι ο 4.785, θα λαμβάνονται τετραψήφιοι αριθμοί, από 0001 μέχρι 4.785, όπως 0894, 0076, 1854, κλπ.)

Στις περιπτώσεις όπου το μέγεθος N του πληθυσμού είναι πολύ μεγάλο ο σχηματισμός όλων των δυνατών διακεκριμένων δειγμάτων δεν είναι τόσο εύκολος.

Για τη διαδικασία της δειγματοληψίας ακολουθείται η εξής εναλλακτική διαδικασία: Μια μονάδα του πληθυσμού επιλέγεται τυχαία, δηλαδή με τρόπο που εξασφαλίζει την ίδια πιθανότητα επιλογής σε κάθε μια από τις N μονάδες του πληθυσμού. Η μέθοδος αυτή αποτελείται από τα εξής βήματα:

1) Αντιστοιχίζουμε σε κάθε μονάδα του πληθυσμού έναν αριθμό από το 1 μέχρι το N και
2) διαλέγουμε μια σειρά n τυχαίων αριθμών από το 1 μέχρι το N με την βοήθεια πινάκων τυχαίων αριθμών.

Πανεπιστήμιο Πειραιώς

Πίνακες τυχαίων αριθμών

Η επιλογή ενός απλού τυχαίου δείγματος γίνεται συνήθως με την βοήθεια πινάκων τυχαίων αριθμών. Αυτοί είναι πίνακες αποτελούνται από τα ψηφία 0-9 και η πιθανότητα επιλογής σε οποιαδήποτε δοκιμή είναι η ίδια, δηλαδή 1/10 για το κάθε ψηφίο.

Για την επιλογή ενός απλού τυχαίου δείγματος μεγέθους n από ένα πληθυσμό μεγέθους N, αντιστοιχίζουμε σε κάθε μια από τις μονάδες του πληθυσμού έναν αριθμό από το 1 μέχρι το N (διαφορετικό για κάθε μονάδα). Διαλέγουμε τυχαία τόσες στήλες όσα τα ψηφία του N και διαβάζουμε προς μια κατεύθυνση, π.χ. προς τα κάτω την συγκεκριμένη ομάδα στηλών επιλέγοντας τους αριθμούς που είναι ≤N.

Βασισμένοι στο απόσπασμα των 1000 ψηφίων από τον πίνακα των τυχαίων αριθμών των Snedecor και Cochran (1967) [1] θα εξηγήσουμε αναλυτικά την διαδικασία.

Έστω N=198 και n=5. Έστω ότι διαλέγουμε τις στήλες 10-12. Ξεκινώντας από την γραμμή 00 και διαβάζοντας προς τα κάτω (εφόσον έχουμε διαλέξη αυτή την κατεύθυνση), οι πρώτοι 5 διακεκριμμένοι αριθμοί είναι το 188, 112, 106, 108 και το 72.

Πανεπιστήμιο Πειραιώς

Τυχαίοι αριθμοί

	00-04	05-09	10-14	15-19	20-24	25-29	30-34	35-39	40-44	45-49
00	54463	22662	65905	70639	79365	67382	29085	69831	47058	08186
01	15389	85205	18850	39226	42249	90669	96325	23248	60933	26927
02	85941	40756	82414	02015	13858	78030	16269	65978	01385	15345
03	61149	69440	11286	88218	58925	03638	52862	62733	33451	77455
04	05219	81619	10651	67079	92511	59888	84502	72095	83463	75577
05	41417	98326	87719	92294	46614	50948	64886	20002	97365	30976
06	28357	94070	20652	35774	16249	75019	21145	05217	47286	76305
07	17783	00015	10806	83091	91530	36466	39981	62481	49177	75779
08	40950	84820	29881	85966	62800	70326	84740	62660	77379	90279
09	82995	64157	66164	41180	10089	41757	78258	96488	88629	37231
10	96754	17676	55659	44105	47361	34833	86679	23930	53249	27083
11	34357	88040	53364	71726	45690	66334	60332	22554	90600	71113
12	06318	37403	49927	57715	50423	67372	63116	48888	21505	80182
13	62111	52820	07243	79931	89292	84767	85693	73947	22278	11551
14	47534	09243	67879	00544	23410	12740	02540	54440	32949	13491
15	98614	75993	84460	62846	59844	14922	48730	73443	48167	34770
16	24856	03648	44898	09351	98795	18644	39765	71058	90368	44104
17	96887	12479	80621	66223	86085	78285	02432	53342	42846	94771
18	90801	21472	42815	77408	37390	76766	52615	32141	30268	18106
19	55165	77312	83666	36028	28420	70219	81369	41943	47366	41067

Εικόνα 2 - Απόσπασμα από τον πίνακα τυχαίων αριθμών των Snedecor και Cochran (1967)

Πανεπιστήμιο Πειραιώς

Παράδειγμα 1 [5]:

Αρχικά καθορίζουμε τον πληθυσμός έτσι ώστε να μπορούμε να σχηματίσουμε το τελικό δείγμα. Υποθέτουμε λοιπόν ότι έχουμε ένα καθορισμένο πληθυσμό μεγέθους 5.

Πληθυσμός: {1, 2, 3, 4, 5}

Με βάση τον πιο πάνω προκαθορισμένο πληθυσμό τα πιθανά τελικά δείγματα μεγέθους 2 είναι τα εξής 10:

{1, 2} , {1, 3} , {1, 4} , {1, 5} , {2, 3} , {2, 4} , {2, 5} , {3, 4} , {3, 5} , {4, 5}

Γενικότερα, αν ο πληθυσμός αποτελείται από N μονάδες και επιθυμούμε να σχηματίσουμε δείγμα μεγέθους n, το πλήθος των δυνατών διακεκριμένων δειγμάτων είναι:

$$\binom{N}{n} = \frac{N!}{n!(N-n)!} = \frac{N(n-1)\ldots(N-n+1)}{n},$$

όπου n! = 1 × 2 × 3 × … × n. Η διαδικασία κατά την οποία επιλέγεται ένα δείγμα μεγέθους n μέσα από έναν πληθυσμό μεγέθους N ονομάζεται απλή τυχαία δειγματοληψία εάν κάθε ένα από τα δείγματα έχει πιθανότητα επιλογής ίση με $1/\binom{N}{n}$

Παράδειγμα 2:

Από ένα σύνολο 1000 φορολογικών δηλώσεων χρειάζεται να επιλεγούν τυχαία οι 20 φορολογικές δηλώσεις προσώπων για να εξεταστούν. Για να επιλεγεί ένα τυχαίο δείγμα από n=20 φορολογικές δηλώσεις από ένα σύνόλο N=1000 φορολογικές δηλώσεις τότε είτε:

1) Γράφουμε τους αριθμούς φορολογικών δηλώσεων σε 1000 κλήρους τους οποίους ρίχνουμε σε ένα κουτί. Επιλέγουμε διαδοχικά χωρίς να κοιτάμε 20 κλήρους χωρίς επανατοποθέτηση κι έχουμε το ζητούμενο τυχαίο δείγμα. (Είναι προφανές πως η συγκεκριμένη τεχνική δεν είναι και τόσο εφαρμόσιμη σε δειγματοληψίες μεγάλου μεγέθους.)
2) Χρησιμοποιούμε έναν πίνακα τυχαίων αριθμών από τον οποίο επιλέγουμε το τυχαίο δείγμα.

Για την επίλυση του πιο πάνω προβήματος θα χρησιμοποιήσουμε την δεύτερη προτεινόμενη λύση. Παίρνουμε ένα παράρτημα ενός πίνακα τυχαίων αριθμλων έτσί ώστε να επιλεγεί το δείγμα των 20 τυχαιίων φορολογικών δηλώσεων.

Αριθμούμε από 1 έως το 1000 τις φορολογικές δηλώσεις πάνω σε μια λίστα όπου είναι κατεγραμμένοι οι αριθμοί φορολογικών δηλώσεων (σε μία αντιστοιχία 1 προς 1). Ξεκινώντας από μια τυχαία θέση στον πίνακα κινούμαστε προς μια τυχαία συγκεκριμένη κατεύθυνση διαβάζοντας αριθμούς. Από αυτή μας την επιλογή προκύπτουν οι πιο κάτω τυχαίοι αριθμοί:

45	126	600	3	85	103	666	32	99	9	454	83	12	23	72	289	509	703	846	902

Υποσημείωση: Αγνοούνται οι αριθμοί που είναι μεγαλύτεροι από το 1000, εάν προκύψουν.

Πλεονεκτήματα και Μειονεκτήματα

<u>Πλεονεκτήματα</u>

- Η απλή τυχαία διαδικασία δειγματοληψίας είναι απλή
- Εύκολη μέτρηση του τυπικού (δειγματοληπτικού) σφάλματος και των ορίων αξιοπιστίας.

Πανεπιστήμιο Πειραιώς

Μειονεκτήματα

- Είναι απαραίτητος ο πλήρης κατάλογος των μονάδων ή των ατόμων του πληθυσμού.
- Το τελικό δείγμα δεν είναι πάντα πλήρης αντιπροσωπευτικό εφόσον επιλέγεται κατά τύχης.
- Ενδεχομένως να είναι δύσκολη και πολυδάπανη η προσέγγιση λόγω της διασπορά των μονάδων του πληθυσμού.

Συστηματική Δειγματοληψία (Systematic Random Sampling)

Η συγκεκριμένη δειγματοληπτική τεχνική εισάγει ένα συστηματικό στοιχείο στην διαδικασία επιλογής των μονάδων του πληθυσμού, για αυτό το λόγο και ονομάζεται συστηματική δειγματοληψία. Στις περιπτώσεις όπου ένα δειγματοληπτικό πλαίσιο είναι διαθέσιμο υπό τη μορφή λίστας ή καταλόγου μπορεί να εφαρμοστεί η συστηματική δειγματοληψία. Αυτός ο τρόπος δειγματοληψίας είναι ταχύτερος και ευκολότερος τις πλύστες φορές. Διεξάγεται μέσω της χρήσης ενός ήδη δημιουργημένου καταλόγου ξεκινώντας από κάποιο τυχαίο αρχικό σημείο και επιλέγοντας μια μονάδα κάθε k (k>0) μονάδες μέχρι να κατασκευασθεί το δείγμα με το δοθέν μέγεθος.

Για παράδειγμα, εάν πρόκειται να επιλεγούν 1000 τηλεθεατές από μία λίστα που περιέχει 10000 εγγραφές (ονοματεπώνυμο και τηλέφωνο), είναι ταχύτερο να επιλεγεί ένας τυχαίος αριθμός μεταξύ του 1 και του 10 και να περιληφθεί στο δείγμα ο αριθμός τηλεφώνου που αντιστοιχεί σ' αυτόν τον αριθμό καθώς και κάθε δέκατη καρτέλα από εκεί και πέρα, από το να επιλεγούν 1000 τυχαίοι αριθμοί και να περιληφθούν οι καρτέλες που αντιστοιχούν σε αυτούς.

Περιγραφή της Διαδικασίας

Σε έναν πληθυσμό μεγέθους N, οι μονάδεις είναι αριθμημένες από τον αριθμό 1 μέχρι και τον αριθμό N.

Έστω k ένας θετικός ακέραιος.

Για την επιλογή ενός 1-ανά-k συστηματικού δείγματος μεγέθους n, διαλέγουμε τυχαία μια μονάδα και περιλαμβάνουμε στο δείγμα αυτήν και κάθε μονάδα του πληθυσμού που απέχει από αυτήν κατά κάποιο πολλαπλάσιο του k. Η επιλογή της πρώτης μονάδας καθορίζει ολόκληρο το δείγμα. Για παράδειγμα, αν ο πληθυσμός αποτελείται από τις τιμές y1, y2, …, yN και από τις πρώτες k μονάδες του επιλεγεί η k0, τότε το δείγμα θα αποτελείται από τις μονάδες yk0, yk0 + k, … , yk0 + (n-1)k. Πιο πριν έχουμε δει και το παράδειγμα των τηλεθεατών που χρησιμοποιεί ενδεικτικούς πραγματικούς αριθμούς.

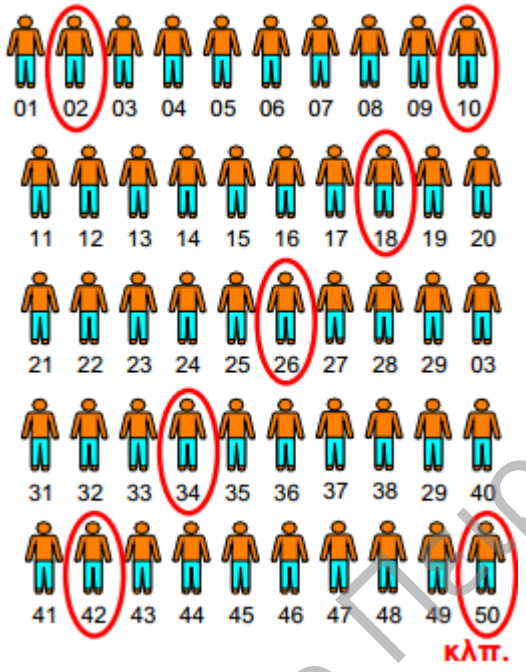
Παράδειγμα [6]

Σκοπός της δειγματοληψίας είναι να εκτιμηθεί η κάλυψη για ιλαρά με εμβολιασμό στα παιδιά που προσέρχονται στο κέντρο υγείας της κωμόπολης Α εφόσον γνωρίζουμε ότι εισέρχονται περίπου 80 παιδιά την ημέρα.

1. Αρχικά ορίζουμε τον πληθυσμού δειγματοληψίας: επισκέψεις των παιδιών στο κέντρο υγείας μεταξύ 1η Ιουνίου και 31 Ιουνίου (80x25=> N = 2000 παιδιά)
2. Έπειτα καθορίζουμε το μέγεθος του δείγματος που επιθυμούμε: n = 250 παιδιά
3. Στη συνέχεια υπολογίζουμε το δειγματοληπτικό κλάσμα (2000/250 => k = 8)
4. Μετά επιλέγουμε το πρώτο παιδί το οποίο θα ληφθεί στο δείγμα με έναν τυχαίο αριθμό από το 1 έως 8 (π.χ. το παιδί 2)

Πανεπιστήμιο Πειραιώς

5. Τέλος, γνωρίζουμε ότι κάθε 8ο παιδί μετά το πρώτο που επιλέγηκε θα συμπεριλαμβάνεται στο δείγμα. (Ακολουθεί εικόνα η οποία απεικονίζει μέρος του σεναρίου που περιγράφηκε)



Εικόνα 3 - Παράδειγμα συστηματικής δειγματοληψίας

Πλεονεκτήματα και Μειονεκτήματα

Πλεονεκτήματα

- Δεν είναι απαραίτητος ο πλήρης κατάλογος των μονάδων του πληθυσμού της δειγματοληψίας
- Είναι μια απλή διαδικασία
- Η μέτρηση του τυπικού σφάλματος και των ορίων αξιοπιστίας είναι εύκολη.

Μειονεκτήματα

- Ανάλογα με τον κανόνα επιλογής του δείγματος ή των χαρακτηριστικών της ακολουθίας των ατόμων, μπορεί να υπάρξει απόκλιση από την αντιπροσωπευτική επιλογή του δείγματος.
- Ενδεχομένως να είναι δύσκολη και πολυδάπανη η προσέγγιση λόγω της διασποράς των μονάδων του πληθυσμού.

Δειγματοληψία με διαστρωμάτωση (Stratified Random Sampling)

Η δειγματοληψία διαστρωμάτωσης είναι η διαδικασία της διαίρεσης των μελών του πληθυσμού σε ομογενείς υποομάδες πριν από τη δειγματοληψία. Τα στρώματα πρέπει να αλληλοαποκλείονται, δηλαδή κάθε στοιχείο του πληθυσμού θα πρέπει να ανατεθεί σε ένα μόνο στρώμα έτσι ώστε να μην υπάρχει πιθανότητα επαναχρησιμοποίησης. Τα στρώματα θα πρέπει επίσης να είναι συλλογικά εξαντλητικά: κανένα στοιχείο του πληθυσμού δεν μπορεί να αποκλειστεί. Στη συνέχεια, εφαρμόζεται απλή τυχαία δειγματοληψία ή συστηματική δειγματοληψία μέσα σε κάθε ένα από τα στρώματα που έχουν καθοριστεί

Πανεπιστήμιο Πειραιώς

πιο πριν. Αυτό συχνά βελτιώνει την αντιπροσωπευτικότητα του δείγματος, μειώνοντας το δειγματοληπτικό σφάλμα.

Συνοπτικά, η συγκεκριμένη μέθοδος δειγματοληψίας γίνεται μέσω χωριστικής διαδικασίας εφόσον διαχωρίζουμε και καθορίζουμε τα υποσύνολα του πληθυσμού έτσι ώστε να μπορούμε να συλλέξουμε στοιχεία και να εξάγουμε αποτελέσματα.

Περιγραφή της Διαδικασίας

Σε πρώτο στάδιο γίνεται χωρισμός του πληθυσμού σε «στρώματα» (κατά προτίμηση με εσωτερική ομοιογένεια αλλά ετερογενή μεταξύ τους)

Ακολουθεί λήψη τυχαίου δείγματος μονάδων σε κάθε ένα από τα προεπιλεγμένα «στρώματα»

Τέλος γίνεται ένας συνδυασμός των αποτελεσμάτων όλων των «στρωμάτων» έτσι ώστε να παρθούν τα αποτελέσματα.

Παράδειγμα [6]

Σε αυτό το παράδειγμα θέλουμε να εκτιμήσουμε την κάλυψη με εμβολιασμό για ιλαρά στα παιδιά που φοιτούν στην Α΄ Δημοτικού στο σύνολο της Ελλάδας.

1. Σαν πρώτο βήμα γίνεται ο χωρισμός του πληθυσμού σε «στρώματα», δηλαδή κατά περιφέρεια της χώρας.
2. Μετά γίνεται λήψη τυχαίου δείγματος παιδιών σε κάθε περιφέρεια.
3. Τέλος, ακολουθεί συνδυασμός των αποτελεσμάτων όλων των περιφερειών.

Ακολουθεί πίνακας με τις μετρήσεις που αναφέρθηκαν πιο πριν. Οι μετρήσεις έγιναν το 2006 σε διάφορες περιοχές της Ελλάδας.

Γεωγραφικό διαμέρισμα	Αριθμός Σχολικών τμημάτων	Αριθμός παιδιών	Δειγματοληπτικό κλάσμα (τμήματα)
Ήπειρος – Ιόνια νησιά	59	611	12,9%
Κρήτη – Νησιά Αιγαίου	55	605	5,1%
Μακεδονία – Θεσσαλία	46	639	2,0%
Αττική	50	840	2,8%
Στερεά - Πελοπόννησος	58	608	3,5%
Θράκη	79	575	19,6%
ΣΥΝΟΛΟ	342	3.878	4,5%

Εικόνα 4 - Παράδειγμα δειγματοληψίας με διαστρωμάτωση.

Πλεονεκτήματα και Μειονεκτήματα

Πλεονεκτήματα

- Υπάρχει μεγαλύτερη ακρίβεια εκτιμήσεων εφόσον υπάρχει μικρότερο τυχαίο σφάλμα και στενότερα όρια αξιοπιστίας

Πανεπιστήμιο Πειραιώς

- Λόγω της επαρκής αντιπροσώπευσης των μονάδων/ατόμων από τα επιμέρους «στρώματα», γίνονται εκτιμήσεις για κάθε «στρώμα».

Μειονεκτήματα

- Ο υπολογισμός του τυπικού σφάλματος και των ορίων αξιοπιστίας είναι πολύπλοκος
- Υπάρχει απώλεια ακρίβειας εάν ο αριθμός των μονάδων/ατόμων στα επιμέρους «στρώματα» είναι μικρός.

Δειγματοληψία κατά συστάδες (Cluster Sampling)

Η δειγματοληψία κατά συστάδες είναι μια τεχνική δειγματοληψίας η οποία χρησιμοποιείται όταν οι «φυσικές» ομάδες είναι εμφανής ότι βρίσκονται σε στατιστικό πληθυσμό. Σε αυτή την τεχνική δειγματοληψίας ο συνολικός πληθυσμός διαιρείται σε «φυσικές» ομάδες (ή συστάδες) και επιλέγεται ένα δείγμα από αυτές. Χρησιμοποιούνται συχνά στην έρευνα μάρκετινγκ.

Το αρχικό στάδιο για την επιλογή του δείγματος περιλαμβάνει την επιλογή των συστάδων/ομάδων από μονάδες (άτομα).

Περιγραφή της Διαδικασίας

1. Αρχικά γίνεται ταξινόμηση του πληθυσμού σε συστάδες (=ομάδες) από μονάδες (άτομα).
2. Στη συνέχεια γίνεται λήψη τυχαίου δείγματος συστάδων.
3. Τέλος, το σύνολο ή μέρος των μονάδων (ατόμων) από τις επιλεγμένες συστάδες περιλαμβάνεται στο τελικό δείγμα.

Παράδειγμα [6]

Σε αυτό το παράδειγμα δειγματοληψίας θέλουμε να εκτιμήσουμε την κάλυψη με εμβολιασμό για ιλαρά στα παιδιά της Α΄ Δημοτικού του Νομού Α, όπου υπάρχουν 150 Δημοτικά σχολεία συνολικά (αρχικές δειγματοληπτικές μονάδες).

Αρχικά ταξινομούμε «φυσικά» τα παιδιά σε σχολεία: 150 σχολεία x περίπου 20 μαθητές ανά σχολείο = 3.000 μαθητές συνολικά.

Ακολουθεί λήψη τυχαίου δείγματος σχολείων: για παράδειγμα επιλέγουμε 30 από τα 150 σχολεία και χρειάζεται κατάλογος μόνο για τα συγκεκριμένα.

Στη συνέχεια γίνεται λήψη δείγματος 10 παιδιών από κάθε σχολείο που επελέγη. Συνολικά έχουμε 30 σχολεία x 10 μαθητές από κάθε σχολείο = 300 μαθητές συνολικά αποτελούν το τελικό δείγμα και χρειάζεται κατάλογος μαθητών μόνο από τα επιλεγμένα σχολεία.

Πλεονεκτήματα και Μειονεκτήματα

Πλεονεκτήματα

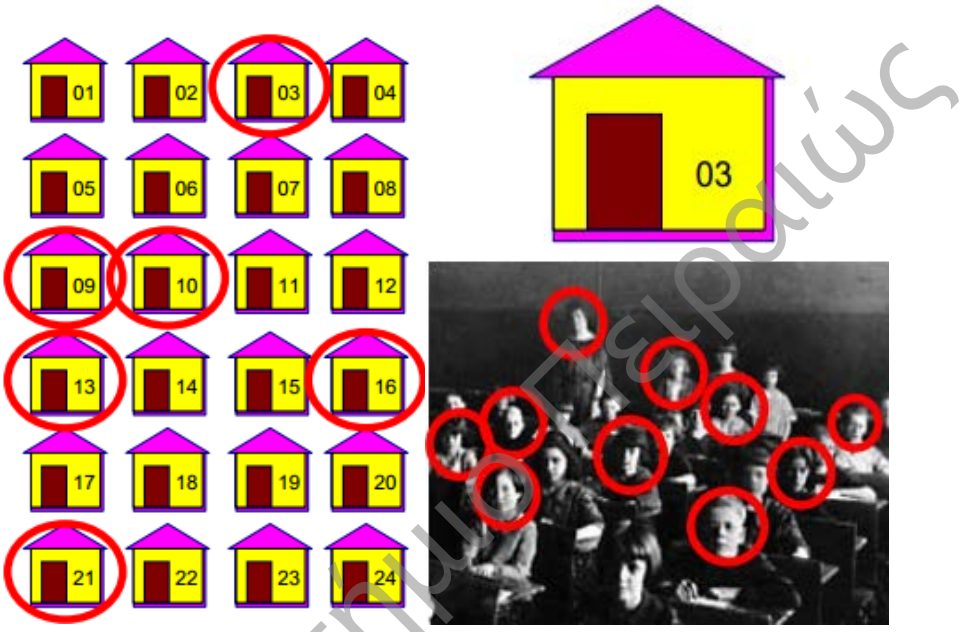
- Η τεχνική αυτή είναι εύκολη, φθηνή και γρήγορη.
- Δεν είναι απαραίτητος ο πλήρης κατάλογος μονάδων του πληθυσμού της δειγματοληψίας

Πανεπιστήμιο Πειραιώς

- Μεγαλύτερη ευκολία οργάνωσης και μικρότερο κόστος λόγω της μικρότερης διασποράς των μονάδων

Μειονεκτήματα

- Υπάρχει μεγαλύτερο τυπικό σφάλμα και ευρύτερα όρια αξιοπιστίας σε σύγκριση με την απλή τυχαία δειγματοληψία (για ίδιο μέγεθος δείγματος).
- Ο υπολογισμός του τυπικού σφάλματος και των ορίων αξιοπιστίας είναι πολύπλοκος.



Εικόνα 5 - Παράδειγμα δειγματοληψίας κατά συστάδες

Πολυσταδιακή Δειγματοληψία (Multistage Random Sampling)

Η πολυσταδιακή δειγματοληψία είναι μια σύνθετη μορφή δειγματοληψίας η οποία είναι βασισμένη στην δειγματοληψία κατά συστάδες, τεχνική που έχουμε δει λίγο πιο πριν. Η δειγματοληψία κατά συστάδες είναι ένα είδος δειγματοληψίας η οποία περιλαμβάνει τη διαίρεση του πληθυσμού σε ομάδες (ή συστάδες - clusters). Στη συνέχεια, μία ή περισσότερες συστάδες επιλέγονται τυχαία και όλες οι μονάδες/άτομα που βρίσκονται εντός του επιλεγμένου συμπλέγματος εμπεριέχονται στο δείγμα [14].

Για παράδειγμα, σκεφτείτε την ιδέα της δειγματοληψίας των κατοίκων της Νέας Υόρκης για συνεντεύξεις πρόσωπο-με-πρόσωπο. Σαφώς, στο πρώτο στάδιο της διαδικασίας, θα θέλαμε να κάνουμε κάποιο είδος Δειγματοληψίας Συμπλέγματος (Cluster Sampling). Θα μπορούσαμε να δοκιμάσουμε δήμους ή εκτάσεις απογραφής σε όλη την πολιτεία. Όμως, στην Δειγματοληψία Συμπλέγματος θα προχωρούσαμε στην μέτρηση του καθενός στις συστάδες που επιλέξαμε. Ακόμα και αν δοκιμάζουμε εκτάσεις απογραφής μπορεί να μην είμαστε σε θέση να μετρήσουμε τον καθένα που είναι στην έκταση απογραφής. Έτσι, θα μπορούσαμε να δημιουργήσουμε μια διαστρωματωμένη διαδικασία δειγματοληψίας εντός των συστάδων.

Πανεπιστήμιο Πειραιώς

Δειγματοληψία χωρίς πιθανότητα

Δειγματοληψία ευκολίας ή συμπτωματική δειγματοληψία

Πολλές φορές μέρος του πληθυσμού είτε δεν είναι διαθέσιμο είτε δεν είναι διατεθειμένο να συμμετάσχει σε έρευνες. Εκτός αυτού, η διαδικασία επιλογής ενός τυχαίου δείγματος από ένα μεγάλο πληθυσμό είναι χρονοβόρα και δαπανηρή, με αποτέλεσμα οι ερευνητές να αναγκάζονται να χρησιμοποιούν μόνο τα διαθέσιμα δείγματα. Λόγω αυτού τα δείγματα δεν είναι πάντα αντικειμενικά και αντιπροσωπευτικά και συχνά μπορεί να θεωρηθούν μεροληπτικά.

Η δειγματοληψία ευκολίας είναι ένας τύπος δειγματοληψίας χωρίς πιθανότητα που περιλαμβάνει το δείγμα που προέρχεται από το τμήμα εκείνο του πληθυσμού το οποίο έχει επιλεγεί επειδή είναι εύκολα διαθέσιμο και βολικό. Το συγκεκριμένο τμήμα πληθυσμού είναι εύκολα προσβάσιμο από ους ερευνητές καθώς επιλέγεται με βάση τις σχέσεις ή τα δίκτυα στα οποία έχουν εύκολη πρόσβαση. Όπως είναι αναμενόμενο, ο ερευνητής χρησιμοποιώντας ένα τέτοιο δείγμα δεν μπορεί επιστημονικά να κάνει γενικεύσεις για το σύνολο του πληθυσμού από το δείγμα αυτό, διότι δεν θα ήταν αρκετά αντιπροσωπευτικό.

Ένα παράδειγμα της συγκεκριμένης μεθόδου θα ήταν η διεξαγωγή μίας έρευνας σε ένα εμπορικό κέντρο κάποια συγκεκριμένη ώρα της ημέρας, είτε πρωί είτε απόγευμα είτε βράδυ αλλά όχι σε διαφορετικές ώρες της ημέρας και όχι αρκετές φορές την εβδομάδα. Σε αυτή την περίπτωση, οι άνθρωποι από τους οποίους θα μπορεί ο ερευνητής να πάρει συνέντευξη θα περιορίζονται σε αυτούς που επισκέπτονται το εμπορικό κέντρο εκείνη τη δεδομένη στιγμή. Με αυτό τον τρόπο όμως το δείγμα δεν θα αντιπροσωπεύει τις απόψεις των υπολοίπων μελών της κοινωνίας σε μια τέτοια περιοχή, που θα έπρεπε επίσης να ληφθούν υπόψιν. Έτσι, το συγκεκριμένο είδος δειγματοληψίας είναι πιο χρήσιμο για πιλοτικές δοκιμές [7].

Η αξιοπιστία των αποτελεσμάτων ενός ερευνητή με δειγματοληψία θα εξαρτηθεί από το αν ο αναγνώστης πεισθεί ότι η επιλογή του δείγματος ισοδυναμεί σε μεγάλο βαθμό με τον πληθυσμό από τον οποίο έχουν καθοριστεί.

Δειγματοληψία αναλογίας ή ποσοστιαία δειγματοληψία

Ο κύριος στόχος της δειγματοληψίας αναλογίας είναι να επικεντρωθεί σε συγκεκριμένα χαρακτηριστικά του πληθυσμού που παρουσιάζουν ενδιαφέρον, με σκοπό να μπορέσει ο ερευνητής να ανταποκριθεί καλύτερα και να απαντήσει σε ερωτήσεις της έρευνας. Και πάλι το δείγμα το οποίο μελετάμε δεν είναι αντιπροσωπευτικό του πληθυσμού, αλλά και για τους ερευνητές που επιδιώκουν την ποιοτική ή σχέδια μικτών ερευνητικές μεθόδους, αυτό δεν θεωρείται ως αδυναμία.

Στην σκόπιμη δειγματοληψία, επιλέγουμε το δείγμα με κάποιο σκοπό για αυτό και συνήθως έχουμε μία ή περισσότερες συγκεκριμένες προκαθορισμένες ομάδες στις οποίες στοχεύουμε.

Ένα παράδειγμα για το συγκεκριμένο είδος δειγματοληψίας θα μπορούσε να ήταν μία έρευνα η οποία στοχεύει σε γυναίκες ηλικίας μεταξύ 20-30 ετών. Έτσι, ο εν λόγω ερευνητής σταματάει τα άτομα που φαίνονται να ανήκουν στην κατηγορία την οποία θέλει να εξετάσει και ρωτάει αν θέλουν να συμμετάσχουν. Ένα από τα πρώτα πράγματα που είναι πιθανό να κάνει ο ερευνητής είναι να βεβαιωθεί ότι ο εναγόμενος πράγματι πληροί τα κριτήρια για να συμπεριφερθεί στο δείγμα. Η σκόπιμη δειγματοληψία μπορεί να είναι πολύ χρήσιμη σε περιπτώσεις όπου θα πρέπει να καταλήξει σε συγκεκριμένο δείγμα και σε μικρό χρονικό διάστημα. Επίσης, μπορεί να εφαρμοστεί εκεί όπου η δειγματοληψία βάση αναλογικότητας δεν είναι το κύριο μέλημα. Με μια ποσοστιαία δειγματοληψία είναι πιθανό να παρθούν οι απόψεις του πληθυσμού που

Πανεπιστήμιο Πειραιώς

μας ενδιαφέρει, αλλά παράλληλα είναι πιθανό να προκύψουν περισσότερες μετρήσεις από συγκεκριμένες υποομάδες του πληθυσμού λόγω του ότι τα άτομα αυτά είναι πιο εύκολα προσβάσιμα.

2.2. Χρονοσειρές

Εισαγωγή στις Χρονοσειρές

Μια χρονοσειρά (Timeseries) είναι μια ακολουθία από σημεία δεδομένων, η οποία μετράται συνήθως σε διαδοχικές χρονικές στιγμές που απέχουν κατά ομοιόμορφα διαστήματα χρόνου μεταξύ τους. Οι παρατηρήσεις που αποτελούν μία χρονοσειρά παίρνονται σε ορισμένες χρονικές στιγμές ή περιόδους που ισαπέχουν μεταξύ τους.

Οι διάφορες n χρονικές στιγμές, οι οποίες μπορεί να αναφέρονται σε έτη, μήνες, μέρες κ.λπ., συμβολίζονται με τον όρο «Xi» ενώ οι διάφορες n τιμές των αντίστοιχων παρατηρήσεων με τον όρο «Yi». Με αυτό τον τρόπο δημιουργούνται n ζεύγη της μορφής M(Xi,Yi) τα οποία μπορούν να απεικονιστούν στο σύστημα αξόνων με τη μορφή μίας γραφικής παράστασης. Με την ένωση όλων των n σημείων προκύπτει το χρονοδιάγραμμα, η μελέτη του οποίου μας δίνει μια γενική εικόνα της διαχρονικής εξέλιξης του φαινομένου ή του χαρακτηριστικού το οποίο είναι υπό έρευνα και για το οποίο πάρθηκαν οι συγκεκριμένες μετρήσεις. Η ανάλυση των χρονοσειρών περιλαμβάνει μεθόδους για την ανάλυση των δεδομένων χρονοσειρών προκειμένου να εξαχθούν χρήσιμα στατιστικά στοιχεία και άλλα χαρακτηριστικά των δεδομένων. Πιο συγκεκριμένα, η ανάλυση αυτή χρησιμοποιείται για να καθορίσουμε μοντέλα τα οποία μετατρέπουν τις διάφορες πληροφορίες από κανονικά χρονικά διαστήματα σε στατιστικά μέτρα.

Η πρόβλεψη των χρονοσειρών είναι η χρήση ενός μοντέλου για την πρόβλεψη μελλοντικών τιμών που βασίζονται σε τιμές που έχουν ήδη παρατηρηθεί και καταγραφεί.

Οι χρονοσειρές είναι ένας πολύ χρήσιμος και παραγωγικός τύπος δεδομένων εφόσον μας βοηθάνε να προβλέψουμε μελλοντικές τιμές, να έχουμε μια καλύτερη κατανόηση του μηχανισμού δημιουργίας των δεδομένων ενώ παράλληλα παρέχουν τον βέλτιστο έλεγχο του συστήματος.

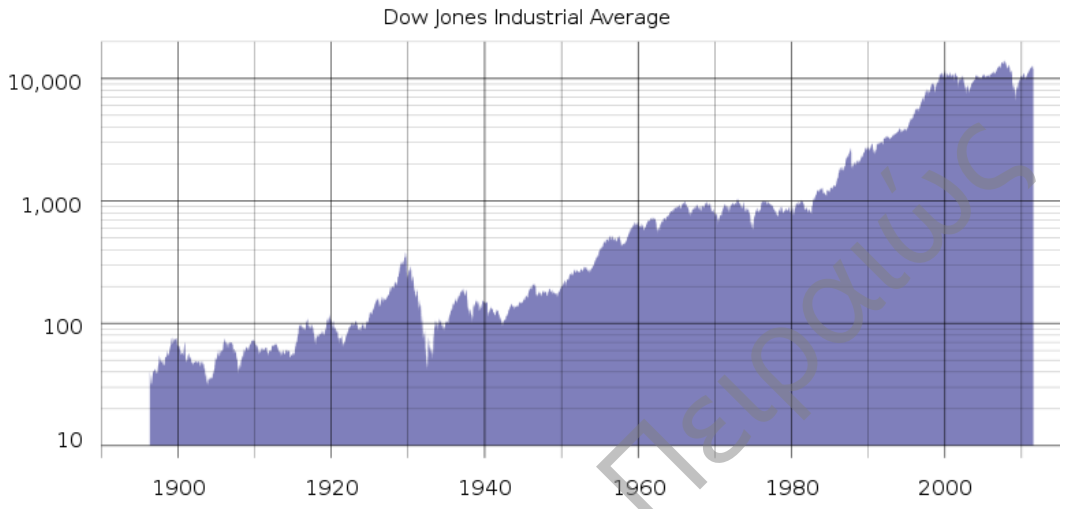
Η πιο χαρακτηριστική ιδιότητα μίας χρονοσειράς είναι ότι τα δεδομένα δε δημιουργούνται ανεξάρτητα και η διασπορά τους ποικίλει στο χρόνο. Για το λόγο αυτό οι στατιστικές διαδικασίες που υποθέτουν ανεξάρτητα και ταυτόσημα κατανεμημένα δεδομένα αποκλείονται από την ανάλυση των χρονοσειρών.

Όπως έχει προαναφερθεί, τα δεδομένα χρονολογικών σειρών έχουν μια φυσική χρονική διάταξη, γεγονός που καθιστά την ανάλυση τους διαφορετική από τα υπόλοιπα κοινά προβλήματα ανάλυσης δεδομένων στα οποία δεν υπάρχει φυσική διάταξη των παρατηρήσεων. Επιπρόσθετα, η ανάλυση δεν περιέχει γεωγραφικά και χωρικά δεδομένα, όπως συμβαίνει στη χωρική ανάλυση των δεδομένων. Γενικότερα, ένα στοχαστικό μοντέλο για μια χρονοσειρά αντικατοπτρίζει το γεγονός ότι οι παρατηρήσεις οι οποίες έγιναν σε κοντινά χρονικά διαστήματα είναι πιο στενά συνδεδεμένες από ότι οι παρατηρήσεις που έγιναν σε πιο μακρινές χρονικές στιγμές.

Η ανάλυση των χρονοσειρών μπορεί να εφαρμοστεί σε πραγματικές τιμές, συνεχόμενα δεδομένα, διακριτά αριθμητικά δεδομένα ή διακριτά συμβολικά στοιχεία (δηλαδή ακολουθίες χαρακτήρων, όπως γράμματα και λέξεις στην αγγλική γλώσσα). Κάποια πολύ γνωστά και χαρακτηριστικά παραδείγματα των χρονοσειρών

Πανεπιστήμιο Πειραιώς

είναι η ημερήσια τιμή κλεισίματος του δείκτη Dow Jones Industrial Average (Εικόνα 1) και ο ετήσιος όγκος της ροής του ποταμού Νείλου στο Ασουάν. Οι χρονοσειρές πολύ συχνά απεικονίζονται με γραφήματα γραμμών και χρησιμοποιούνται στις στατιστικές, στην επεξεργασία σήματος, στην αναγνώριση προτύπων, στην Οικονομετρία, στα οικονομικά μαθηματικά, στην πρόγνωση του καιρού, στην πρόγνωση σεισμών, στο ηλεκτροεγκεφαλογράφημα, στον μηχανικό ελέγχου, στην αστρονομία και στη μηχανική επικοινωνιών.



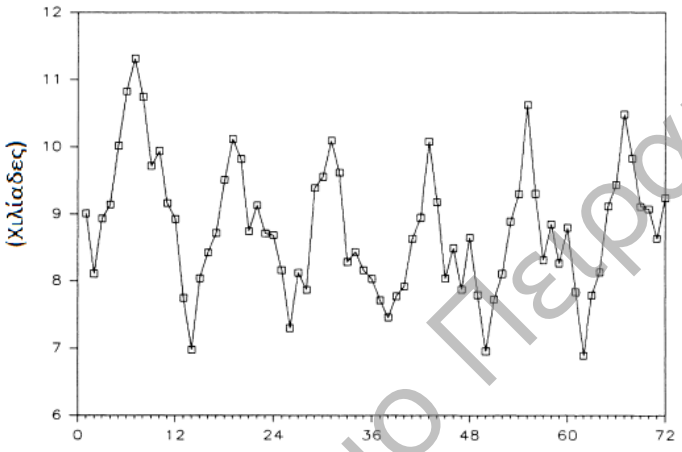
Εικόνα 6 – Ιστορική λογαριθμική γραφική παράσταση του DJIA από το 1896 έως τον Ιούλιο 2011

Μια χρονοσειρά είναι μία σειρά από ένα σύνολο παρατηρήσεων, όπου η κάθε μία έχει καταγραφεί σε μία συγκεκριμένη στιγμή t. Μία διακριτή χρονοσειρά (Εικόνα 3) είναι αυτή στην οποία το σύνολο των φορών «Τ0» κατά το οποίο λήφθηκαν οι παρατηρήσεις, είναι ένα διακριτό σύνολο όπως για παράδειγμα όταν οι παρατηρήσεις γίνονται ανά σταθερά χρονικά διαστήματα. Από την άλλη, μία συνεχής χρονοσειρά (Εικόνα 2) είναι αυτή κατά την οποία οι παρατηρήσεις που λαμβάνονται καταγράφονται συνεχώς βασισμένες σε κάποιο χρονικό διάστημα, όπως για παράδειγμα Τ0 = [0,1]. Στη συνέχεια ακολουθούνε δύο αντιπροσωπευτικά παραδείγματα για τις διακριτές και τις συνεχής χρονοσειρές. Συγκεκριμένα, το σενάριο για το οποίο καταγράφηκαν οι διακριτές τιμές είναι οι το πλήθος των θανάτων από ατυχήματα κάθε μήνα στην Αμερική μεταξύ των χρονολογιών 1973 και 1978. Το σενάριο για το οποίο καταγράφηκαν οι συνεχής τιμές είναι οι νίκες των παιχνιδιών που σημειώθηκαν από την Εθνική ομάδα και από την Αμερικανική ομάδα χειροσφαίρισης μεταξύ των χρονολογιών 1933 και 1980.

Παράδειγμα Διακριτής Χρονοσειράς: Θάνατοι από ατύχημα κάθε μήνα στην Αμερική 1973-1978

Πανεπιστήμιο Πειραιώς

	1973	1974	1975	1976	1977	1978
Jan.	9007	7750	8162	7717	7792	7836
Feb.	8106	6981	7306	7461	6957	6892
Mar.	8928	8038	8124	7776	7726	7791
Apr.	9137	8422	7870	7925	8106	8129
May	10017	8714	9387	8634	8890	9115
Jun.	10826	9512	9556	8945	9299	9434
Jul.	11317	10120	10093	10078	10625	10484
Aug.	10744	9823	9620	9179	9302	9827
Sep.	9713	8743	8285	8037	8314	9110
Oct.	9938	9129	8433	8488	8850	9070
Nov.	9161	8710	8160	7874	8265	8633
Dec.	8927	8680	8034	8647	8796	9240



Εικόνα 7 – Διακριτή Χρονοσειρά

Το πρώτο βήμα για την σωστή ανάλυση μιας χρονοσειράς είναι η επιλογή του κατάλληλου μαθηματικού μοντέλου (ή ομάδας μοντέλων) για τα δεδομένα. Για να καταστεί δυνατή η ενδεχομένως απρόβλεπτη φύση των μελλοντικών παρατηρήσεων είναι φυσικό να υποθέσουμε ότι κάθε παρατήρηση xi αποτελεί συνειδητοποιημένη αξία μιας τυχαίας μεταβλητή Xi.

Οι μέθοδοι για αναλύσεις χρονοσειρών μπορούν να χωριστούν σε δύο κατηγορίες: τις μεθόδους που βασίζονται στο πεδίο της συχνότητας και τις μεθόδους που βασίζονται στο πεδίο του χρόνου. Η πρώτη κατηγορία περιλαμβάνει φασματική ανάλυση ενώ η δεύτερη κατηγορία περιλαμβάνει την αυτόματη συσχέτιση και την ανάλυση πολλαπλής συσχέτισης. Στη περίπτωση της συσχέτισης του πεδίου χρόνου η ανάλυση μπορεί να γίνει με κάποιο τρόπο φλιταρίσματος, χρησιμοποιώντας κλίμακα συσχέτισης, μετριάζοντας έτσι την ανάγκη λειτουργίας σε πεδίο συχνότητας.

Επιπλέον, οι τεχνικές ανάλυσης χρονοσειρών μπορεί να διαιρεθούν σε παραμετρικές και μη παραμετρικές μεθόδους. Οι παραμετρικές προσεγγίσεις υποθέτουν ότι η υποκείμενη στάσιμη στοχαστική διαδικασία έχει μια ορισμένη δομή η οποία μπορεί να περιγραφεί χρησιμοποιώντας ένα μικρό αριθμό παραμέτρων (για παράδειγμα, χρησιμοποιώντας ένα αυτοπαλινδρομικό ή ένα κινούμενο μέσο μοντέλο). Σε αυτές τις προσεγγίσεις, στόχος είναι η εκτίμηση των παραμέτρων του μοντέλου το οποίο περιγράφει τη στοχαστική διαδικασία. Αντίθετα, στις μη-παραμετρικές προσεγγίσεις εκτιμάται ρητά η συνδιακύμανση ή το φάσμα της διαδικασίας, χωρίς να γίνει υπόθεση ότι η διαδικασία έχει κάποια συγκεκριμένη δομή.

Πανεπιστήμιο Πειραιώς

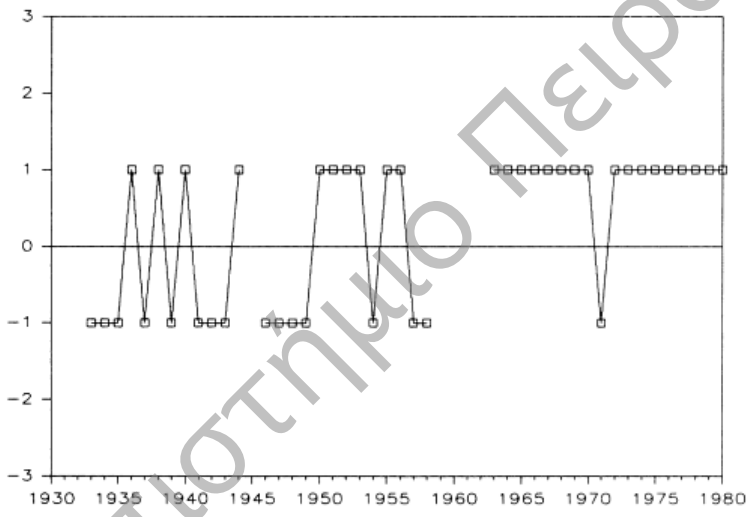
Τέλος, οι μέθοδοι ανάλυσης χρονοσειρών μπορούν ακόμη να διαιρεθούν σε γραμμικές και μη γραμμικές όπως επίσης και σε μονοπαραγοντικές και πολυπαραγοντικές.

Παράδειγμα Συνεχής χρονοσειράς: Παιχνίδια χειροσφαίρισης 1933-1980

$$x_t = \begin{cases} 1 & \text{εάν η Εθνική ομάδα νίκησε τη χρονιά t} \\ -1 & \text{εάν η Αμερικάνικη ομάδα νίκησε τη χρονιά t} \end{cases}$$

$t-1900$	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48
x_t	−1	−1	−1	1	−1	1	−1	1	−1	−1	−1	1	†	−1	−1	−1
$t-1900$	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64
x_t	−1	1	1	1	1	−1	1	1	−1	−1	*	*	*	*	1	1
$t-1900$	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80
x_t	1	1	1	1	1	1	−1	1	1	1	1	1	1	1	1	1

† = κανένα παιχνίδι
* = δύο προγραμματισμενα παιχνίδια



Εικόνα 8 – Συνεχής χρονοσειρά

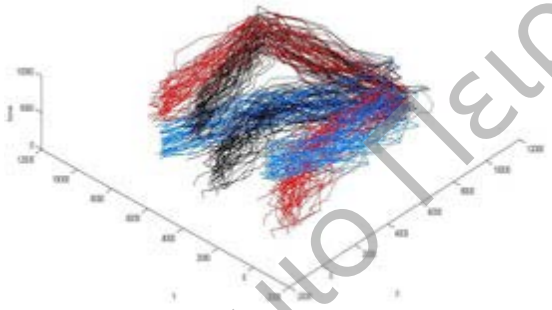
Πανεπιστήμιο Πειραιώς

2.3. Τροχιές

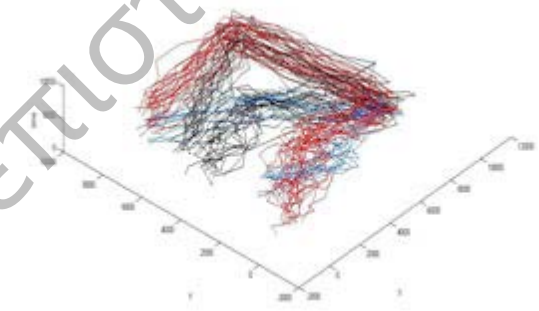
Εισαγωγή στις τροχιές

Οι τροχιές αποτελούν τη διαδρομή που ακολουθεί ένα κινούμενο αντικείμενο μέσα στο χώρο ως μία συνάρτηση του χρόνου. Το αντικείμενο θα μπορούσε να είναι ένα πλοίο ή ένα σύνολο πλοίων για τα οποία καταγράφονται οι διαδρομές που κάνουν καθημερινά μέσα στη θάλασσα. Αξίζει να σημειωθεί ότι όταν μιλάμε για τροχιές μιλάμε για «αντικείμενα» που η κίνηση τους μπορεί να εντοπιστεί μέσω GPS ή έχουν πάνω τους ένα μηχανισμό GPS μέσω του οποίου καταγράφεται η κίνηση τους.

Σήμερα, η καταγραφή της πορείας ενός κινούμενου αντικειμένου έχει γίνει πολύ σημαντική και αποτελεί θέμα για πολλές μελέτες. Συγκεκριμένα υπάρχουν τεράστιες βάσεις δεδομένων οι οποίες έχουν καταχωρημένες τιμές για τα σημεία πολλών αντικειμένων, η ακολουθία των οποίων αποτελεί την τροχιά του αντικειμένου. Οι συγκεκριμένες βάσεις δεδομένων ονομάζονται «Βάσεις Δεδομένων Τροχιάς» (TD-Trajectory Databases).



Εικόνα 9 - Βαση Δεδομένων Τροχιάς



Εικόνα 10 - Υποσύνολο της αρχικής Βάσης Δεδομένων Τροχιάς

Τα τελευταία χρόνια έχει παρουσιαστεί ένα έντονο ενδιαφέρον για την εξόρυξη των δεδομένων όσο αφορά τις βάσεις δεδομένων που περιέχουν δεδομένα τροχιών. Η δειγματοληψία των τροχιών είναι ένα δύσκολο έργο με μεγάλες δυνατότητες εφαρμογών σε αυτόν τον τομέα.

Για τη δειγματοληψία της τροχιάς έχουν προταθεί και χρησιμοποιηθεί ποικίλοι τρόποι. Ένας ενδιαφέρον τρόπος που έχει προταθεί πρόσφατα, και παρουσιάζεται σε μελέτη που θα αναλυθεί πιο μετά (στο κεφάλαιο 3), χωρίζεται σε τρία στάδια. Αρχικά, υιοθετείται μία συμβολική αναπαράσταση των τροχιών – η οποία και επιτρέπει να μοντελοποιηθούν όλες οι τροχιές που υπάρχουν καταχωρημένες στη βάση

Πανεπιστήμιο Πειραιώς

δεδομένων με έναν τρόπο παρόμοιο με τα διανύσματα. Πέρα από ατή την αναπαράσταση προτείνεται και μία μέθοδος αναπαράστασης κάθε τροχιάς μέσω μίας συνεχόμενης συνάρτησης που περιγράφει την αντιπροσωπευτικότητα του κάθε συστατικού της με σεβασμό προς ολόκληρη τη βάση δεδομένων. Στο τελευταίο μέρος, εισάγετε μία αυτόματη μέθοδος για τη δειγματοληψία της τροχιάς – η οποία ονομάζεται T-Sampling- και είναι βασισμένη στην αντιπροσωπευτικότητα της τροχιάς.

2.4. Web-clicks (κλικ σε ιστοσελίδες)

Στις μέρες μας, όπου το διαδίκτυο και η χρήση του είναι μια από τις πιο συχνές και σημαντικές ασχολίες του ανθρώπου, έχουν προκύψει διάφορες καινοτόμες τεχνικές διαχείρισης των κινήσεων των χρηστών. Ένα από αυτά είναι η καταγραφή του πλήθους των κλικ που γίνονται σε κάποια συγκεκριμένη ιστοσελίδα από τους διάφορους χρήστες. Η πιο πάνω διαδικασία αποτελεί άλλη μία μέθοδο δειγματοληψίας με απώτερο σκοπό την διεξαγωγή μελέτης και την εξαγωγή συμπερασμάτων.

Η ανάλυση των κλικ (click analytics) είναι ένας ειδικός τύπος web analytics και χρησιμοποιείται για τη μέτρηση, συλλογή, ανάλυση και καταγραφή των δεδομένων στον Παγκόσμιο Ιστό για τους σκοπούς της κατανόησης και βελτιστοποίηση της χρήσης web. Ο ειδικός αυτός τύπος δίνει έμφαση στα κλικς και οι εκδότες μιας ιστοσελίδας χρησιμοποιούν click analytics για τον προσδιορισμό της απόδοσης της συγκεκριμένης ιστοσελίδας, σε σχέση με το που κάνουν κλικ οι χρήστες του site [16].

Η ανάλυση των κλικ σε μία ιστοσελίδα δεν είναι απλώς ένα εργαλείο για τη μέτρηση της κυκλοφορίας του Ιστού, αλλά μπορεί να χρησιμοποιηθεί και ως εργαλείο για τις επιχειρήσεις και την έρευνα της αγοράς, για να αξιολογήσουν και να βελτιώσουν την αποτελεσματικότητα μιας ιστοσελίδας. Ο κύριος στόχος της μέτρησης των κλικ σε μία ιστοσελίδα είναι να γίνεται γνωστό στους ιδιοκτήτες κάθε ιστοσελίδας πόσοι χρήστες επισκέπτονται κάθε μέρα τον ιστιότοπο τους. Με αυτό τον τρόπο θα γνωρίζουν εάν χρειάζονται βελτιώσεις ή/και διαφήμιση έτσι ώστε να γίνει πιο γνωστή η σελίδα τους ή εάν κάποια αλλαγή που έκαναν τους επέφερε πιο πολλούς επισκέπτες ανά μέρα.

Επίσης, η διαδικασία ανάλυσης των κλικ μπορεί να γίνει είτε σε πραγματικό χρόνο είτε σε «εξωπραγματικό» χρόνο, ανάλογα με το είδος των πληροφοριών που ζητούνται. Ένα παράδειγμα είναι οι συντάκτες των πρωτοσέλιδων σε ιστοσελίδες μέσων ενημέρωσης υψηλής κυκλοφορίας που συνήθως θέλουν να παρακολουθούν τις σελίδες τους σε πραγματικό χρόνο, για τη βελτιστοποίηση του περιεχομένου. Από την άλλη, σχεδιαστές ή άλλου τύπου ενδιαφερόμενοι μπορεί να αναλύσουν τα κλικς σε ένα ευρύτερο χρονικό πλαίσιο για να τους βοηθήσουν στην αξιολόγηση των επιδόσεων όσο αφορά τα στοιχεία σχεδίασης, διαφημίσης κλπ.

Όπως έχει αναφερθεί πιο πάνω, τα δεδομένα σχετικά με τα κλικ μπορούν να συγκεντρωθεούν με τουλάχιστον δύο τρόπους. Στην ιδανική περίπτωση, ένα κλικ είναι "καταγράφεται" όταν συμβαίνει (είναι δηλαδή συνδεδμένα κλικ και ώρα), και αυτή η μέθοδος απαιτεί κάποια λειτουργικότητα που παραλαμβάνει τις σχετικές πληροφορίες, ακριβώς όταν λαμβάνει χώρα το συμβάν. Εναλλακτικά, μπορεί κάποιος να υποθέσει ότι μια προβολή σελίδας είναι αποτέλεσμα ενός κλικ, και ως εκ τούτου, καταγράφεται ένα προσομοιωμένο κλικ το οποίο οδήγησε σε αυτήν την προβολή της σελίδας.

Για την παρακολούθηση των επισκεπτών και των πωλήσεων με τη βοήθεια κάποιων ήδη έτοιμων υπηρεσιών, όπως για παράδειγμα το Google Analytics που προσφέρεται από το Google και δημιουργεί λεπτομερή στατιστικά στοιχεία σχετικά με την κυκλοφορία ενός δικτυακού τόπου.

Πανεπιστήμιο Πειραιώς

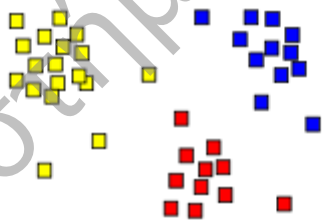
Για την καταγραφή των προσβάσεων σε μία ιστοσελίδα (κλικς χρηστών) υπάρχουν και πάλι ήδη υπάρχουσες εφαρμογές οι οποίες δημιουργούν και παρέχουν διαγράμματα με το πλήθος των κλικ βασισμένα σε διάφορες κατηγορίες.

2.5. Ομαδοποίηση (Clustering)

Η μέθοδος της ομαδοποίησης είναι μία τεχνική για την εξόρυξη δεδομένων που χρησιμοποιείται για να τοποθετήσει τα στοιχεία δεδομένων σε σχετικές ομάδες, χωρίς όμως να υπάρχει γνώση εκ των προτέρων για τους ορισμούς της ομάδας. Αποτελεί την ανάλυση των συστάδων (clusters) και την ομαδοποίηση ενός συνόλου αντικειμένων με τέτοιο τρόπο ώστε τα αντικείμενα στην ίδια ομάδα – που ονομάζεται συστάδα – να είναι πιο παρόμοια μεταξύ τους παρά με άλλα αντικείμενα σε άλλες συστάδες. Είναι μία από τις πιο σημαντικές εργασίες σχετικά με την εξόρυξη δεδομένων και μια κοινή τεχνική για την ανάλυση των στατιστικών στοιχείων, που χρησιμοποιούνται σε πολλούς τομείς όπως είναι η αναγνώριση προτύπων και η ανάκτηση πληροφοριών. Η ανάλυση της ομαδοποίησης μπορεί να επιτευχθεί με διάφορους αλγορίθμους [14] .

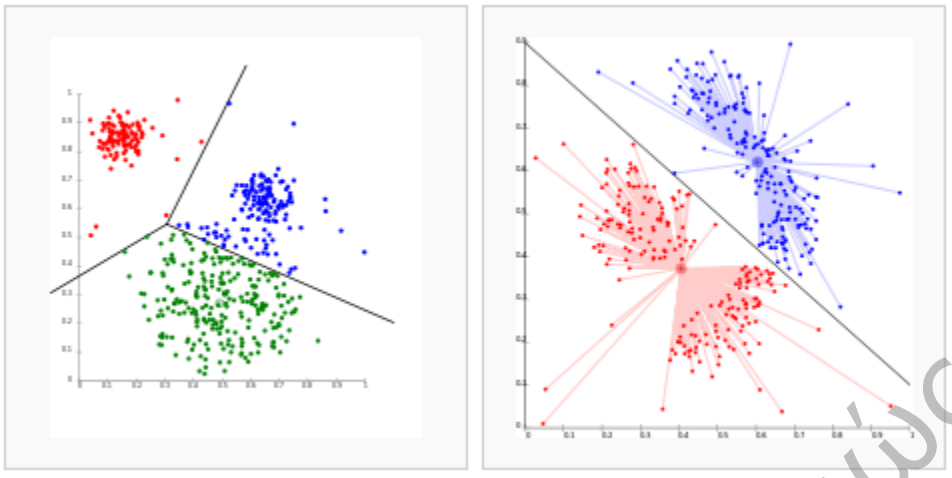
Η μέθοδος αυτή μπορεί να χρησιμοποιηθεί για την αξιολόγηση τεχνικών δειγματοληψίας. Συγκεκριμένα, εφόσον υπάρχουν τα δεδομένα (π.χ. είναι καταχωρημένα σε κάποια βάση δεδομένων) τότε εφαρμόζεται η μέθοδος Clustering σε αυτά τα αρχικά δεδομένα, και έπειτα εφαρμόζεται ξανά η ίδια εξόρυξη (mining) στο δείγμα, έτσι ώστε να βρεθούν στο τέλος όσο το δυνατόν ίδια πρότυπα (patterns).

Κάποιες από τις πιο δημοφιλείς τεχνικές ομαδοποίησης είναι η «k-means clustering» και η «expectation maximization» (EM clustering).



Εικόνα 11 - Ανάλυση ομαδοποίησης

Πανεπιστήμιο Πειραιώς



Εικόνα 12 - Παράδειγμα τεχνικής "k-means clustering"

Για τους σκοπούς του Data Mining στην εφαρμογή Δεδομένων (κεφάλαιο 4) θα χρησιμοποιήσουμε την τεχνική “TwoStep Cluster Analysis” η οποία αποτελεί ανάλυση των συστάδων σε δύο βήματα [20]. Συγκεκριμένα, η εν λόγω διαδικασία είναι διερευνητική και είναι σχεδιασμένη για να αναδείξει τις φυσικές ομάδες (clusters) μέσα σε από σύνολο δεδομένων που διαφορετικά δεν θα ήταν εμφανής. Ο αλγόριθμος που χρησιμοποιείται από τη διαδικασία αυτή έχει πολλά επιθυμητά χαρακτηριστικά που το διαφοροποιούν από τις παραδοσιακές τεχνικές ομαδοποίησης:

- Χειρισμός των κατηγορηματικών και συνεχόμενων μεταβλητών.
- Αυτόματη επιλογή του αριθμού των clusters.

Πανεπιστήμιο Πειραιώς

ΕΝΟΤΗΤΑ 3

Σχετικές Μελέτες

3.1. Δειγματοληψία σε τροχιές κινούμενων αντικειμένων

Στην ενότητα αυτή θα δούμε και θα αναλύσουμε σχετικές μελέτες που έγιναν βασισμένες στη μέθοδο δειγματοληψίας των τροχιών. Μέσω αυτών των μελετών θα δούμε πως η δειγματοληψία οδηγεί στην ανακάλυψη προτύπων συμπεριφοράς κινούμενων αντικειμένων και πως η διαδικασία αυτή μπορεί να εκμεταλλευτεί σε διάφορα πεδία.

Τα τελευταία χρόνια έχουν γίνει αρκετές μελέτες βασισμένες στις μεθόδους δειγματοληψίας, με απώτερο σκοπό την εξαγωγή γενικευμένων αποτελεσμάτων μέσα από τις μετρήσεις που καταγράφονται από το δείγμα (είτε αυτό είναι άτομα είτε είναι αντικείμενα). Η εφαρμογή των τεχνικών δειγματοληψίας εφαρμόζονται επίσης σε σύνθετους τύπους δεδομένων, των οποίων το κοινό χαρακτηριστικό είναι ότι είναι χρονοσημασμένα, ακολουθιακά δεδομένα. Ένα αντιπροσωπευτικό παράδειγμα αυτού του τύπου δεδομένων θα ήταν η καταγραφή της πορείας κινούμενων αντικειμένων στο χωροχρόνο, η πλοήγηση χρηστών σε ιστοσελίδες του διαδικτύου, οι ακολουθίες αγοραπωλησιών και οι τιμές κλεισίματος των μετοχών ενός χρηματιστηρίου.

Στις μέρες μας η δειγματοληψία της τροχιάς αντικειμένων είναι πολύ σημαντική εφόσον είναι μέρος εργασιών που εκτελούνται καθημερινά, όπως το GPS, τα GSM δίκτυα αλλά και οι διάφορες τεχνικές εντοπισμού μέσω των ηλεκτρονικών υπολογιστών.

Δειγματοληψία τροχιών χωρίς επίβλεψη

Η ανάλυση των δεδομένων και η γνώση για την εξερεύνηση και την ανακάλυψη σε βάσεις δεδομένων τροχιάς οδηγεί στην ανακάλυψη προτύπων συμπεριφοράς κινούμενων αντικειμένων, διαδικασία που μπορεί να εκμεταλλευτεί σε διάφορα πεδία. Σε πολλές σχετικές εργασίες έχουν προταθεί κατά καιρούς διάφορες προσεγγίσεις που προσπαθούν να αναλύσουν τα δεδομένα τροχιών είτε σε διερευνητικό επίπεδο είτε μέσω ανακάλυψης/εξόρυξης μοτίβων όπως είναι μία συστάδα ή ένα σύμπλεγμα από κινούμενα αντικείμενα [12].

Όλες οι πιο πάνω προσεγγίσεις λειτουργούν συνήθως σε μεγάλες βάσεις δεδομένων τροχιών και το ερώτημα που τίθεται είναι εάν είναι δυνατόν να εξαχθούν τα ίδια μοτίβα χρησιμοποιώντας πολύ μικρότερα αντιπροσωπευτικά υποσύνολα. Και πάνω σε αυτό τίθεται ακόμα ένα σημαντικό ερώτημα, εάν δηλαδή μπορεί να οριστεί ένα κατάλληλο υποσύνολο μέσα από μια πραγματική βάση δεδομένων τροχιών το οποίο θα μπορεί να συλλάβει τα ίδια μοτίβα.

Το πρόβλημα με τη δειγματοληψία τροχιών είναι μεγάλη πρόκληση λόγω της πολυπλοκότητας των κινούμενων αντικειμένων, εφόσον οι τεχνικές για τα σταθερά αντικείμενα δεν μπορούν να εφαρμοστούν.

Πανεπιστήμιο Πειραιώς

Για αυτό το λόγο έχει γίνει μία σχετική μελέτη και έχουν καταγραφεί τα συμπεράσματα έτσι ώστε να αντιμετωπισθεί το πρόβλημα της δειγματοληψίας τροχιάς.

Η προσέγγιση αυτή αποτελείται από τρία βασικά βήματα. Αρχικά, υιοθετείται μία συμβολική αναπαράσταση των τροχιών – η οποία και επιτρέπει να μοντελοποιηθούν όλες οι τροχιές που υπάρχουν καταχωρημένες στη βάση δεδομένων με έναν τρόπο παρόμοιο με τα διανύσματα. Η συμβολική αυτή αναπαράσταση των τροχιών διατηρεί το κινούμενο μοτίβο της κάθε τροχιάς, επιταχύνει τους υπολογισμούς και επιπλέον δεν υπάρχουν απώλειες όσο αφορά την κινητικότητα των μοτίβων.

Πέρα από ατή την αναπαράσταση προτείνεται και μία μέθοδος αναπαράστασης κάθε τροχιάς μέσω μίας συνεχόμενης συνάρτησης που περιγράφει την αντιπροσωπευτικότητα του κάθε συστατικού της με σεβασμό προς ολόκληρη τη βάση δεδομένων. Η συγκεκριμένη ιδέα εγκρίνει τη χρήση ενός αλγορίθμου συγχώνευσης ο οποίος εντοπίζει την μέγιστη χρονική περίοδο όπου το πρότυπο της κινητικότητας της κάθε τροχιάς διατηρείται.

Στο τελευταίο μέρος, εισάγεται μία αυτόματη μέθοδος για τη δειγματοληψία της τροχιάς – η οποία ονομάζεται T-Sampling- και είναι βασισμένη στην αντιπροσωπευτικότητα της τροχιάς. Με αυτή τη μέθοδο δεν λαμβάνεται υπόψη μόνο η μεγαλύτερη αλλά και η μικρότερη αντιπροσωπευτικότητα.

Στην συγκεκριμένη έρευνα παρέχεται και μία ενότητα που περιλαμβάνει την πειραματική μελέτη προκειμένου να αξιολογηθεί η προσέγγισή σε πραγματικές και συνθετικές βάσεις δεδομένων τροχιών (TD). Ειδικότερα, έχει χρησιμοποιηθεί το πραγματικό σύνολο δεδομένων των φορτηγών της Αθήνας, το οποίο αποτελείται από 112,300 θέσεις GPS που έχουν παρθεί από 50 φορτηγά που μεταφέρουν τσιμέντο στη μητροπολιτική περιοχή της Αθήνας, κατανεμημένες σε 1100 τροχιές.

Για περαιτέρω πειραματισμό, έχουν χρησιμοποιηθεί επίσης συνθετικά σύνολα δεδομένων που δημιουργούνται από μια προσαρμοσμένη γεννήτρια που βασίζεται στο δημοφιλές GSTD. Συγκεκριμένα, αυτή η γεννήτρια παράγει σύνολα δεδομένων τροχιάς με βάση μια δεδομένη κατανομή των χωροχρονικών σημείων εστίασης (focal points), για να επισκέπτονται από κάθε τροχιά σε μια συγκεκριμένη σειρά. Το παραγόμενο σύνολο δεδομένων τότε σχηματίζει ένα φυσικό σύμπλεγμα δεδομένου ότι όλες οι τροχιές ακολουθούν περίπου την ίδια συμπεριφορά. Για παράδειγμα, το σχήμα 8(α) απεικονίζει την δυσδιάστατη προβολή ενός συμπλέγματος που παράγεται με την παραπάνω γεννήτρια χρησιμοποιώντας τα σημεία 1 έως 5 ως εστιακά σημεία.

Σε αυτή την εργασία, έχει προταθεί μια νέα λύση στο δύσκολο πρόβλημα της δειγματοληψίας τροχιάς, όπου η πρόκληση είναι να κατασκευαστεί ένα δείγμα με την επιλογή εκπροσώπων ανάμεσα σε ένα μεγάλο σύνολο τροχιών, με ένα ανεξέλεγκτο τρόπο, για γενικό σκοπό. Εξ' όσων έχει μελετηθεί, δεν υπάρχει καμία σχετική εργασία που αντιμετωπίζει αυτό το πρόβλημα εκτός από διερευνητικές προσεγγίσεις, επιβλεπόμενες από το χρήστη.

Τμηματοποίηση και δειγματοληψία της τροχιάς ενός κινούμενου αντικειμένου

Η καταγραφή και παρακολούθηση της τροχιάς ενός αντικειμένου χρησιμοποιείται σε καθημερινή βάση, με την χρήση του GPS (Global Positioning System) ως το πιο χαρακτηριστικό παράδειγμα στις κινητές συσκευές. Η χρήση της τεχνολογίας GPS βοηθάει στον εντοπισμό της θέσης της κινητής συσκευής παρουσιάζοντας το σημείο πάνω στο χάρτη, μέσω της οθόνης της συγκεκριμένης συσκευής. Λόγω της αυξημένης ζήτησης της συγκεκριμένης λειτουργίας, σημειώθηκε και τεράστια αύξηση στη χρήση των βάσεων δεδομένων

Πανεπιστήμιο Πειραιώς

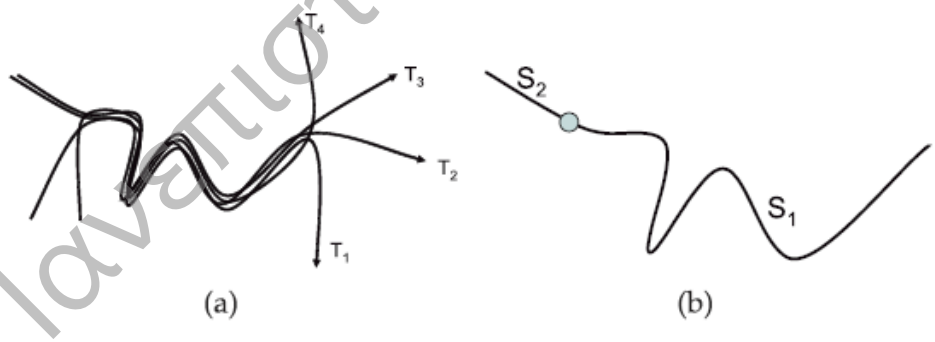
κινούμενων αντικειμένων (MOD). Έτσι, αυτή η έκρηξη των δεδομένων οδηγεί στο ενδιαφέρον για την εξόρυξη δεδομένων και στην ανακάλυψη της γνώσης μέσω της αντίληψης των κινούμενων δεδομένων.

Σε αυτό το σημείο θα αναφερθούμε σε μία πρόσφατη σχετική έρευνα που έχει γίνει με τη χρήση χρονοσημασμένων, ακολουθιακών δεδομένων. Συγκεκριμένα, το θέμα του άρθρου είναι «Segmentation and Sampling of Moving Object Trajectories Based on Representativeness» και αναφέρεται στην τμηματοποίηση και δειγματοληψία της τροχιάς ενός κινούμενου αντικειμένου βάσει της αντιπροσωπευτικότητας των τιμών που καταγράφονται.

Για να επιτευχθεί ο στόχος μελετάται ο συνδυασμός τριών διαφορετικών πτυχών: οι εναλλακτικές αναπαραστάσεις των διαδρομών των κινούμενων αντικειμένων (εκτός από τις συνήθεις αλληλουχίες των 3D ευθύγραμμων τμημάτων) σύμφωνα με τις συναφείς πληροφορίες που μπορεί να προέλθουν αυτόματα από το σύνολο του πληθυσμού τροχιών, η «αντιπροσωπευτική» ή μη «αντιπροσωπευτική» διορατικότητα μέσα από τον κατακερματισμό της τροχιάς και η δειγματοληψία τροχιών.

Στο συγκεκριμένο έγγραφο προτείνεται μία μέθοδος τμηματοποίησης και δειγματοληψίας της τροχιάς βασισμένη στην αντιπροσωπευτικότητα των υπό-τροχιών που βρίσκονται αποθηκευμένες στις βάσεις δεδομένων κινούμενων αντικειμένων (Moving Object Databases - MOD). Σκοπός της μελέτης αυτής είναι η εύρεση των πιο αντιπροσωπευτικών τροχιών του κινούμενου αντικειμένου έτσι ώστε να μπορέσει να γίνει κατανόηση, έρευνα, ανάλυση και περιήγηση στο χωροχρονικό περιεχόμενο. Σε αυτή την προσέγγιση, ο αναλυτής μπορεί να αιτηθεί τις k-κορυφαίες αντιπροσωπευτικές υπό-τροχιές.

Στο πιο κάτω σχήμα εικονογραφείται ένα παράδειγμα MOD το οποίο αποτελείται από τροχιές T1, T2, T3 και T4 (στο αριστερό μέρος) και από τις 2 πιο αντιπροσωπευτικές υπό-τροχιές S1 και S2 (στο δεξί μέρος), οι οποίες περιγράφουν καλύτερα το MOD.



Εικόνα 13 – Τροχιές κινούμενων αντικειμένων

Αρχικά, εκτελείται κάποιος καινοφανής αλγόριθμος παγκόσμιας ψήφους ο οποίος είναι βασισμένος στην τοπική πυκνότητα στις ομοιότητες των πληροφοριών μίας τροχιάς. Η μέθοδος αυτή εφαρμόζεται σε κάθε τμήμα της τροχιάς σχηματίζοντας έναν περιγραφέα τοπικής τροχιάς που αντιπροσωπεύει το τμήμα της ιδανικής γραμμής. Η αλληλουχία του παροχέα αυτού πάνω από την τροχιά δίνει το γενικότερη εικόνα της τροχιάς, όπου οι υψηλότερες τιμές αντιστοιχούν στα πιο αντιπροσωπευτικά κομμάτια. Στη συνέχεια, και αφότου έχει σχηματιστεί ο εν λόγω περιγραφέας, εφαρμόζεται ένας καινοφανής αλγόριθμος στο σήμα το οποίο αυτόματα υπολογίζει τον αριθμό των διαχωρισμάτων και των διαχωριστικών συνόρων,

Πανεπιστήμιο Πειραιώς

προσδιορίζοντας έτσι τα ομοιογενή διαχωριστικά όσο αφορά την αντιπροσωπευτικότητα. Στο τέλος, μία μέθοδος δειγματοληψίας, που εφαρμόζεται πάνω από τα τμήματα που προέκυψαν, αποδίδει τις πιο αντιπροσωπευτικές υπό-τροχιές στη βάση δεδομένων τροχιών (MOD).

Σε πρόσφατες σχετικές εργασίες οι οποίες είτε προσπαθούν να αναλύσουν τα δεδομένα των τροχιών είτε τα διάφορα αντιληπτά μοτίβα κίνησης χρησιμοποιείται η τεχνική της διχοτόμησης της τροχιάς έτσι ώστε να επιτευχθεί καλύτερη οργάνωση της βάσης δεδομένων ή ώστε να εξαχθούν κοινές τοπικές συμπεριφορές για ομαδοποίηση και ταξινόμηση. Συγκεκριμένα, για να επιτευχτεί ο στόχος έχουν εφαρμοστεί πολλοί διαφορετικοί τρόποι όπως είναι η χρήση των μικρότερων οριοθετημένων τριγώνων (MBRs – Minimum Bounding Rectangles) για την τμηματοποίηση της τροχιάς του αντικειμένου, χρήση συναρτήσεων απόστασης για την ομαδοποίηση των τροχιών κτλ. Όσο αφορά τώρα το κομμάτι της δειγματοληψίας, υπάρχουν ενδιαφέρουσες προσεγγίσεις όπως είναι η διερεύνηση, τεχνικές βασισμένες στην ομαδοποίηση όπως και τεχνικές κατά προσέγγιση. Όσα έχουμε προαναφέρει σε αυτό το σημείο συμπεριλαμβάνουν λύσεις για επιλογή μίας τροχιάς που δεν περιλαμβάνουν εντοπισμό αντιπροσωπευτικών υπό-τροχιών.

Δειγματοληψία τροχιάς για άμεση παρατήρηση της κυκλοφορίας

Στον τομέα των δικτύων επικοινωνίας, η μέτρηση της κυκλοφορίας είναι πολύ σημαντική για τον έλεγχο και την μηχανική λειτουργία και θα έπρεπε μέσω αυτής να μπορεί να αποκτηθεί η χωρική ροή της κυκλοφορίας μέσω της διεύθυνσης του Διαδικτύου (domain). Ένα παράδειγμα θα ήταν τα μονοπάτια (paths) που ακολουθούνται από πακέτα μεταξύ εισόδου και εξόδου της διεύθυνσης του Διαδικτύου. Από τις εν λόγω πληροφορίες μπορούν να επωφεληθούν πολλές εργασίες κατανομής των πόρων όπως επίσης και οι εργασίες του σχεδιασμού της χωρητικότητας. Επίσης, οι μετρήσεις της κυκλοφορίας πρέπει να λαμβάνονται χωρίς κάποιο μοντέλο δρομολόγησης και χωρίς γνώσεις της κατάστασης του δικτύου. Τα πιο πάνω επιτρέπουν την διαδικασία μέτρησης της κυκλοφορίας να είναι ελαστική σε τυχόν βλάβες του δικτύου και σε καταστάσεις αβεβαιότητας [11].

Πανεπιστήμιο Πειραιώς

ΕΝΟΤΗΤΑ 4

Εφαρμογή Δεδομένων

4.1. Τροχιές Πλοίων
4.2. Τιμές Κλεισίματος Μετοχών

Η ενότητα αυτή αποτελεί το πρακτικό κομμάτι και αποτελείται από την εφαρμογή δύο διαφορετικών ειδών δεδομένων, δι-διάστατα και τρι-διάστατα . Και στις δύο περιπτώσεις θα ασχοληθούμε με χρονοσειρές, μέσω των οποίων θα προσπαθήσουμε να εξάγουμε μία γενική συμπεριφορά. Τα δεδομένα μας αφορούν τροχιών πλοίων (πληροφορίες για τα σημεία στον χωροχρόνο κατά τη διάρκεια 3 ημερών μεταξύ των ημερομηνιών 31 Δεκεμβρίου 2008 και 02 Ιανουαρίου 2009) και χρονοσειρές από τιμές κλεισίματος διαφόρων μετοχών από το 2011 μέχρι και το 2014.

4.1. Τροχιές πλοίων

Σε αυτό το σημείο θα χρησιμοποιήσουμε δυσδιάστατα Timeseries με μη σταθερή χρονική απόσταση του ενός δείγματος από το άλλο [9]. Το συγκεκριμένο σύνολο δεδομένων έχει συλλεχθεί από την IMIS Hellas S.A.» και έχει δωρηθεί στο «Infolab» για σκοπούς ερευνών. Οι συγκεκριμένες τιμές αντιπροσωπεύουν τις τιμές τροχιών πλοίων, πληροφορίες για τα σημεία στον χωροχρόνο μεταξύ των ημερομηνιών 31 Δεκεμβρίου 2008 ώρα 19:29:30 και 02 Ιανουαρίου 2009 ώρα 17:10:06. Επίσης το γεωγραφικό πλάτος κυμαίνεται μεταξύ 34.96521579176599A° και 38.91619977511049A° ενώ το γεωγραφικό μήκος κυμαίνεται μεταξύ 21.09214590973562A° και 29.185142738919286A°.

Κάθε εγγραφή στο σύνολο δεδομένων έχει τη μορφή «obj_id, traj_id, t, lon, lat», όπου κάθε "obj_id" είναι μοναδικό και αντιπροσωπεύει ένα πλοίο. Το «traj_id» έχει την τιμή 1 σε όλες τις εγγραφές, εφόσον κάθε πλοίο έχει μόνο μία τροχιά. Επίσης, το «t» αντιπροσωπεύει την ώρα στη μορφή yyyy-MM-dd HH:mm:ss. Τέλος, οι μεταβλητές «lon» και «lat» αντιπροσωπεύουν τις συντεταγμένες στο Παγκόσμιο Γεωδαιτικό σύστημα WGS84 [1].

Κάπου εδώ θα θέλαμε να διευκρινίσουμε τους λόγους για τους οποίους είναι χρήσιμο να καταγραφούν τέτοιου είδους γεγονότα έτσι ώστε να γίνει σαφές το τι ψάχνουμε και που θέλουμε να καταλήξουμε με αυτή μας την έρευνα. Τα συγκεκριμένα δεδομένα μαζεύτηκαν έτσι ώστε να μπορεί να γίνει αναπαράσταση της διαδρομής των σκαφών για τον βέλτιστο και ασφαλή σχεδιασμό της διαδρομής των πλοίων μέσα στο Αιγαίο. Επίσης, με την καταγραφή του χωροχρόνου ενός πλοίου μπορούν να παράγονται ειδοποιήσεις για την πιθανότητα μιας σύγκρουσης ή προσάραξης ατυχήματος σε πραγματικό χρόνο, λαμβάνοντας υπόψη το ενδεχόμενο ανθρώπινου λάθους το οποίο μπορεί να επηρεάσει την τροχιά του πλοίου ή τους κινδύνους που συνδέονται με άλλα πλοία, σε συνδυασμό με τη θέση του και την προγραμματισμένη διαδρομή, το φορτίο του και τις μετεωρολογικές συνθήκες.

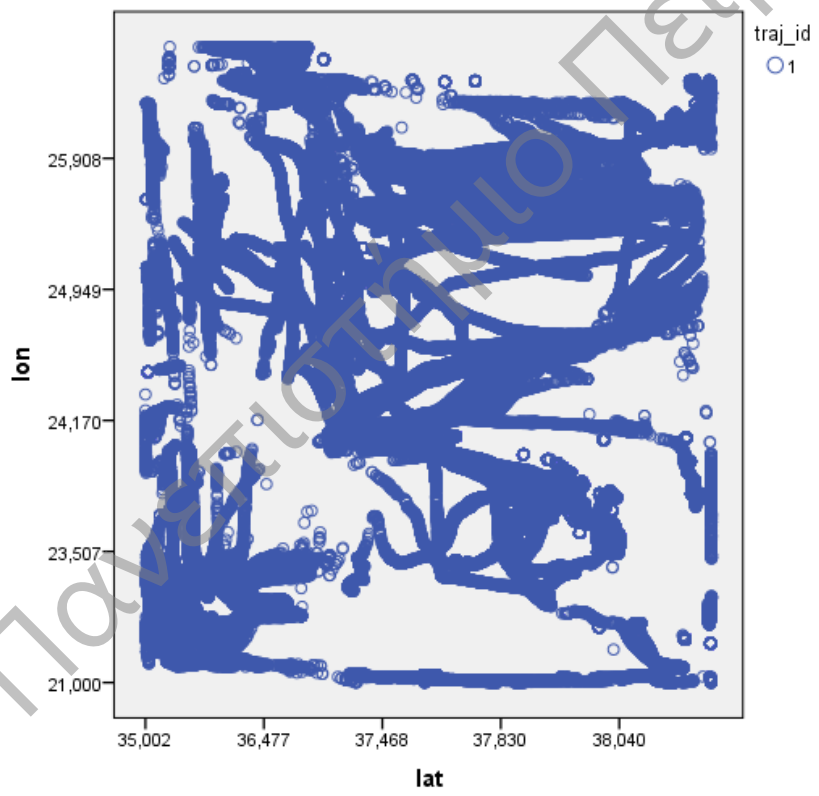
Πανεπιστήμιο Πειραιώς

Στη συνέχεια θα ακολουθήσει εφαρμογή των διαφορών τεχνικών δειγματοληψίας, με και χωρίς πιθανότητα, με βάση τις τιμές που μόλις έχουμε αναφέρει. Αξίζει να σημειωθεί ότι όλα τα στοιχεία του χωροχρόνου για τα πλοία είναι καταγραμμένα σε έναν κατάλογο, ο οποίος θα χρησιμοποιηθεί ως το δειγματοληπτικό πλαίσιο. Ο κατάλογος που έχουμε στα χέρια μας αποτελείται από 3097121 εγγραφές (γραμμές).

Πλήρες Δείγμα

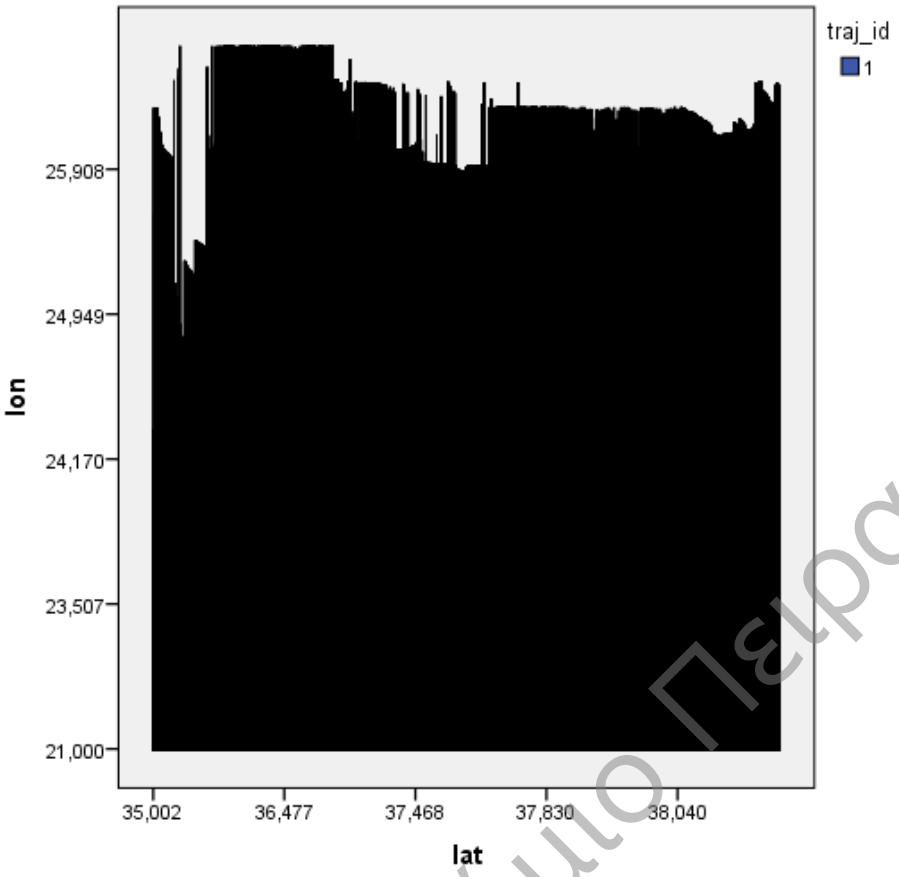
Σε πρώτο στάδιο παίρνουμε το πλήρες δείγμα με τις τροχιές πλοίων για ανάλυση, προσπαθώντας να βγάλουμε κάποια γενικά συμπεράσματα μέσα από τα σχήματα που θα παραχθούν. Για τη δημιουργία των σχημάτων χρησιμοποιήσαμε το data mining λογισμικού SPSS.

Στις γραφικές παραστάσεις που ακολουθούν βλέπουμε τα αποτελέσματα από την ανάλυση του ολοκληρωμένου δείγματος που αναπαριστάται από ένα γράφημα γράφημα ομαδοποιημένης διασποράς (grouped scatter). Οι μεταβλητές είναι οι εξής: «lat» που απευθύνεται στο γεωγραφικό πλάτος, «lon» που απευθύνεται στο γεωγραφικό μήκος και «obj_id» που απευθύνεται στον μοναδικό αριθμό πλοίου. Σε αυτή τη γραφική ομαδοποιούμε τη διασπορά και παρατηρούμε το εύρος τιμών όσο αφορά το γεωγραφικό πλάτος και το γεωγραφικό μήκος.



Εικόνα 14 – Γράφημα για Τροχιές Πλοίων

Πανεπιστήμιο Πειραιώς



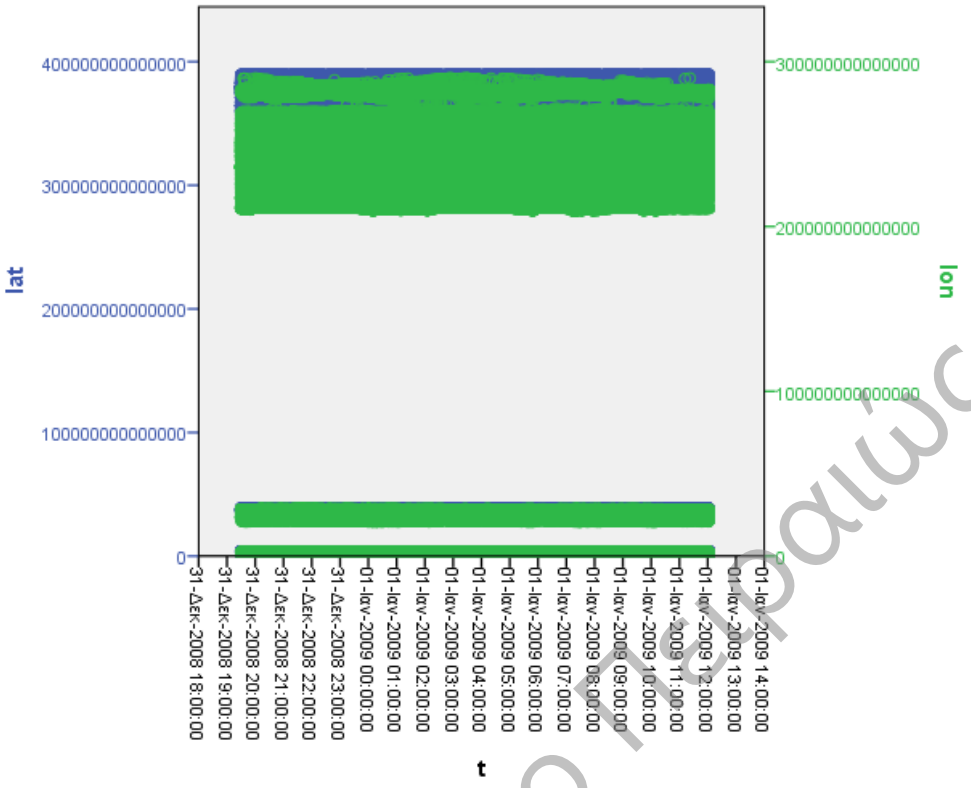
Εικόνα 15 Γράφημα για Τροχιές Πλοίων)

Μία παρατήρηση για τις πιο πάνω γραφικές παραστάσεις είναι οι πληροφορίες που ήδη προαναφέραμε, ότι δηλαδή το γεωγραφικό μήκος κυμαίνεται μεταξύ 21.09214590973562A° και 29.185142738919286A° ενώ το γεωγραφικό πλάτος κυμαίνεται μεταξύ 34.96521579176599A°.

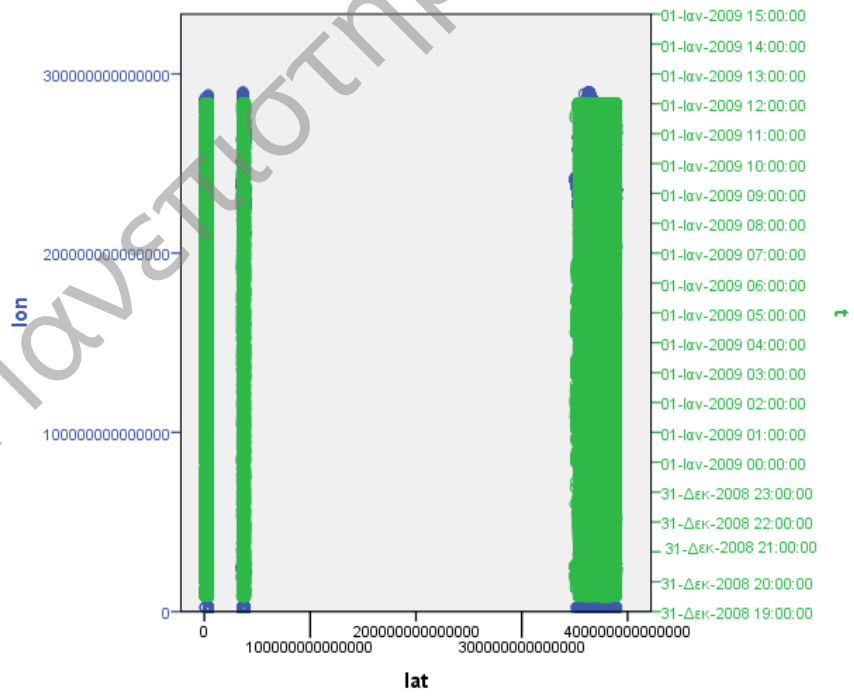
Στις δύο γραφικές παραστάσεις που προηγούνται αναλύουμε τα δεδομένα τροχιάς με βάση τον μοναδικό αριθμό πλοίου. Λόγω του γεγονότος ότι η τροχιά που σχηματίζει κάθε πλοίο δεν έχει να κάνει με τον μοναδικό του αριθμό δεν μπορούμε να εξάγουμε κάποιο περεταίρω συμπέρασμα βάση της συγκεκριμένης ανάλυσης.

Στη συνέχεια αναλύσαμε βλέπουμε τα αποτελέσματα από την ανάλυση του ολοκληρωμένου δείγματος που αναπαριστάται από ένα διπλών αξόνων. Σε αυτό γράφημα παρουσιάζουμε τα δεδομένα τροχιάς με βάση την ημερομηνία και την ώρα καταγραφής των σημείων. Οι γραφικές παραστάσεις π0ου ακολουθούν δεν δείχνουν κάτι το οποίο μπορεί να μας οδηγήσει σε κάποιο συμπέρασμα εφόσον τα σημεία είναι σε όλο το κομμάτι του σχήματος. Αυτό συμβαίνει επειδή για κάθε μέρα για κάθε πλοίο έχουμε εκατοντάδες εγγραφές.

Πανεπιστήμιο Πειραιώς



Εικόνα 16 - Τροχιές Πλοίων σε βάση ήμερομηνίας και ώρας



Εικόνα 17 - Τροχιές Πλοίων σε βάση ήμερομηνίας και ώρας

Πανεπιστήμιο Πειραιώς

Απλή Τυχαία Δειγματοληψία

Σε αυτό το είδος δειγματοληψίας επιλέγονται τυχαία από τον κατάλογο τόσα μέλη ώστε να δημιουργηθεί το πλήθος ίσο με το μέγεθος του δείγματος που χρειαζόμαστε. Όλα τα μέλη της λίστας έχουν την ίδια πιθανότητα να εκλεγούν. Για την επιλογή των τυχαίων αριθμών (με βάση τη θέση τους στον κατάλογο) θα χρησιμοποιήσουμε τη συνάρτηση RANDBETWEEN(), η οποία παρέχεται από το λογισμικό Excel της Microsoft.

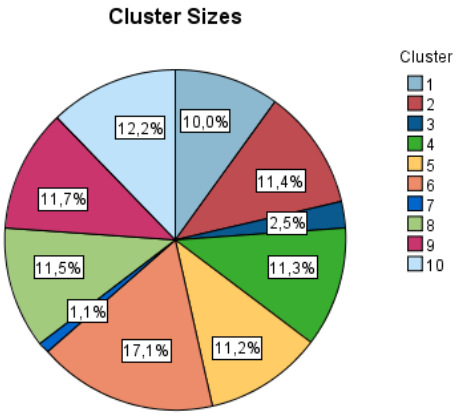
Σε αυτό το σημείο θα κάνουμε δειγματοληψία σε επίπεδο τροχιάς – όπου κάθε τροχιά αποτελείται από εκατοντάδες εγγραφές. Η λίστα με τους μοναδικούς αριθμούς πλοίων που έχουν επιλεγεί τυχαία για τη δειγματοληψία είναι οι ακόλουθες: 103, 136, 158, 192, 196, 267,304, 419, 729, 755.

Αρχικά εφαρμόσαμε τη δειγματοληψία Ομαδοποίησης (Clustering). Συγκεκριμένα, με τη βοήθεια του λογισμικού SPSS (που έχουμε χρησιμοποιήσει και για την προηγούμενη φάση) θα εφαρμόσουμε Two-Step Clustering. Η διαδικασία των δύο σταδίων θα μας επιτρέψει να χρησιμοποιήσουμε κατηγορικές μεταβλητές για τη διαμόρφωση ομάδων. Η συγκεκριμένη διαδικασία είναι επίσης χρήσιμη γιατί γνωρίζουμε εκ τν προτέρων το πλήθος των clusters, εφόσον θα δημιουργήσουμε τις ομάδες μας βάση μοναδικού αριθμού πλοίου. Ακόμη, δίνει τη δυνατότητα επιλογής των στατιστικών στοιχείων που μπορούν να χρησιμοποιηθούν για να συγκριθούν οι διάφορες λύσεις.

Για σκοπούς της κατηγοριοποίησης και δημιουργίας ομάδων έχουμε αποφασίσει να κάνουμε χρήση μοναδικού αριθμού πλοίου, ως τον βασικό παράγοντα. Έτσι, κατά την ταξινόμηση έχουν δημιουργηθεί δέκα ομάδες (clusters), μία για κάθε πλοίο. Κάθε cluster διαφορετικό πλήθος δεδομένων – με το μικρότερο cluster να αποτελείται από 1759 εγγραφές (1759 για γεωγραφικό μήκος και 1759 για γεωγραφικό πλάτος εφόσον μιλάμε για τροχιές) και το μεγαλύτερο να αποτελείται από 27194 έγγραφες.

Έπειτα από την εφαρμογή της μεθόδου Two-Step Clustering στο δείγμα μας, έχουμε καταλήξεις στα εξής αποτελέσματα: Το μεγαλύτερο cluster καταλαμβάνει το 17,1% του συνολικού πλήθους εγγραφών και απευθύνεται στο πλοίο με το μοναδικό αριθμό 267, ενώ το μικρότερο cluster καταλαμβάνει μόλις το 1,1% του συνολικού πλήθους εγγραφών και απευθύνεται στο πλοίο με το μοναδικό αριθμό 158.

Πανεπιστήμιο Πειραιώς



Size of Smallest Cluster	1759 (1,1%)
Size of Largest Cluster	27194 (17,1%)
Ratio of Sizes: Largest Cluster to Smallest Cluster	15,46

Εικόνα 18 - Μέγεθος των Clusters για Πλοία

Η γραφική παράσταση που ακολουθεί αποτελεί ένα γράφημα διπλών αξόνων που απεικονίζει το γεωγραφικό μήκος και πλάτος βάση του μοναδικού ID κάθε πλοίου. Συγκεκριμένα:

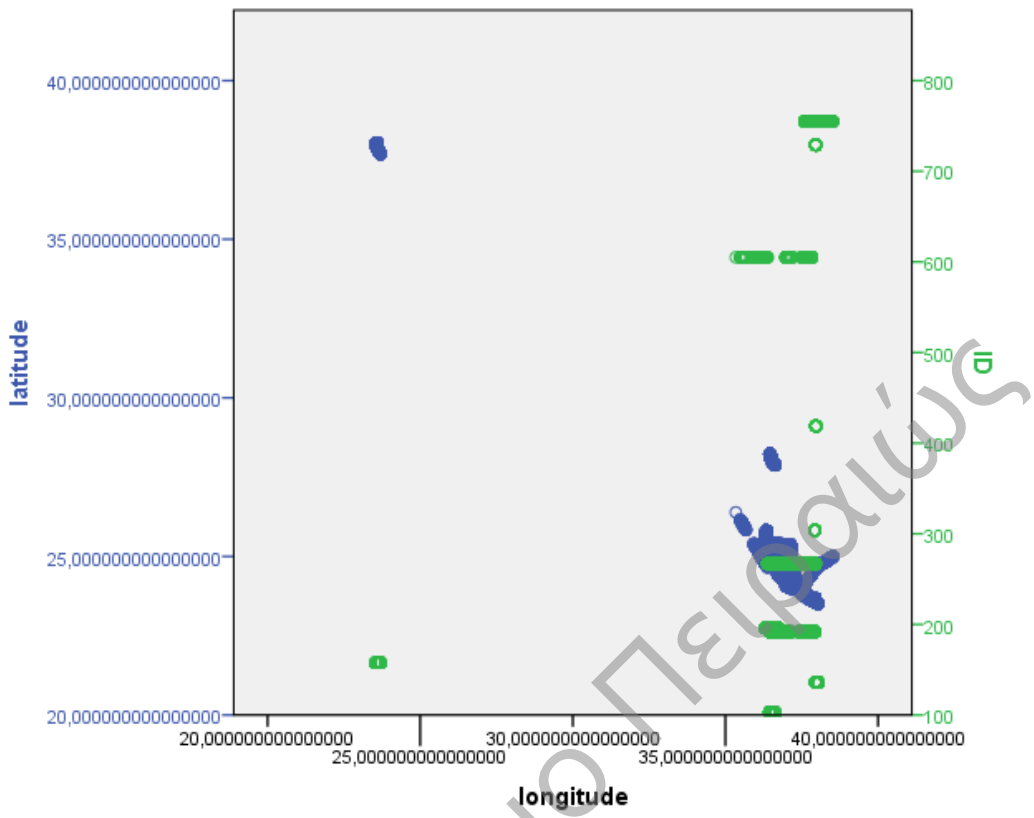
- Το πλοίο με μοναδικό αριθμό 103 σχηματίζει την πιο πάνω τροχιά. Οι τιμές για το γεωγραφικό του μήκος κυμαίνονται μεταξύ 36,4 – 36,65 ενώ οι τιμές για το γεωγραφικό του πλάτος κυμαίνονται μεταξύ 27,8 – 28,3.

- Το πλοίο με μοναδικό αριθμό 158 σχηματίζει την πιο πάνω τροχιά. Οι τιμές για το γεωγραφικό του μήκος κυμαίνονται μεταξύ 23,5 – 23,7 ενώ οι τιμές για το γεωγραφικό του πλάτος κυμαίνονται μεταξύ 37,7 – 38.

- Το πλοίο με μοναδικό αριθμό 729 σχηματίζει την πιο πάνω τροχιά. Οι τιμές για το γεωγραφικό του μήκος κυμαίνονται μεταξύ 37,9564 – 37,95665 ενώ οι τιμές για το γεωγραφικό του πλάτος κυμαίνονται μεταξύ 23,60385 – 23,6041.

- Το πλοίο με μοναδικό αριθμό 196 σχηματίζει την πιο πάνω τροχιά. Οι τιμές για το γεωγραφικό του μήκος κυμαίνονται μεταξύ 36,3 – 36,8 ενώ οι τιμές για το γεωγραφικό του πλάτος κυμαίνονται μεταξύ 24,5 – 26.

Πανεπιστήμιο Πειραιώς



Εικόνα 19 - Τροχιές Πλοίων σε βάση μοναδικού ID

Συγκρίνοντας την Εικόνα 17 με την Εικόνα 22 παρατηρούμε πως αν και το δείγμα μας στην συγκεκριμένη δειγματοληψία είναι μικρό η διασπορά των πλοίων είναι παρόμοια, πράγμα που καθιστά τη δειγματοληψία μας επιτυχή.

Δειγματοληψία με διαστρωμάτωση

Σε αυτό το είδος δειγματοληψίας γίνεται διαίρεση των μελών του πληθυσμού σε ομογενείς υποομάδες πριν από τη δειγματοληψία. Κάθε στοιχείο του πληθυσμού θα πρέπει να ανατεθεί σε ένα μόνο στρώμα έτσι ώστε να μην υπάρχει πιθανότητα επαναχρησιμοποίησης.

Σε πρώτο στάδιο γίνεται χωρισμός του πληθυσμού σε «στρώματα» (κατά προτίμηση με εσωτερική ομοιογένεια αλλά ετερογενή μεταξύ τους) ενώ ακολουθεί λήψη τυχαίου δείγματος μονάδων σε κάθε ένα από τα προεπιλεγμένα «στρώματα». Τέλος γίνεται ένας συνδυασμός των αποτελεσμάτων όλων των «στρωμάτων» έτσι ώστε να παρθούν τα αποτελέσματα.

Σε αυτή την περίπτωση και εφόσον έχουμε μόνο πληροφορίες για το μοναδικό αριθμό πλοίου, το γεωγραφικό πλάτος και την ημερομηνία είναι δύσκολο να γίνει διαστρωμάτωση. Η συγκεκριμένη τεχνική θα μπορούσε να εφαρμοστεί εάν για παράδειγμα υπήρχε μία επιπλέον κατηγορία «είδος πλοίου» π.χ. ερευνητικά, πολεμικά κτλ ούτως ώστε να μπορούσαν τα πλοία μας να χωριστούν σε ομάδες.

Πανεπιστήμιο Πειραιώς

Δειγματοληψία κατά συστάδες

Σε αυτό το είδος δειγματοληψίας ο συνολικός πληθυσμός διαιρείται στις «φυσικές» ομάδες και επιλέγεται ένα δείγμα από αυτές.

Αρχικά θα γίνει ταξινόμηση του πληθυσμού σε συστάδες (=ομάδες) από τις τροχιές ενώ στη συνέχεια θα γίνει λήψη τυχαίου δείγματος συστάδων. Τέλος, το σύνολο ή μέρος των τροχιών από τις επιλεγμένες συστάδες περιλαμβάνεται στο τελικό δείγμα.

Εφόσον για το συγκεκριμένο δείγμα μας δεν υπάρχουν «φυσικές» ομάδες ούτε η εν λόγω διαδικασία δειγματοληψίας μπορεί να εφαρμοστεί. Όπως και πιο πριν, μοναδικό αριθμό πλοίου, το γεωγραφικό πλάτος και την ημερομηνία είναι δύσκολο να γίνει διαστρωμάτωση. η συγκεκριμένη τεχνική θα μπορούσε να εφαρμοστεί εάν για υπήρχε μία επιπλέον κατηγορία η οποία θα μπορούσε να διαχωρίσει το δείγμα σε ομάδες.

Δειγματοληψία αναλογίας ή ποσοστιαία δειγματοληψία

Και τώρα θα εφαρμόσουμε μία τεχνική δειγματοληψίας χωρίς πιθανότητα – εφόσον οι τρεις τεχνικές που προηγήθηκαν είναι τεχνικές με πιθανότητα.

Σε αυτό το είδος επιλέγουμε το δείγμα με κάποιο σκοπό και έτσι έχουμε συνήθως μία ή περισσότερες συγκεκριμένες προκαθορισμένες ομάδες στις οποίες στοχεύουμε.

Δυστυχώς και πάλι δεν υπάρχει κάποιο χαρακτηριστικό βάση του οποίου θα μπορούσαμε να διαχωρίσουμε το δείγμα μας σε ομάδες.

Επιλογή Κατάλληλης Δειγματοληπτικής Διαδικασίας

Για το συγκεκριμένο δείγμα, το οποίο αποτελείται από τροχιές πλοίων, έχουμε καταλήξει στο συμπέρασμα ότι οι πλύστες δειγματοληπτικές διαδικασίες δεν είναι εφαρμόσιμες λόγω του ότι τα στοιχεία που έχουμε δεν μπορούν να διαχωριστούν σε ομάδες. Ο λόγος που συμβαίνει αυτό είναι ότι το δείγμα μας αποτελείται από το γεωγραφικό μήκος και πλάτος, μοναδικό αριθμό πλοίου και ημερομηνία καταγραφής των σημείων. Τα χαρακτηριστικά αυτά δεν αλληλο-αποκλείονται κι έτσι δεν μπορούν να διαχωριστούν. Για να γίνουμε πιο σαφής, εάν για παράδειγμα διαχωρίσουμε τα δεδομένα μας βάση μοναδικού αριθμού πλοίου αυτό δε σημάνει πως κάποιο άλλο πλοίο δεν έχει στην τροχιά του τα ίδια σημεία με το πρώτο.

Η μόνη δειγματοληπτική διαδικασία που καταφέραμε να εφαρμόσουμε είναι η απλή τυχαία δειγματοληψία, η οποία μας έφερε κάποιο επιθυμητό αποτέλεσμα, όμως γενικά είναι μια δαπανηρή διαδικασία.

Πανεπιστήμιο Πειραιώς

4.2. Τιμές κλεισίματος Μετοχών

Για την έμπρακτη πτυχή αυτής της εργασίας θα παρουσιάσουμε πραγματικά δεδομένα μέσω των οποίων θα εξάγουμε κάποια γενικευμένα συμπεράσματα. Για την συλλογή των δεδομένων δεν έχει διεξαχθεί κάποια προσωπική δειγματοληψία αλλά θα χρησιμοποιηθούν ήδη δημοσιευμένα στοιχεία από τις τιμές κλεισίματος των μετοχών για κάποιες εταιρείες τα τελευταία τρία χρόνια [8].

Για την ανάλυση μας αυτή θα χρησιμοποιήσουμε αναλυτικές τιμές κλεισίματος διαφόρων μετοχών από το 2011 μέχρι και το 2014. Οι τιμές αυτές θα επεξεργαστούν με τη βοήθεια της χρήσης κάποιων από τις μεθόδους δειγματοληψίας που έχουμε δει πιο πριν.

Για την μελέτη μας θα παρακολουθήσουμε εταιρείες μετοχών που ανήκουν στις κατηγορίες της κύριας αγοράς, της χαμηλής διασποράς και στην κατηγορία προς κλείσιμο. Η Κύρια Αγορά απευθύνεται κυρίως σε μεσαίες και μεγάλες εταιρείες κεφαλαιοποίησης με προοπτικές ανάπτυξης, ενώ παρέχει πρόσβαση σε διασυνοριακή άντληση κεφαλαίων και στη συμμετοχή των μεριδίων στους ευρωπαϊκές δείκτες. Στην κατηγορία της χαμηλής διασποράς ανήκουν οι εταιρείες που έχουν χαμηλή κατοχή των μετοχών τους από μετόχους καθώς επίσης και χαμηλή διάθεση κεφαλαίου σε αγορά πολλών ειδών μετοχών, για λόγους ασφαλείας. Τέλος, στην κατηγορία «Προς Κλείσιμο» ανήκουν οι εταιρείες που δεν τα πάνε καθόλου καλά και οδεύουν προς κλείσιμο.

Πλήρες Δείγμα

Για αυτό το δείγμα ήτανε αδύνατον να συγκεντρώσουμε τις τιμές κλεισίματος για όλες τις μετοχές εφόσον δεν υπήρχανε κάπου συγκεντρωμένες. Για το λόγο αυτό, σε αυτό το κομμάτι θα συμπεριλάβουμε το γενικό δείκτη τιμών για τη διάρκεια των τεσσάρων χρόνων – όπως τον βρήκαμε διαδικτυακά.

Εικόνα 20 - Γενικός Δείκτης τιμών

Πανεπιστήμιο Πειραιώς

Απλή Τυχαία Δειγματοληψία

Σε αυτό το είδος δειγματοληψίας επιλέγονται τυχαία από τον κατάλογο τόσα μέλη ώστε να δημιουργηθεί το πλήθος ίσο με το μέγεθος του δείγματος που χρειαζόμαστε. Όλα τα μέλη της λίστας έχουν την ίδια πιθανότητα να εκλεγούν. Για την επιλογή των τυχαίων αριθμών (με βάση τη θέση τους στον κατάλογο) θα χρησιμοποιήσουμε τη συνάρτηση RANDBETWEEN(), η οποία παρέχεται από το λογισμικό Excel της Microsoft.

Για σκοπούς τυχαίας επιλογής των μετοχών έχουμε αντιστοιχίσει τα διάφορα ονόματα των μετοχών που βρίσκονται διαθέσιμα στο διαδίκτυο. Έπειτα κάναμε χρήση της RANDBETWEEN() και έχουμε καταλήξει στις εξής μετοχές:

- Alpha Astika Akinita
- Kathimerini
- Alsinco
- PCSystems
- Alpha Bank
- Trapeza Ellados

Κάπου εδώ να αναφερθεί ότι κάθε μετοχή αποτελείται από 1248 εγγραφές για τη χρονική περίοδο 2009 – 2014 (26/6/2009 έως 26/6/2014).

Σε αυτό το σημείο και αφότου απομονώσαμε μόνο τα χαρακτηριστικά που μας ενδιέφεραν (ημερομηνία και τιμή κλεισίματος κάθε μετοχής ανά ημέρα) κάναμε χρήση του data mining λογισμικού SPSS.

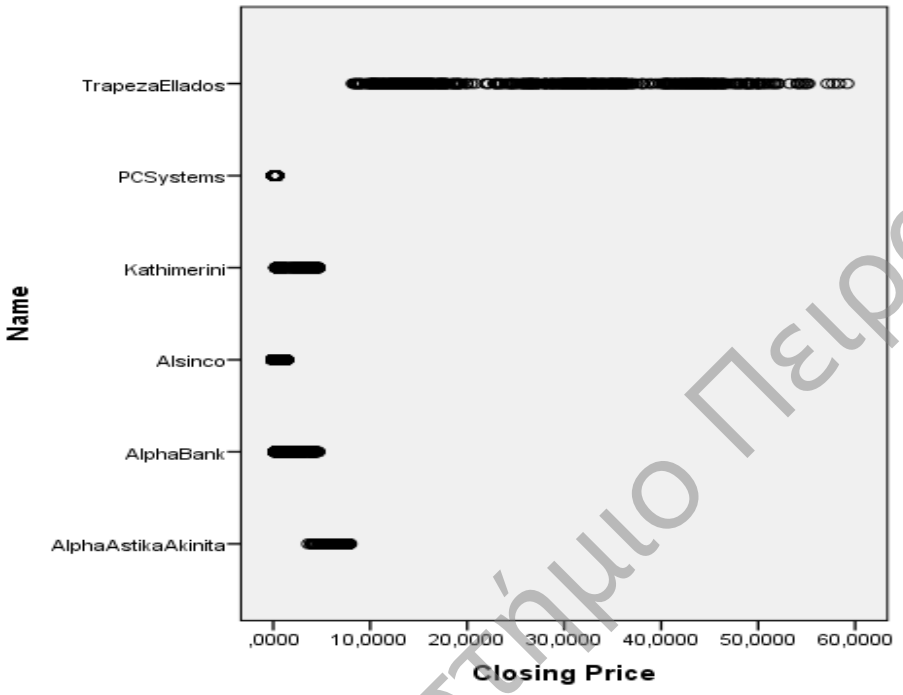
Πιο κάτω ακολουθεί η γραφική παράσταση η οποία απεικονίζει τις τιμές κλεισίματος για κάθε μία από τις 6 μετοχές κατά τη διάρκεια των 4 χρόνων. Συγκεκριμένα:

- Η μετοχή της Alpha Astika Akinita: Οι τιμές για το 2010 ξεκινάνε από τα 8 ευρώ, έπειτα παρατηρείται μία μεγάλη πτώση στο τέλος του 2012 και μία ανάκαμψη μέχρι το τέλος του 2014. Στατιστικά η τιμή της συγκεκριμένης μετοχής έχει πέσει κατά 1,5 ευρώ από το 2010 μέχρι και το 2014.
- Η μετοχή της Kathimerinis: Οι τιμές για το 2010 ξεκινάνε από τα 4,5 ευρώ, έπειτα παρατηρείται μία μεγάλη πτώση η οποία ξεκινάει στα μέσα του 2011 και συνεχίζεται εώς και τα μέσα του 2012. Έπειτα η πορεία της μετοχής θα χαρακτηριζόταν σταθερή. Γενικά η τιμή της συγκεκριμένης μετοχής έχει πέσει κατά 3,5 ευρώ από το 2010 μέχρι και το 2014.
- Η μετοχή της PC Systems: Οι τιμές για το 2010 ξεκινάνε από τα 0,35 ευρώ, έπειτα παρατηρείται μία σταθερή πτώση και κάποια τακτικά σκαμπανεβάσματα. Η τιμή της συγκεκριμένης μετοχής έχει πέσει κατά5 0,30 ευρώ περίπου από το 2010 μέχρι και το 2014, μία πτώση της τάξεως του 75%.
- Η μετοχή της Alsinco: Οι τιμές για το 2010 ξεκινάνε από τα 1,35 ευρώ ενώ στη συνέχεια παρατηρείται μία μεγάλη και σταθερή πτώση που οδηγεί τη μετοχή σχεδόν σε μηδενική τιμή κάπου στα μέσα του 2013. Στην πορεία η μετοχή ανακάμπτει ελάχιστα εφόσον κυμαίνεται 0,4 ευρώ κατά το 2014. Η τιμή της συγκεκριμένης μετοχής έχει πέσει κατά 0,95 ευρώ από το 2010 μέχρι και το 2014, ποσοστό που ξεπερνά το -60%.
- Η μετοχή της Trapeza Ellados: Οι τιμές για το 2010 ξεκινάνε από τα 40 ευρώ ενώ στη συνέχεια παρατηρείται μία μεγάλη αύξηση, η οποία αγγίζει τα 60 ευρώ. Στη συνέχεια σημειώνεται σταθερή

Πανεπιστήμιο Πειραιώς

πτώση και έπειτα άνοδος. Η τιμή της συγκεκριμένης μετοχής από το 2010 έως το 2014 έχει πέσει κατά 25 ευρώ, ποσοστό που αγγίζει το -60%. Παρόλα αυτά η τιμή της μετοχής εξακολουθεί να είναι υψηλή.

- Η μετοχή της Alpha: Οι τιμές για το 2010 ξεκινάνε από τα 2,5 ευρώ ενώ στη συνέχεια παρατηρείται μία μεγάλη αύξηση, η οποία αγγίζει τα 4,5 ευρώ. Στη συνέχεια σημειώνεται σταθερή πτώση και έπειτα μικρή άνοδος. Η τιμή της συγκεκριμένης μετοχής από το 2010 έως το 2014 έχει πέσει κατά 1,5 ευρώ, ποσοστό που αγγίζει το -70%.



Εικόνα 21 - Τιμές Κλεισίματος Μετοχών

Δειγματοληψία με διαστρωμάτωση

Η δειγματοληψία αυτή ανήκει στην κατηγορία δειγματοληψίας με πιθανότητα. Σε αυτό το είδος δειγματοληψίας γίνεται διαίρεση των μελών του πληθυσμού σε ομογενείς υποομάδες πριν από τη δειγματοληψία. Κάθε στοιχείο του πληθυσμού θα πρέπει να ανατεθεί σε ένα μόνο στρώμα έτσι ώστε να μην υπάρχει πιθανότητα επαναχρησιμοποίησης.

Σε πρώτο στάδιο γίνεται χωρισμός του πληθυσμού σε «στρώματα» (κατά προτίμηση με εσωτερική ομοιογένεια αλλά ετερογενή μεταξύ τους) ενώ ακολουθεί λήψη τυχαίου δείγματος μονάδων σε κάθε ένα από τα προεπιλεγμένα «στρώματα». Τέλος γίνεται ένας συνδυασμός των αποτελεσμάτων όλων των «στρωμάτων» έτσι ώστε να παρθούν τα αποτελέσματα.

Εφόσον δεν είχαμε κάποιο χαρακτηριστικό που να μπορεί να χρησιμοποιηθεί για ομαδοποίηση των χρονοσημασμένων δεδομένων, προσθέσαμε ένα. Συγκεκριμένα, κάνουμε χρήση του χαρακτηριστικού «Κατηγορία» το οποίο συμπεριλαμβάνει τις εξής τρεις κατηγορίες:

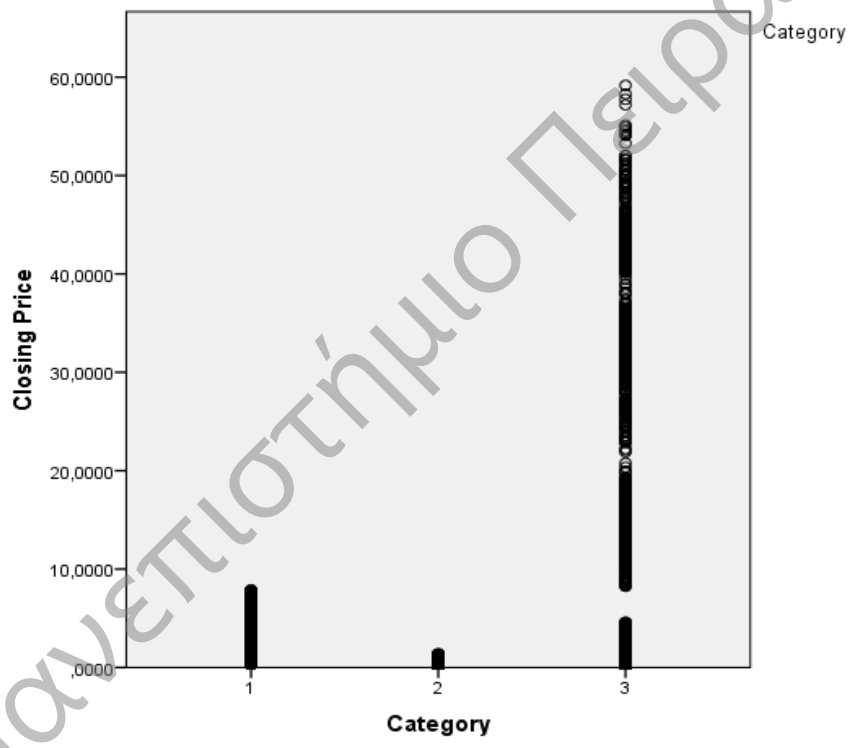
- Χαμηλή Διασπορά (Κατηγορία 1)

Πανεπιστήμιο Πειραιώς

- Alpha Astika Akinita
- Kathimerini

- Προς Κλείσιμο (Κατηγορία 2)
  - Alsinco
  - PCSystems
- Κύρια Αγορά (Κατηγορία 3)
  - Alpha Bank
  - Trapeza Ellados

Έτσι, κατά την ομαδοποίηση έχουν δημιουργηθεί τρεις ομάδες (clusters), ένα για κάθε μία από της κατηγορίες. Κάθε cluster αποτελείται από ίσο πλήθος δεδομένων, δηλαδή κάθε cluster αποτελείται από το 33,3% του συνολικού πλήθους δεδομένων. Εφόσον κάθε κατηγορία αποτελείται από 2 διαφορετικές μετοχές, το πλήθος εγγραφών κάθε ομάδας είναι 2496.



Εικόνα 22 - Τιμές Κλεισίματος Μετοχών βάση Κατηγορίας

Παρατηρώντας τα αποτελέσματα θα τα χαρακτηρίζαμε αναμενόμενα εφόσον η ομάδα με τις ψηλότερες τιμές κλεισίματος είναι η ομάδα των μετοχών που βρίσκονται στην Κύρια Αγορά. Εφόσον η συγκεκριμένη κατηγορία μετοχών απευθύνεται κυρίως σε μεσαίες και μεγάλες εταιρείες κεφαλαιοποίησης με προοπτικές ανάπτυξης, ενώ παρέχει πρόσβαση σε διασυνοριακή άντληση κεφαλαίων και στη συμμετοχή των μεριδίων στους ευρωπαϊκές δείκτες, είναι λογικό οι τιμές των μετοχών να βρίσκονται σε υψηλά επίπεδα. Επίσης, η ομάδα με τις χαμηλότερες τιμές κλεισίματος είναι η ομάδα των μετοχών που οδεύουν προς Κλείσιμο. Και εδώ να σημειώσουμε πως προσδοκώμενο για μία μετοχή που πάει για κλείσιμο είναι να έχει μία πολύ χαμηλή – έως μηδενική - τιμή πώλησης.

Πανεπιστήμιο Πειραιώς

Ακολουθεί ένα πίνακας με τους μέσους όρους τιμών κλεισίματος των μετοχών για κάθε μία από τις τρεις κατηγορίες. Ο πίνακας δημιουργήθηκε μέσω του λογισμικού SPSS και της τεχνικής Two-Step Clustering [20]. Η διαδικασία των δύο σταδίων θα μας επιτρέψει να χρησιμοποιήσουμε κατηγορικές μεταβλητές για τη διαμόρφωση ομάδων. Η συγκεκριμένη διαδικασία είναι επίσης χρήσιμη γιατί γνωρίζουμε εκ των προτέρων το πλήθος των clusters, εφόσον θα δημιουργήσουμε τις ομάδες μας βάση μοναδικού αριθμού πλοίου. Ακόμη, δίνει τη δυνατότητα επιλογής των στατιστικών στοιχείων που μπορούν να χρησιμοποιηθούν για να συγκριθούν οι διάφορες λύσεις.

Cluster	1	2	3
Size	33,4% (2498)	33,3% (2496)	33,3% (2496)
Inputs	Category 3 (100,0%)	Category 1 (100,0%)	Category 2 (100,0%)
	Closing Price 12,56	Closing Price 4,65	Closing Price 0,28

Εικόνα 23 - Ομάδες μετοχών

Επιλογή Κατάλληλης Δειγματοληπτικής Διαδικασίας

Για το συγκεκριμένο δείγμα, το οποίο αποτελείται από τις τιμές κλεισίματος μετοχών, και αποτελούν χρονοσειρές, έχουμε καταλήξει στο συμπέρασμα ότι η δειγματοληπτική διαδικασία με διαστρωμάτωση είναι η πιο κατάλληλη.

Κάνοντας χρήση της Δειγματοληψίας με διαστρωμάτωση καταλήξαμε στο συμπέρασμα που αναμέναμε, ότι δηλαδή οι μετοχές που βρίσκονται στην «Κύρια Αγορά» έχουν υψηλές τιμές κλεισίματος και ιδιαίτερα πολύ υψηλότερες από τις τιμές των μετοχών που βρίσκονται στις άλλες δύο κατηγορίες - «Χαμηλή Διασπορά» και «Προς Κλείσιμο». Επίσης, οι μετοχές που ανήκουν στην κατηγορία «Προς Κλείσιμο» έχουν πολύ χαμηλές τιμές κλεισίματος.

Πανεπιστήμιο Πειραιώς

ΕΝΟΤΗΤΑ 5

Συμπεράσματα

5.1. Συμπεράσματα

Η τελευταία μας ενότητα αποτελεί την σύνοψη όλων των προηγούμενων κεφαλαίων. Σε αυτό το κομμάτι θα παρουσιαστούν τα αποτελέσματα και τα συμπεράσματα των όσων έχουν προηγηθεί.

5.1. Συμπεράσματα

Σε αυτή τη διπλωματική εργασία έχουμε μελετήσει τη διαδικασία της δειγματοληψίας και τις διαφορετικές τις μεθόδους για τη συγκέντρωση στατιστικών στοιχείων, οι οποίες μπορούν να χωριστούν σε δύο μεγάλες ομάδες, τις απογραφές και τις δειγματοληπτικές έρευνες. Η διαδικασία επιλογής διακρίνεται σε 2 βασικά είδη δειγματοληψίας: την τυχαία δειγματοληψία ή αλλιώς δειγματοληψία με πιθανότητα (Probability Sampling) και τη συστηματική δειγματοληψία ή αλλιώς δειγματοληψία χωρίς πιθανότητα (Nonprobability sampling).

Εφόσον έχει γίνει αναφορά, περιγραφή και έχουν δοθεί παραδείγματα για τις διάφορες τεχνικές δειγματοληψίας σε σύνθετους τύπους δεδομένων έχει γίνει και εφαρμογή τους σε πραγματικές τιμές που έχουν παρθεί από υπάρχουσες βάσεις δεδομένων.

Η δειγματοληψία ως διαδικασία είναι πολύ σημαντική για την μελέτη και την παρακολούθηση συνόλων και βοηθάει στην εξαγωγή συμπερασμάτων για γενικές συμπεριφορές του κάθε συνόλου. Μπορεί επίσης να προβλέψει μελλοντικές συμπεριφορές και να προτρέψει ανεπιθύμητα γεγονότα.

Μέσω της εφαρμογής κάποιων από τις τεχνικές δειγματοληψίας στους σύνθετους τύπους δεδομένων (τροχιές πλοίων και τιμές κλεισίματος μετοχών) έχουμε καταλήξει στο συμπέρασμα ότι δεν μπορούν όλες οι τεχνικές να είναι κατάλληλες σε κάθε περίπτωση. Συγκεκριμένα, και στους δύο περιπτώσεις έχουμε αρχικά εφαρμόσει την απλή τυχαία δειγματοληψία και δεν απέφερε κάποιο ουσιαστικό αποτέλεσμα ενώ στη συνέχεια εφαρμόσαμε τη δειγματοληψία Ομαδοποίησης (Clustering) και μας έδωσε την ευκαιρία να εξάγουμε αποτελέσματα και να κάνουμε κάποιες συγκρίσεις.

Κλείνοντας, η επιλογή των διάφορων τεχνικών δειγματοληψίας σε σύνθετους τύπους δεδομένων πρέπει να γίνει προσεκτικά ανάλογα με το τι είναι το ζητούμενο, ποιο είναι το δείγμα αλλά και βάση του πώς είναι η δομή των δεομένων που θα εξετασθούν.

Πανεπιστήμιο Πειραιώς

ΕΝΟΤΗΤΑ 6

Βιβλιογραφικές Αναφορές

6.1. Βιβλιογραφία

[1] *«Μέθοδοι Συγκέντρωσης Δεδομένων»*. Statistics Scientist. [Online]. Διαθέσιμο στο: <http://statistics.scientist.gr/122.pdf> [Πρόσβαση: 16 Ιουν. 2014]

[2] *«Time Series»*. Wikipedia. [Online]. Διαθέσιμο στο: <http://en.wikipedia.org/wiki/Time_series> [Πρόσβαση: 10 Ιουν. 2014]

[3] *«Time Series: Theory and Methods»*. Peter J.Brockwell, Richard A. Davis, Google Books. [Online]. Διαθέσιμο στο: <http://books.google.com.cy/books?id=_DcYu_EhVzUC&printsec=frontcover&dq=time+series&hl=el&sa=X&ei=LFaZU9OaMYeo0AXotIGIBA&ved=0CCUQ6AEwAQ#v=onepage&q=time%20series&f=false> [Πρόσβαση: 12 Ιουν. 2014]

[4] *«Δειγματοληψία»*. Eudoxus. [Online]. Διαθέσιμο στο: <https://static.eudoxus.gr/books/65/chapter-11765.pdf> [Πρόσβαση: 12 Ιουν. 2014]

[5] *«ΑΠΛΗ ΤΥΧΑΙΑ ΔΕΙΓΜΑΤΟΛΗΨΙΑ»*. Οικονομικό Πανεπιστήμιο Αθηνών . [Online]. Διαθέσιμο στο: <http://www.stat-athens.aueb.gr/~exek/Sampling-Techniques/chapter2.pdf> [Πρόσβαση: 12 Ιουν. 2014]

[6] *«Δειγματοληψία στην επιδημιολογική επιτήρηση και διεύρυνση επιδημιών»*. Εθνική Σχολή Δημόσιας Υγείας. Πρόγραμμα εκπαίδευσης στην επιδημιολογική επιτήρηση και διερεύνηση επιδημιών. ΕΣΔΥ-ΚΕΕΛΠΝΟ, 2008 Διαθέσιμο στο: <http://www.nsph.gr/files/011_Ygeias_Paidiou/Epidimiologiki_epitirisi_mathimata/Digmatolipsia.pdf > [Πρόσβαση: 23 Ιουν. 2014]

[7] *«Accidental sampling»*.Wikipedia.[Online].Διαθέσιμο στο: <http://en.wikipedia.org/wiki/Accidental_sampling> [Πρόσβαση: 23 Ιουν. 2014]

[8] *«Οικονομία & Αγορές - Χρηματιστήριο»*. naftemporiki.gr. [Online]. Διαθέσιμο στο: <http://www.naftemporiki.gr/> [Πρόσβαση: 25 Ιουν. 2014]

Πανεπιστήμιο Πειραιώς

[9] *«Imis3Dats»*.Chorochronos.org.[Online].Διαθέσιμο στο: <http://www.chorochronos.org/?q=node/8> [Πρόσβαση: 25 Ιουν. 2014]

[10] *«Plot Lat/Long Points on Map by Coordinates»*. Darrin ward [Online]. Διαθέσιμο στο: <http://www.darrinward.com/lat-long> [Πρόσβαση: 26 Ιουν. 2014]

[11] N. G. Duffield, & M. Grossglauser. *«Trajectory Sampling for Direct Traffic Observation»*. IEEE/ACM Transactions on Networking (TON), Volume 9 Issue 3, June 2001, Pages 280-292. Darrin ward [Online]. Διαθέσιμο στο: <http://www.csd.uoc.gr/~hy558/papers/duffield01trajectory.pdf> [Πρόσβαση: 26 Ιουν. 2014]

[12] Nikos Pelekis, Ioannis Kopanakis, Costas Panagiotakis, & Yiannis Theodoridis. *«Unsupervised Trajectory Sampling»*. Proceedings of the 2010 European conference on Machine learning and knowledge discovery in databases part III, pages:17-33. Διαθέσιμο στο: Machine Learning and Knowledge Discovery in Databases.

[13] *«Real-Time Website Geo Visitor Tracker»*. Track my IP [Online]. Διαθέσιμο στο: <http://www.tracemyip.org/tools/website-visitors-counter-traffic-tracker-statistics/> [Πρόσβαση: 30 Ιουν. 2014]

[14] *«Cluster Analysis»*. Wikipedia. [Online]. Διαθέσιμο στο: <http://en.wikipedia.org/wiki/Cluster_analysis> [Πρόσβαση: 10 Ιουλ. 2014]

[15] *«Research Methods Knowledge Base»*. Social Research Methods. [Online]. Διαθέσιμο στο: <http://www.socialresearchmethods.net/kb/sampprob.php> [Πρόσβαση: 15 Ιουλ. 2014]

[16] *«Studies in Sampling Techniques and Time Series Analysis»*. Rajesh Singh, Florentin Smarandache. Gallup.[Online].Διαθέσιμο στο: <http://www.gallup.unm.edu/~smarandache/TimeSeriesAnalysis.pdf> [Πρόσβαση: 20 Ιουλ. 2014]

[17] *«Web analytics»*. Wikipedia. [Online]. Διαθέσιμο στο: <http://en.wikipedia.org/wiki/Web_analytics> [Πρόσβαση: 23 Ιουλ. 2014]

[18] *«Tracking Clicks on Ads With Google Analytics»*. Imavex. [Online]. Διαθέσιμο στο: <http://www.imavex.com/tracking-clicks-on-ads-with-google-analytics.blog> [Πρόσβαση: 23 Ιουλ. 2014]

[19] *«World Geodetic System»*. Wikipedia. [Online]. Διαθέσιμο στο: <http://en.wikipedia.org/wiki/World_Geodetic_System> [Πρόσβαση: 24 Ιουλ. 2014].

[20] *«TwoStep Cluster Analysis».* IBM. [Online]. Διαθέσιμο στο: <http://pic.dhe.ibm.com/infocenter/spssstat/v22r0m0/index.jsp?topic=%2Fcom.ibm.spss.statistics.help%2Fspss%2Fbase%2Fidh_twostep_main.htm> [Πρόσβαση: 01 Σεπτ. 2014].

Πανεπιστήμιο Πειραιώς

[21] «*Segmentation and Sampling of Moving Object Trajectories Based on Representativeness»,* Panagiotakis, C., Pelekis, N., Kopanakis, I., Ramasso, E., Theodoridis, Y. Journal, IEEE Transactions on Knowledge and Data Engineering, (Volume:24 , Issue: 7), pages: 1328 – 1343. July 2012. Διαθέσιμο στο: <http://ieeexplore.ieee.org/xpl/login.jsp?tp=&arnumber=5710924&url=http%3A%2F%2Fieeexplore.ieee.org%2Fiel5%2F69%2F4358933%2F05710924.pdf%3Farnumber%3D5710924> [Πρόσβαση: 09 Σεπτ. 2013]

6.2. Παράρτημα Χρήσιμων Ορολογιών

[1] Παγκόσμιο Γεωδαιτικό σύστημα WGS84: είναι ένα πρότυπο για χρήση στη χαρτογραφία, τη γεωδαισία και την πλοήγηση. Αποτελείται από ένα πρότυπο σύστημα συντεταγμένων για τη Γη, μια τυπική επιφάνεια αναφοράς σφαιροειδή (το δεδομένο ή ελλειψοειδές αναφοράς) για τα ανεπεξέργαστα δεδομένα υψομέτρου, και μία βαρυτική ισοδυναμική επιφάνεια (γεωειδές) που καθορίζει την ονομαστική στάθμη της θάλασσας [19].

Πανεπιστήμιο Πειραιώς