
ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΑΓΡΟΝΟΜΩΝ ΚΑΙ ΤΟΠΟΓΡΑΦΩΝ ΜΗΧΑΝΙΚΩΝ

ΤΟΜΕΑΣ ΤΟΠΟΓΡΑΦΙΑΣ


# **“ Τεχνικές και Μέθοδοι Χωρικής Ανάλυσης Ιατρικών Δεδομένων: Περιβαλλοντικοί παράγοντες που επηρεάζουν την Παχυσαρκία και τη Λοίμωξη από *Clostridium difficile* στην πολιτεία της Καλιφόρνια ”**

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ


**Μυλωνά Κ. Ευαγγελία**

**Επιβλέπων :** Νικόλαος Δουλάμης
Επίκουρος Καθηγητής Ε.Μ.Π.



Αθήνα, Ιούλιος 2018

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΑΓΡΟΝΟΜΩΝ ΚΑΙ ΤΟΠΟΓΡΑΦΩΝ ΜΗΧΑΝΙΚΩΝ

ΤΟΜΕΑΣ ΤΟΠΟΓΡΑΦΙΑΣ

# **“ Τεχνικές και Μέθοδοι Χωρικής Ανάλυσης Ιατρικών Δεδομένων: Περιβαλλοντικοί παράγοντες που επηρεάζουν την Παχυσαρκία και τη Λοίμωξη από *Clostridium difficile* στην πολιτεία της Καλιφόρνια ”**

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

**Ευαγγελία Κ. Μυλωνά**

**Επιβλέπων :** Νικόλαος Δουλάμης
Επίκουρος Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 17η Ιουλίου 2018.

| | | |
|---|---|---|
| ..... | ..... | ..... |
| Νικόλαος Δουλάμης | Μαρίνος Κάβουρας | Αναστάσιος Δουλάμης |
| Επίκουρος Καθηγητής Ε.Μ.Π. | Καθηγητής Ε.Μ.Π. | Επίκουρος Καθηγητής Ε.Μ.Π. |

.....

**Ευαγγελία Κ. Μυλωνά**

Διπλωματούχος Αγρονόμος και Τοπογράφος Μηχανικός Ε.Μ.Π.

# Ευχαριστίες

Θα ήθελα να ευχαριστήσω θερμά τον Επίκουρο Καθηγητή Νικόλαο Δουλάμη, για την ευκαιρία που μου έδωσε να ασχοληθώ με το παρόν θέμα, για τη πολύτιμη βοήθειά του και την εμπιστοσύνη που μου έδειξε κατά την εκπόνηση της διπλωματικής μου εργασίας.

Θα ήθελα να ευχαριστήσω επίσης τον Καθηγητή Ιατρικής του Brown University Δρ. Ελευθέριο Μυλωνάκη, για τις πολύτιμες συμβουλές του σε θέματα ιατρικής καθώς επίσης και τα μέλη του εργαστηρίου του για τον χρόνο που δαπάνησαν όλο αυτό το διάστημα.

Τέλος, θα ήθελα να ευχαριστήσω τον σύντροφό μου για την πολύτιμη στήριξη και βοήθειά του και την οικογένειά μου που με στήριξε όλα αυτά τα χρόνια κατά τη διάρκεια των σπουδών μου.

# Περίληψη

Σκοπός της παρούσας διπλωματικής ήταν η εύρεση κατάλληλων τεχνικών και μεθόδων για την χωρική και στατιστική ανάλυση δεδομένων που αφορούν την έρευνα της Ιατρικής. Συγκεκριμένα οι δύο κατηγορίες που αναλύθηκαν ήταν τα κρούσματα παχυσαρκίας και ο αριθμός ασθενών που παρουσίασε λοίμωξη από το βακτήριο *Clostridium difficile*, στην Καλιφόρνια των ΗΠΑ, το έτος 2015.

Για την ανάλυση της παχυσαρκίας εξετάζεται η επίδραση του επιπέδου εκπαίδευσης και του εισοδήματος των κατοίκων της Καλιφόρνια καθώς και η επιρροή των αθλητικών εγκαταστάσεων που υπάρχουν στην πολιτεία, δεδομένα που εξήχθησαν μέσω του Foursquare API με τη χρήση της γλώσσας Python.

Για την ανάλυση των κρουσμάτων λοίμωξης από το βακτήριο *Clostridium difficile* που σημειώθηκαν στα νοσοκομεία της Καλιφόρνια, εξετάστηκε το ποσοστό επίδρασης των μεταβλητών που αφορούν τη χωρητικότητα των νοσοκομείων, την ύπαρξη μονάδων νοσοκομειακής περίθαλψης ηλικιωμένων και η απόστασή τους από αυτά και τέλος η θερμοκρασία της περιοχής.

Και για τις δύο περιπτώσεις πραγματοποιήθηκε στατιστική και χωρική ανάλυση των δεδομένων με τις μεθόδους παλινδρόμησης OLS και GWR αντίστοιχα. Επίσης, λόγω της φύσης των δεδομένων, χρειάστηκε να γίνουν δοκιμές διαμερισμού της πολιτείας ώστε να επιλεγεί η καταλληλότερη μεθοδολογία. Αποτέλεσμα των αναλύσεων ήταν η ένδειξη της σημαντικότητας μεθόδου GWR σε ιατρικές μελέτες.

**Λέξεις Κλειδιά:** Χωρική Ανάλυση, Γραμμική Παλινδρόμηση, Γεωγραφικά Σταθμισμένη Παλινδρόμηση, HotSpot Ανάλυση, Python, Παχυσαρκία, Λοίμωξη από Clostridium Difficile

## Abstract

The aim of this diploma thesis was to explore and compare techniques and methods for the spatial and statistical analysis of data in the field of medical research. Specifically, the two data categories analyzed were the prevalence of obesity and the prevalence of *Clostridium difficile* infection, in the U.S. state of California in 2015.

In the analysis of obesity, we examined the effects of the population's education level and income, as well as, the influence of sport facilities on the prevalence of obesity in the state of California. Sport facilities' location and check-in count were extracted from the Foursquare API using Python.

In the analysis of the hospital onset *Clostridium difficile* infection prevalence in the state of California, we examined the effects of hospital capacity, the number of long term care facilities and their distance from the healthcare facilities, and the temperature of the study area.

For both cases, a statistical and spatial analysis was performed with OLS and GWR regression methods. Due to the nature of the data, various spatial partitioning schemes were tested, to select the most appropriate methodology. Our analysis indicated the significance of the GWR method in clinical research.



**Keywords:** Spatial Analysis, Linear Regression, Geographically Weighted Regression, HotSpot Analysis, Python, Obesity, Clostridium Difficile Infection

# Περιεχόμενα

# Πίνακας Εικόνων

# Περιεχόμενα Πινάκων

# Κεφάλαιο 1ο - Εισαγωγή

## 1.1 Συστήματα Γεωγραφικών Πληροφοριών

Οι εξελίξεις στην πληροφορική που αφορούν στη διαχείριση μεγάλων βάσεων δεδομένων, και στη συσχέτιση γεωμετρικής και θεματικής πληροφορίας ποικίλης μορφής και προέλευσης, έφεραν τη νέα γενιά ψηφιακών συστημάτων, τα Συστήματα Γεωγραφικών Πληροφοριών (Σ.Γ.Π.), γνωστά διεθνώς ως GIS (Geographic Information System) [1].

“Τα Γεωγραφικά Συστήματα Πληροφοριών είναι μια οργανωμένη συλλογή μηχανικών υπολογιστικών συστημάτων (hardware), λογισμικού (software), χωρικών δεδομένων και ανθρώπινου δυναμικού και έχουν ως σκοπό τη συλλογή, καταχώρηση, ενημέρωση, διαχείριση, ανάλυση και απόδοση κάθε μορφής πληροφορίας που αφορά το γεωγραφικό περιβάλλον.” [2].

Σύμφωνα με τον Burrough “Τα Γεωγραφικά Συστήματα Πληροφοριών αποτελούν ένα ισχυρό σύνολο εργαλείων για την συλλογή, αποθήκευση, ανάληψη ανά πάσα στιγμή, μετασχηματισμό και απεικόνιση χωρικών στοιχείων του πραγματικού κόσμου” [3]. Συνεπώς ένα ΣΓΠ :

- Έχει τη δυνατότητα να αποθηκεύει, να διαχειρίζεται και να ενσωματώνει μεγάλο όγκο χωρικών δεδομένων.
- Αποτελεί το πιο κατάλληλο εργαλείο χωρικής ανάλυσης, εστιαζόμενο ειδικά στη χωρική διάσταση των στοιχείων.
- Αποτελεί ένα πολύ αποτελεσματικό μηχανισμό για την επίλυση χωρικών προβλημάτων μέσα από την οργάνωση, διαχείριση και μετασχηματισμό μεγάλου όγκου στοιχείων, έτσι ώστε η πληροφορία να είναι προσιτή στο σύνολο των χρηστών.

Οι βασικές διαδικασίες για την ολοκλήρωση και την εφαρμογή των ΓΣΠ είναι [2]:

- ο καθορισμός του προβλήματος
- η διαδικασία από στοιχεία σε πληροφορία
- τα συμπεράσματα

Τα συστήματα αυτά υποστηρίζουν πολλές μορφές χωρικών δεδομένων, πολλούς τύπους αρχείων και διάφορα Συστήματα Διαχείρισης Βάσεων Δεδομένων. Στα ΣΓΠ τα δεδομένα χωρίζονται σε δύο κατηγορίες, στα διανυσματικά (vector) και στα κανονικοποιημένα (raster)

δεδομένα. Τα διανυσματικά δεδομένα αναπαριστούν τριών ειδών χωρικές οντότητες με βάση το σχήμα τους :

- Σημεία (Points)
- Γραμμές (Polylines)
- Πολύγωνα (Polygons)

Στα κανονικοποιημένα δεδομένα κάθε ψηφίδα (pixel) αποτελεί ένα τετράγωνο με συγκεκριμένες διαστάσεις που αναπαριστά ένα τμήμα της επιφάνειας της γης. Τέτοια δεδομένα είναι οι δορυφορικές εικόνες ή οι ορθοφωτοχάρτες [4].

## 1.2 Χωρική Ανάλυση

Η στατιστική ανάλυση είναι ένας κλάδος που βοηθά στη μελέτη και κατανόηση φαινομένων ή ιδιοτήτων πολυπληθών ομάδων. Σε περιπτώσεις, όμως, όπου η ανάλυση των φαινομένων εξαρτάται από χαρακτηριστικά χωρικών δεδομένων, όπως η γεωγραφική θέση, δεν αρκεί η κλασική στατιστική, διότι αποκρύπτει σημαντικές πληροφορίες που αφορούν τις χωρικές σχέσεις αυτών. Σε αυτές τις περιπτώσεις υπεισέρχεται η χωρική ανάλυση, η οποία με αντίστοιχες μεθόδους, όπως η τοπική παλινδρόμηση, αναδεικνύει τις χωρικές διεργασίες και σχέσεις, και σε συνδυασμό με άλλες φυσικές και κοινωνικές επιστήμες επιτρέπει την ανάπτυξη μεθόδων ανάλυσης των δεδομένων [5, 6].

Η χωρική παρατήρηση έχει γνωστή θέση στο χώρο και αυτό το χαρακτηριστικό την διαφοροποιεί από οποιαδήποτε άλλη παρατήρηση. Από τη θέση κάθε παρατήρησης, που μπορεί να οριστεί με γεωγραφικές συντεταγμένες σε ένα σύστημα χαρτογραφικής προβολής, απορρέουν κάποιες επιπλέον πληροφορίες, όπως η γειτνίασή της και η απόστασή της από άλλες παρατηρήσεις. Συνεπώς, τα τρία ιδιαίτερα χαρακτηριστικά των χωρικών δεδομένων είναι:

- η θέση
- η γειτνίαση
- η απόσταση

Στα παραπάνω χαρακτηριστικά θα μπορούσε να προστεθεί και ο χρόνος μιας και συχνά η χρονική διάσταση είναι σημαντική στη μελέτη φαινομένων με χωρική διάσταση. Ωστόσο, ο χρόνος αφορά και πολλά μη χωρικά δεδομένα [4].

Σύμφωνα με τον Unwin (1981), χωρική ανάλυση είναι η μελέτη της κατανομής των σημείων, γραμμών περιοχών και επιφανειών ενός χάρτη [7]. Καθώς ο ορισμός αυτός, είναι πολύ γενικός, ο Bailey (1990) ορίζει την ανάλυση χώρου ως μια συνολική δυνατότητα διαχείρισης και μετασχηματισμού των χωρικών στοιχείων σε διαφορετικές μορφές, δίνοντας τους διαφορετική έννοια [8]. Αποτελεί μία ποσοτική ανάλυση ή μελέτη των χωρικών φαινομένων που βρίσκονται στο γεωγραφικό χώρο εισάγοντας έτσι την έννοια της ανάλυσης χωρικών δεδομένων [9]. Κατά τον Κουτσόπουλο (1990) η ανάλυση χώρου είναι επίσης, η διαδικασία μετάβασης από στοιχεία σε πληροφορίες [10].

Η Χωρική Ανάλυση (Spatial Analysis) αποτελείται από ένα σύνολο ποσοτικών μεθόδων και τεχνικών που μελετούν χωρικές οντότητες και φαινόμενα χρησιμοποιώντας τις τοπολογικές, γεωμετρικές, ή γεωγραφικές ιδιότητές τους. Κύριος ρόλος της, είναι η τροφοδότηση της διαδικασίας του χωρικού σχεδιασμού. Σε αυτό το πλαίσιο, το ενδιαφέρον των επιστημόνων, που μελετούν το χώρο, εστιάζεται στην ανάλυση και στον προσδιορισμό της αντιστοιχίας και συσχέτισης μεταξύ δύο ή περισσότερων χωρικών κατανομών.

Σήμερα, η σύγχρονη χωρική ανάλυση, αξιοποιώντας την πρόοδο της Τεχνολογίας των Πληροφοριακών Συστημάτων (information Systems), εστιάζει στις βασισμένες σε υπολογιστή τεχνικές κυρίως λόγω του μεγάλου όγκου των δεδομένων, της πολυπλοκότητας της αναλυτικής επεξεργασίας, αλλά και των αυξημένων δυνατοτήτων των Γεωγραφικών Συστημάτων Πληροφοριών (Geographic Information Systems). Διατυπώνει και ελέγχει υποθέσεις γύρω από τη μαθηματική σχέση ή τους μηχανισμούς που προξενούν την αντιστοιχία που μελετάται, ενώ σε άλλες περιπτώσεις, η ανάλυση είναι ανιχνευτική και αναζητά επαγωγικές γενικεύσεις για τη συμμεταβλητότητα των προτύπων [11].

Σύμφωνα με τους Fotheringham et al. (2000) [12] η ποσοτική γεωγραφία περιλαμβάνει μία ή περισσότερες από τις παρακάτω δραστηριότητες:

- την ανάλυση αριθμητικών χωρικών δεδομένων
- την ανάπτυξη χωρικής θεωρίας
- τον ορισμό και έλεγχο μαθηματικών μοντέλων χωρικών διεργασιών.

## 1.3 Σκοπός της Διπλωματικής

Σκοπός της παρούσας διπλωματικής ήταν η εύρεση κατάλληλων τεχνικών και μεθόδων για την χωρική και στατιστική ανάλυση δεδομένων που αφορούν την έρευνα της Ιατρικής. Συγκεκριμένα οι δύο κατηγορίες που αναλύθηκαν ήταν τα κρούσματα παχυσαρκίας και ο αριθμός ασθενών που παρουσίασε λοίμωξη από το βακτήριο Difficile Infection, στην Καλιφόρνια των ΗΠΑ, το έτος 2015.

Αρχικά έγινε μία ανασκόπηση των διαθέσιμων τεχνικών και εργαλείων που προσφέρονται για τέτοιου είδους αναλύσεις από το ArcMap 10.4.1 και στη συνέχεια πραγματοποιήθηκε η συλλογή των δεδομένων από επίσημους κυβερνητικούς οργανισμούς. Μεγάλο μέρος της εργασίας αποτελεί και η δημιουργία δεδομένων από διαθέσιμες πηγές. Η δυνατότητα που προσφέρουν οι ιστοσελίδες για εξαγωγή στοιχείων, μέσω κατάλληλης γλώσσας προγραμματισμού, αποτελεί σημαντικό εργαλείο στα χέρια ενός Τοπογράφου Μηχανικού.

Τέλος, πραγματοποιήθηκαν διάφορες δοκιμές στατιστικής και χωρικής ανάλυσης των μεταβλητών προς εξέταση. Στο σημείο αυτό κρίθηκε εμφανής η σημασία και η σπουδαιότητα της Γεωγραφικά Σταθμισμένης Παλινδρόμησης σε ιατρικές αναλύσεις που εστιάζουν στην επιρροή περιβαλλοντικών παραγόντων.

### 1.3.1 Η χρήση των ΣΓΠ στην Δημόσια Υγεία

Το GIS στις ιατρικές εφαρμογές έχει την βάση του στην ιατρική γεωγραφία, η οποία βρίσκεται στη βιβλιογραφία αρκετών αρχαίων πολιτισμών, συμπεριλαμβανομένης της Κίνας, της Ελλάδας και της Ινδίας, με ίσως την πρώτη εργασία του ιατρού Ιπποκράτη τον 5ο αιώνα π.Χ., που ήταν από τους πρώτους που παρακολούθησαν τις σχέσεις μεταξύ της ανθρώπινης υγείας και του περιβάλλοντος [13].

Τα τελευταία χρόνια παρατηρείται η ανάπτυξη εφαρμογών συστημάτων γεωγραφικών πληροφοριών (ΣΓΠ) στη δημόσια υγεία. Η βιβλιογραφία για το GIS και την ιατρική γεωγραφία έχει επικεντρωθεί κυρίως στο πως μπορούν τα ΣΓΠ να εφαρμοστούν ως εργαλεία ανάλυσης και απεικόνισης για να εξεταστούν οι πτυχές των ασθενειών και των υπηρεσιών υγείας [14].

Στην πιο βασική του χρήση στην δημόσια υγεία, ένα ΣΓΠ απαντά σε ερωτήματα θέσης. Αυτό μπορεί να σημαίνει παραδείγματος χάρη "Που ξεκινούν οι ασθένειες;" ή "Που ζουν οι άνθρωποι;". Επίσης απαντούν σε ερωτήματα προτύπων, όπως "Που εντοπίζονται υψηλές συγκεντρώσεις κρουσμάτων;". Η συχνότητα εμφάνισης ενός φαινομένου και η επικράτηση μίας τιμής αποτελούν σημαντικά θεμέλια για τα ΣΓΠ σε εφαρμογές υγείας. Η συχνότητα συμβάλει στην εξέταση νέων περιπτώσεων ασθένειας, ενώ η επικράτηση είναι η επίπτωση των υφιστάμενων περιπτώσεων μιας νόσου σε μία συγκεκριμένη χρονική περίοδο [15]. Επιπλέον τα ΣΓΠ απαντούν σε ερωτήματα τάσεων - σε ποιες περιοχές παρατηρείται αύξηση ή μείωση ενός φαινομένου σε ένα χρονικό διάστημα- , συνθηκών - ποιες περιοχές ικανοποιούν τις εκάστοτε συνθήκες- και επιπτώσεων.

Με τη χρήση του λογισμικού ArcGIS επιτυγχάνεται η καλύτερη κατανόηση των τομέων αναγκών. Μέσω του λογισμικού είναι δυνατή η αντιμετώπιση προβλημάτων δημόσιας υγείας με τη βοήθεια έξυπνων χαρτών και χωρικής ανάλυσης. Επίσης, δίνεται η δυνατότητα διαχείρισης πανδημιών, πρόληψης χρόνιων ασθενειών, καθώς και η δυνατότητα ανάλυσης διάφορων περιβαλλοντικών και κοινωνικών παραγόντων που συμβάλλουν στην εξάπλωση μιας νόσου [16].

Στα πλεονεκτήματα των ΣΓΠ στη δημόσια υγεία συμπεριλαμβάνονται [17]:

- Η χρήση γεωγραφίας και GIS ως κοινή αναλυτική πλατφόρμα ξεπερνά την απλή θεματική χαρτογραφία
- Η ποιότητα των δεδομένων (μοναδικότητα, πληρότητα, συνέπεια, ακρίβεια, εγκυρότητα, επικαιρότητα) υποστηρίζεται σε διάφορες πηγές
- Η μείωση της αλληλεπικάλυψης των προσπαθειών καταχώρησης των δεδομένων, το οποίο επιφέρει και μείωση κόστους διαχείρισης των δεδομένων

- Επιτρέπεται η ταυτόχρονη αντιμετώπιση πολλαπλών θεμάτων δημόσιας υγείας από πιο συστηματική άποψη
- Η δυνατότητα χρήσης της δύναμης των χαρτών για τον προγραμματισμό και τη λήψη αποφάσεων

# Κεφάλαιο 2ο - Τεχνικές Ανάλυσης

## 2.1 Χωρική Αυτοσυσχέτιση

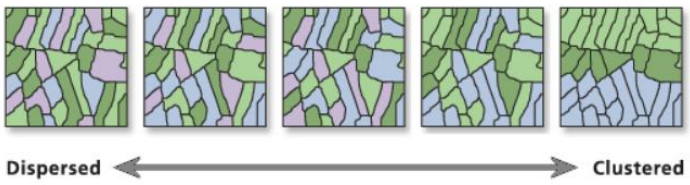
Στη χωρική στατιστική τα δεδομένα προς ανάλυση δεν είναι ανεξάρτητα μεταξύ τους, όπως απαιτείται στην κλασική στατιστική. Το ενδιαφέρον του κλάδου αυτού, είναι η κατανόηση και η επεξήγηση της χωρικής συμπεριφοράς των μεταβλητών που αντιπροσωπεύουν κάποιο φαινόμενο.

Οι δυνατότητες οπτικοποίησης και η γρήγορη ανάκτηση δεδομένων μέσω των ΣΓΠ, έχουν δημιουργήσει την ανάγκη για νέες τεχνικές διερευνητικής ανάλυσης δεδομένων, οι οποίες επικεντρώνονται στις χωρικές πτυχές των δεδομένων. Η αναγνώριση των προτύπων χωρικής συσχέτισης αποτελεί ένα σημαντικό μέλημα [18].

Χωρική αυτοσυσχέτιση είναι η συσχέτιση μεταξύ των τιμών μιας μεταβλητής που οφείλεται αυστηρά στην εγγύτητα των τιμών αυτών στο γεωγραφικό χώρο, εισάγοντας μια απόκλιση από την υπόθεση ανεξάρτητων παρατηρήσεων της κλασικής στατιστικής [19].

Ως χωρική αυτοσυσχέτιση μπορεί να ορισθεί η ύπαρξη ομοιότητας ενός αντικειμένου με τα γειτονικά του αντικείμενα στο χώρο. Στα χωρικά πρότυπα όπου οι γειτονικές παρατηρήσεις τείνουν να έχουν παρόμοια υψηλές ή χαμηλές τιμές υποδηλώνουν θετική χωρική αυτοσυσχέτιση, ενώ όταν οι υψηλές και οι χαμηλές τιμές εναλλάσσονται μεταξύ παρακείμενων περιοχών, η χωρική αυτοσυσχέτιση είναι αρνητική. Όταν η χωρική αυτοσυσχέτιση δεν έχει στατιστική σημαντικότητα τότε η χωρική κατανομή είναι τυχαία.

Στην Εικόνα 1, φαίνεται η κατανομή των χωρικών προτύπων κατά την οποία η διασπορά σχηματίζει είτε ομαδοποιήσεις (clustered) στο χώρο, είτε διασκορπισμένα πρότυπα (dispersed). Ένα πρότυπο που παρουσιάζεται σε ενδιάμεσο σημείο των δύο παραπάνω άκρων, αποτελεί τυχαίο (random) πρότυπο κατανομής.




*Εικόνα 1 : Χωρικά πρότυπα [20]*

Οι χωρικές αυτοσυσχέτισεις ασχολούνται ταυτόχρονα τόσο με τη θέση όσο και με ποσοτικά χαρακτηριστικά. Για να προσδιοριστεί εάν οι τιμές σε μία χαρτογράφηση αποκλίνουν σημαντικά από ένα πρότυπο στο οποίο οι τιμές εκχωρούνται τυχαία, απαιτείται κάποιο είδος δείκτη σύγκρισης [21]. Οι δείκτες μέτρησης χωρικής αυτοσυσχέτισης μπορεί να είναι:

- Γενικοί, όπως ο Global Moran's I και ο Getis and Ord G και G*, οι οποίοι αναγνωρίζουν χωρικά πρότυπα και τάσεις
- Τοπικοί, όπως ο Local Moran's I και ο Getis and Ord Gi και Gi*, οι οποίοι προσδιορίζουν τη θέση και το μέγεθος των χωρικών ομάδων.

### 2.1.1 Δείκτης μέτρησης χωρικής αυτοσυσχέτισης Moran's I

Ο ολικός δείκτης Moran's I (Global Moran's I) αποτελεί έναν από τους παλαιότερους και πιο ευρέως χρησιμοποιούμενους δείκτες για την εξέταση ύπαρξης χωρικής αυτοσυσχέτισης των ποσοτικών δεδομένων μιας μεταβλητής. Ο δείκτης αυτός βασίζεται στον πρώτο ορισμό του δείκτη Moran's (1948) [22],αλλά και στον συντελεστή συσχέτισης Pearson και δίνεται από τη σχέση:

$$I = \frac{n}{W}\frac{\sum_i^n \sum_j^n w_{ij} z_i z_j}{\sum_i^n {z_i}^2} \qquad (2.1)$$

Η Σχέση (2.1) προτάθηκε από τους Cliff και Ord (1973, 1981) [23, 24] και υπολογίζεται η διαφορά $z_i$ του μέσου όρου όλων των τιμών μιας μεταβλητής από την τιμή κάθε χωρικού στοιχείου ($x_i - \bar{x}$) καθώς και η διαφορά $z_j$ του μέσου όρου από την τιμή κάθε γείτονα j του στοιχείου ($x_j - \bar{x}$). Με τον τρόπο αυτό συγκρίνονται οι διαφορές μεταξύ τους, ενώ με βάση τη χωρική εγγύτητα μεταξύ των παρατηρήσεων υπεισέρχονται τα βάρη $w_{ij}$. Το n είναι ο συνολικός αριθμός των στοιχείων και το W το άθροισμα των στοιχείων του πίνακα βαρών.

Οι τιμές που λαμβάνει ο ολικός δείκτης Moran's I είναι από -1 έως +1, όπου τιμές κοντά στο +1 υποδηλώνουν ισχυρή θετική χωρική συσχέτιση (Clustered), τιμές κοντά στο -1 υποδηλώνουν ισχυρή αρνητική χωρική αυτοσυσχέτιση (Dispersed) και τιμές κοντά στο 0 υποδηλώνουν απουσία χωρικής αυτοσυσχέτισης (Random).

Ο υπολογισμός των βαρών γίνεται είτε με βάση τη φυσική γειτνίαση μεταξύ των παρατηρήσεων, είτε με βάση την ευθεία απόσταση, είτε με βάση τον αριθμό των κοντινότερων γειτόνων. Σε οποιαδήποτε περίπτωση τα βάρη μπορεί να είναι δυαδικά, όπου την τιμή 1 λαμβάνουν οι γείτονες και την τιμή 0 οι μη γείτονες, κανονικοποιημένα, τα οποία ορίζονται με διαίρεση κάθε βάρους με το άθροισμα όλων των βαρών, ή να υπολογίζονται συναρτήσει της αντίστροφης απόστασης μεταξύ των παρατηρήσεων [4].

Για την εξέταση τοπικών μεταβολών μέσα σε σχηματισμούς εξάρτησης, ο Anselin (1995), πρότεινε τον τοπικό δείκτη Moran's I [18]:

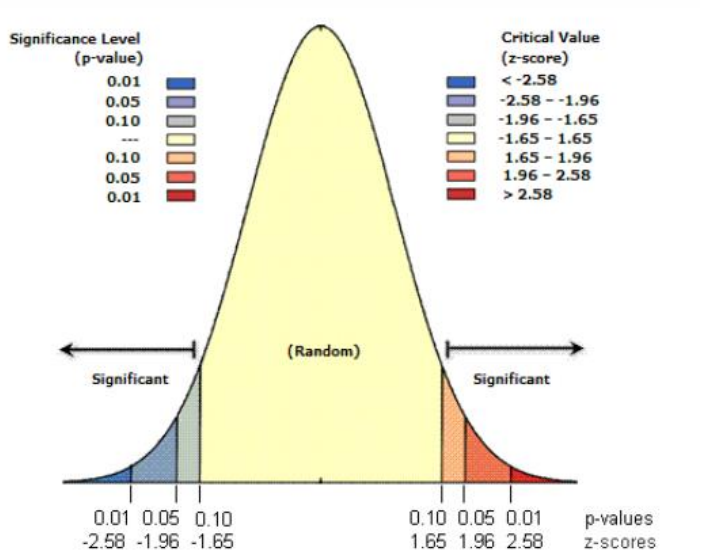
$$I = \frac{z_i}{S_i^2} \sum_{j=1, j\neq i}^{n} w_{ij} z_j \qquad (2.2)$$

όπου $S_i^2 = \frac{1}{n-1} \sum_{j=1, j\neq i}^{n} z_j^2$ [25], το οποίο παραμένει σταθερό για όλα τα τοπικά $I_i$ μιας μεταβλητής. Τα $z_i = (x_i - \bar{x})$ και $z_j = (x_j - \bar{x})$ είναι οι μέσοι όροι των τιμών μιας μεταβλητής από την τιμή κάθε χωρικού στοιχείου και από την τιμή κάθε γείτονα j του στοιχείου, αντίστοιχα, ενώ το $w_{ij}$ αποτελεί τον συντελεστή βάρους και είναι ανάλογος της απόστασης μεταξύ των θέσεων i και j.

Ο τοπικός δείκτης Moran's I μπορεί να έχει τιμές μεγαλύτερες του +1 και μικρότερες του -1, σε αντίθεση με τον ολικό, ωστόσο η ερμηνεία τους είναι όμοια. Η θετική τιμή του τοπικού $I_i$ υποδεικνύει χωρική συγκέντρωση παρόμοιων υψηλών ή χαμηλών τιμών, ενώ η αρνητική υποδεικνύει συγκέντρωση ανόμοιων τιμών, όπου εμφανίζονται οι σημαντικές διαφορές των τιμών γειτονικών σημείων.

Οι περισσότερες στατιστικές δοκιμές ξεκινούν με τον προσδιορισμό μιας μηδενικής υπόθεσης. Πραγματοποιείται έλεγχος με αρχική υπόθεση $H_0=0$ δηλαδή ο δείκτης Moran's I να είναι 0, που σημαίνει ότι τα δεδομένα δεν ακολουθούν κάποιο χωρικό πρότυπο αλλά είναι τυχαία κατανεμημένα στο χώρο, έναντι της αμφίδρομης εναλλακτικής υπόθεσης $H_1\neq0$ και συνεπώς ο

δείκτης Moran's I είναι διάφορος του 0, γεγονός που καταδεικνύει ότι τα δεδομένα εμφανίζουν συγκεκριμένη δομή και κατανομή. Εάν τα δεδομένα ακολουθούν κανονική κατανομή, τότε ο τοπικός δείκτης Moran's I μπορεί να τυποποιηθεί σε ένα z-score. Το στατιστικό Z αποτελεί την τυπική απόκλιση και ως κρίσιμη τιμή ορίζεται το 1,96 με επίπεδο σημαντικότητας p ίσο με 0,05, δηλαδή 95% βεβαιότητα. Όταν το Z έχει τιμή μεγαλύτερη από την κρίσιμη +1,96 ή μικρότερη από -1,96, τότε η μηδενική υπόθεση $H_0$ απορρίπτεται και επομένως ο δείκτης Moran's I μπορεί να θεωρηθεί στατιστικά σημαντικός και το εξεταζόμενο σημείο ομαδοποιείται με τα γειτονικά του σε επίπεδο σημαντικότητας 0,05.




*Εικόνα 2: Spatial Autocorrelation (Global Moran's I) [20]*

### 2.1.2 Δείκτης μέτρησης χωρικής αυτοσυσχέτισης Getis - Ord G

Οι Getis και Ord (1992) πρότειναν μία ομάδα μέτρων χωρικής αυτοσυσχέτισης που συμβολίζονται με το λατινικό γράμμα G. Αυτά τα στατιστικά στοιχεία έχουν ορισμένα χαρακτηριστικά που επιτρέπουν τη μέτρηση χωρικής εξάρτησης σε μεταβλητές χωρικών δεδομένων. Ο ολικός δείκτης G βασίζεται σε όλα τα ζεύγη τιμών ($x_i$, $x_j$) έτσι ώστε τα i και j να βρίσκονται σε απόσταση d μεταξύ τους και ορίζεται από τη σχέση (2.3) [26] :

$$G(d) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(d)x_i x_j}{\sum_{i=1}^{n}\sum_{j=1}^{n} x_i x_j}, \; j \neq l \qquad (2.3)$$

όπου, $w_{ij}$ είναι το βάρος για κάθε σημείο j της γειτονιάς εκτός από το ίδιο το i (το βάρος παίρνει την τιμή 1 για σημεία εντός γειτονιάς και την τιμή 0 για τα υπόλοιπα σημεία) και $x_i$, $x_j$ είναι οι τιμές της μεταβλητής X στα σημεία i, j αντίστοιχα.

Ο τοπικός δείκτης Gi ύστερα από αναθεώρηση των Ord και Getis (1995) [27] δίνεται από την παρακάτω σχέση (2.4), όπου λαμβάνονται υπόψη τα βάρη και οι στατιστικές που σχετίζονται με την αυτοσυσχέτιση Moran's I:

$$G_i(d) = \frac{\sum_{j=1}^{n} w_{ij}(d)x_j - \bar{x}\sum_{j=1}^{n} w_{ij}(d)}{s\sqrt{\frac{(n-1)\sum_{j=1}^{n} w_{ij}^2(d) - [\sum_{j=1}^{n} w_{ij}(d)]^2}{n-2}}}, \text{ j≠I} \quad (2.4)$$

όπου, $x_i$, $x_j$ είναι οι τιμές της μεταβλητής X στα σημεία i, j αντίστοιχα, $\bar{x}$ είναι ο μέσος της μεταβλητής και s η τυπική απόκλιση.

Ο ολικός και ο τοπικός δείκτης G στον οποίο η παρατήρηση στο σημείο *i* συμπεριλαμβάνεται στον υπολογισμό του δείκτη συμβολίζονται ως G*(d) και G*i(d), αντίστοιχα και ορίζονται ως εξής:

$$G^*(d) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(d)x_i x_j}{\sum_{i=1}^{n}\sum_{j=1}^{n} x_i x_j} \quad (2.5)$$

$$G_i^*(d) = \frac{\sum_{j=1}^{n} w_{ij}(d)x_j - \bar{x}\sum_{j=1}^{n} w_{ij}(d)}{s\sqrt{\frac{n\sum_{j=1}^{n} w_{ij}^2(d) - [\sum_{j=1}^{n} w_{ij}(d)]^2}{n-1}}} \quad (2.6)$$

όπου, $x_j$ είναι η τιμή του στοιχείου j , , $w_{ij}$ είναι το βάρος μεταξύ των χαρακτηριστικών i και j, n το πλήθος των παρατηρήσεων και $\bar{x}$ είναι ο μέσος της μεταβλητής και s η τυπική απόκλιση:

$$\bar{x} = \frac{\sum_{j=1}^{n} x_j}{n} \quad (2.7)$$

$$s = \sqrt{\frac{\sum_{j=1}^{n} x_j^2}{n} - (\bar{x})^2} \quad (2.8)$$

Ο ολικός δείκτης Getis – Ord G ή G* παρέχει μία ένδειξη ύπαρξης ή απουσίας χωρικών προτύπων, ενώ για την ανίχνευση της θέσης των χωρικών προτύπων υψηλών ή χαμηλών τιμών

μιας μεταβλητής πρέπει να υπολογιστούν οι τοπικοί δείκτες Gi ή Gi* για κάθε σημείο i. Οι προκύπτουσες τιμές z και p δείχνουν που υπάρχει χαρακτηριστική χωρική συγκέντρωση. Ένα χαρακτηριστικό μπορεί να μην είναι στατιστικά σημαντικό Hot spot παρότι έχει υψηλή τιμή. Για να θεωρηθεί στατιστικά σημαντικό πρέπει η θετική τιμή του Gi του χαρακτηριστικού με υψηλή τιμή να περιβάλλεται από χαρακτηριστικά με εξίσου υψηλές τιμές. Μια αρνητική τιμή του Gi που είναι στατιστικά σημαντική δηλώνει ότι γύρω από το σημείο i σε απόσταση d υπάρχει μια γειτονιά χαμηλών τιμών (cold spots). Γίνεται η υπόθεση ότι το σύνολο των τιμών $x_i$ εντός της γειτονιάς d είναι ένα τυχαίο δείγμα των τιμών της μεταβλητής X και ακολουθεί κανονική κατανομή (Getis and Ord, 1992). Το τοπικό άθροισμα για ένα χαρακτηριστικό και τους γείτονές του συγκρίνεται αναλογικά με το άθροισμα όλων των χαρακτηριστικών. Όταν το τοπικό άθροισμα είναι πολύ διαφορετικό από το αναμενόμενο και όταν αυτή η διαφορά είναι πολύ μεγάλη για να είναι το αποτέλεσμα τυχαίας πιθανότητας, προκύπτει ένα στατιστικά σημαντικό z-score.

### 2.1.3 Hot Spot Analysis

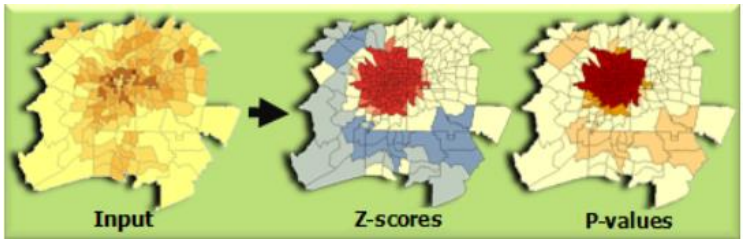
Σε πληθώρα προβλημάτων προκύπτει η ανάγκη εύρεσης περιοχών που παρουσιάζουν υψηλή χωρική συγκέντρωση στοιχείων με υψηλές ή χαμηλές τιμές. Στην παρούσα εργασία υπήρξε η ανάγκη προσδιορισμού των ασθενειών στο χώρο.

Προτού πραγματοποιηθεί η ανίχνευση θερμών σημείων (hot spots) και ψυχρών σημείων (cold spots), ελέγχεται αν τα δεδομένα είναι ομαδοποιημένα με κάποια τεχνική χωρικής αυτοσυσχέτισης [28]. Στο λογισμικό ArcGIS η διαδικασία αυτή πραγματοποιείται με τη μέθοδο Global Moran’s I (Spatial autocorrelation) ή με τη μέθοδο Getis-Ord General G (High/Low Clustering). Η ανάλυση Hotspot πραγματοποιείται με τη μέθοδο Getis – Ord Gi* (Hot Spot Analysis) είτε με τη μέθοδο Anselin Local Moran’s I (Cluster and Outlier Analysis).

Για την εκτέλεση των παραπάνω μεθόδων θα πρέπει να καθορισθεί η σχέση χωρικής συγγένειας των χαρακτηριστικών (Conceptualization of Spatial Relationship). Στο στάδιο αυτό προσδιορίζεται το ποιος θεωρείται γείτονας και καθορίζεται το βάρος του κάθε χαρακτηριστικού. Στο ArcGIS διατίθενται οι εξής μέθοδοι προσδιορισμού αυτών των σχέσεων [29]:

- Inverse Distance: Τα γειτονικά χαρακτηριστικά έχουν μεγαλύτερη επιρροή στους υπολογισμούς για έναν χαρακτηριστικό στόχο από ότι τα απομακρυσμένα. Επί της ουσίας με τη χρήση της ανάστροφης απόστασης η επιρροή ενός σημείου σε ένα άλλο φθίνει κατά την απομάκρυνση του.
- Inverse Distance Square: η κλίση είναι εντονότερη από την προηγούμενη μέθοδο και η σημαντική επιρροή στους υπολογισμούς είναι από τους κοντινούς γείτονες του χαρακτηριστικού στόχου.
- Fixed Distance Band: Προσδιορίζεται μία κρίσιμη απόσταση και στους υπολογισμούς λαμβάνουν βάρος και ασκούν επιρροή τα χαρακτηριστικά τα οποία βρίσκονται εντός αυτής. Τα χαρακτηριστικά εκτός της κρίσιμης απόστασης λαμβάνουν μηδενικό βάρος και δεν επηρεάζουν τους υπολογισμούς.
- Zone of Indifference: Τα χαρακτηριστικά τα οποία βρίσκονται εντός της καθορισμένης κρίσιμης απόστασης λαμβάνουν μοναδιαίο βάρος, ενώ εκείνα τα οποία βρίσκονται εκτός το βάρος τους μειώνεται ανάλογα με την απόσταση από το χαρακτηριστικό στόχο.
- Contiguity Edges only: Μόνο οι γειτονικές ιδιότητες των πολυγώνων που μοιράζονται ένα όριο ή αλληλεπικαλύπτονται θα επηρεάσουν τους υπολογισμούς για το χαρακτηριστικό πολυγώνου στόχο.
- Contiguity Edges Corners: Τα χαρακτηριστικά των πολυγώνων που μοιράζονται ένα όριο, μοιράζονται έναν κόμβο ή αλληλεπικαλύπτονται, επηρεάζουν τους υπολογισμούς για το χαρακτηριστικό πολυγώνου στόχο.
- Get Spatial Weights from File: Οι χωρικές σχέσεις καθορίζονται από συγκεκριμένο αρχείο χωρικών βαρών.

Οι μέθοδοι υπολογισμού των αποστάσεων κάθε χαρακτηριστικού από τα γειτονικά του πραγματοποιείται είτε με Ευκλείδεια Απόσταση, όπου προσδιορίζεται η ευθεία απόσταση μεταξύ δύο σημείων, είτε με την Μανχάταν Απόσταση, όπου η απόσταση μεταξύ δύο σημείων μετράται κατά μήκος ορθογώνιων αξόνων και υπολογίζεται το άθροισμα της απόλυτης διαφοράς μεταξύ των συντεταγμένων x και y.




*Εικόνα 3: Hot Spot Analysis (Getis-Ord Gi*) [29]*

## 2.2 Μέθοδοι Παλινδρόμησης

Η παλινδρόμηση χρησιμοποιούνταν και χρησιμοποιείται για δύο κυρίως λόγους, την πρόβλεψη και την ερμηνεία φαινομένων. Ο λόγος λοιπόν για τον οποίο χρησιμοποιείται κάθε φορά η συγκεκριμένη μέθοδος στην ανάλυση, είναι αυτός που καθορίζει και τη μεθοδολογία που ακολουθείται [30].

Με την παλινδρόμηση εξετάζεται ουσιαστικά, η σχέση αιτίας – αποτελέσματος μεταξύ ενός φαινομένου και των παραγόντων που το επηρεάζουν. Βάσει μιας εξαρτημένης μεταβλητής ελέγχεται η επιρροή ανεξάρτητων μεταβλητών του υπό εξέταση φαινομένου. Τα αποτελέσματα της παλινδρόμησης επιτρέπουν την εξαγωγή χρήσιμων συμπερασμάτων για τους παράγοντες που ενδεχομένως επηρεάζουν ένα φαινόμενο, οδηγώντας στην βέλτιστη κατανόηση του.

Ένα στατιστικό μοντέλο βαθμονομείται με κάποια από τις παρακάτω τεχνικές παλινδρόμησης ανάλογα με τη φύση του φαινομένου που μελετάται [4]:

- Γραμμική, όταν η εξαρτημένη μεταβλητή είναι λόγος (ratio) και ακολουθεί κανονική κατανομή
- Poisson, όταν η εξαρτημένη μεταβλητή αφορά αριθμό (counts) σπάνιων συμβάντων και ακολουθεί κατανομή Poisson
- Λογιστική, όταν η εξαρτημένη μεταβλητή λαμβάνει δύο τιμές (yes/no) και ακολουθεί διωνυμική κατανομή.

Η γραμμική παλινδρόμηση (linear regression) είναι μία από τις κύριες μεθόδους χωρικής ανάλυσης. Για τη βαθμονόμηση ενός μοντέλου μπορούν να χρησιμοποιηθούν διαφορετικές τεχνικές γραμμικής παλινδρόμησης, ωστόσο στη παρούσα εργασία θα αναλυθούν η απλή και πολλαπλή γραμμική παλινδρόμηση και η γεωγραφικά σταθμισμένη παλινδρόμηση.

### 2.2.1 Γραμμική Παλινδρόμηση

Στη περίπτωση δύο φαινομένων Χ και Υ, τα οποία συσχετίζονται μεταξύ τους είναι δυνατόν να υπολογιστεί η μεταξύ τους συνδιακύμανση. Στην περίπτωση εξέτασης δύο μεταβλητών, μίας εξαρτημένης και μίας ανεξάρτητης, αναφερόμαστε σε απλή γραμμική παλινδρόμηση, ενώ στην εξέταση συσχέτισης περισσότερων από δύο ανεξάρτητων μεταβλητών αναφερόμαστε σε πολλαπλή γραμμική παλινδρόμηση.

Στα πλαίσια της απλής γραμμικής παλινδρόμησης στόχος είναι ο προσδιορισμός της γραμμικής εξίσωσης:

$$Y = b_0 + b_1 X \quad (2.9)$$

όπου, Υ και Χ είναι η εξαρτημένη και η ανεξάρτητη μεταβλητή αντίστοιχα, $b_0$ είναι σταθερός όρος και $b_1$ είναι η παράμετρος η οποία αφορά τη σχέση μεταξύ των δύο μεταβλητών (συντελεστής παλινδρόμησης). Επειδή όμως το μοντέλο είναι στοχαστικό και όχι ντετερμινιστικό στις εκτιμώμενες τιμές της εξαρτημένης μεταβλητής θα πρέπει να συνυπολογισθούν και τα σφάλματα. Η ζητούμενη εξίσωση γίνεται της μορφής:

$$\widehat{Y_i} = \widehat{b_0} + \widehat{b_1} X_i \quad (2.10)$$

όπου, $\widehat{Y_i}$ είναι η εκτιμώμενη τιμή του Υ για οποιαδήποτε τιμή του Χ.

Ένας από τους πιο προσφιλείς τρόπους εκτίμησης των παραμέτρων $b_0$ και $b_1$ είναι η Μέθοδος Ελαχίστων Τετραγώνων (Ordinary Least Squares – OLS), σύμφωνα με την οποία ελαχιστοποιείται το άθροισμα των τετραγώνων των διαφορών μεταξύ των παρατηρήσεων και των εκτιμώμενων τιμών της εξαρτημένης μεταβλητής. Τα $b_0$ και $b_1$ δίνονται από τις σχέσεις (2.11) και (2.12) [31]:

$$\widehat{b_1} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \quad (2.11)$$

$$\widehat{b_0} = \frac{\sum_{i=1}^{n} y_i}{n} - b_1 \frac{\sum_{i=1}^{n} x_i}{n} \quad (2.12)$$

Αν το μοντέλο περιλαμβάνει αρκετές ανεξάρτητες μεταβλητές, τότε η εξίσωση πολλαπλής παλινδρόμησης με τη μέθοδο ελαχίστων τετραγώνων ορίζεται ως εξής :

$$\widehat{y_i} = \widehat{b_0} + \sum_{k=1}^{m} \widehat{b_k}\, x_{ki} \quad (2.13)$$

όπου m είναι ο αριθμός των ανεξάρτητων μεταβλητών. Σύμφωνα με τις παραδοχές του μοντέλου παλινδρόμησης, τα σφάλματα θα πρέπει να είναι γραμμικά, με μηδενικό μέσο όρο και σταθερή διακύμανση, να ακολουθούν κανονική κατανομή και να παίρνουν ανεξάρτητες μεταξύ τους τιμές οι οποίες θα είναι πρέπει να είναι ανεξάρτητες και από τις τιμές των x. Αντίστοιχα η εξαρτημένη μεταβλητή πρέπει να αποτελεί γραμμική συνάρτηση της ανεξάρτητης ή των ανεξάρτητων μεταβλητών, να έχει επίσης σταθερή διακύμανση και να ακολουθεί κανονική κατανομή. Επειδή στην πράξη η διακύμανση του σφάλματος δεν είναι γνωστή, ορίζονται διαστήματα εμπιστοσύνης ώστε να ελεγχθούν οι υποθέσεις για το αν οι εκτιμήσεις των παραμέτρων $b_0$ και $b_1$ είναι στατιστικά σημαντικές.

Η ακρίβεια της προσαρμογής του μοντέλου ελέγχεται από τον συντελεστή προσδιορισμού $R^2$. Ο προσδιοριστικός συντελεστής $R^2$ εκφράζει το ποσοστό της ολικής μεταβλητότητας της Y που εξηγείται από τη συνδυασμένη επίδραση όλων των μεταβλητών $X_i$ ,που συμμετέχουν στην περιγραφή της εξίσωσης της γραμμικής παλινδρόμησης επί της εξαρτημένης Y. Ο συντελεστής $R^2$ λαμβάνει τιμές από 0, όπου δεν προκύπτει κανένα ποσοστό προσαρμοστικότητας, έως 1 όπου παρατηρείται άριστη προσαρμοστικότητα. Επειδή όμως ο συντελεστής αυτός δεν είναι απόλυτα αξιόπιστος όταν το πλήθος των παρατηρήσεων είναι μικρό, αντικαθίσταται από το προσαρμοσμένο συντελεστή Adjusted $R^2$, όπου συνυπολογίζονται οι βαθμοί ελευθερίας k του μοντέλου, και η τιμή του είναι πάντοτε μικρότερη από το $R^2$ [32].

Στο λογισμικό ArcGIS ένα χρήσιμο εργαλείο, για την ανάλυση γραμμικής παλινδρόμησης με τη μέθοδο ελαχίστων τετραγώνων, είναι το Exploratory Regression, το οποίο αξιολογεί όλους τους πιθανούς συνδυασμούς των υποψήφιων επεξηγηματικών μεταβλητών, αναζητώντας εκείνα τα μοντέλα OLS, τα οποία εξηγούν καλύτερα την εξαρτημένη μεταβλητή στο πλαίσιο των καθορισμένων από το χρήστη κριτηρίων.

Το εργαλείο αυτό, εντοπίζει τα μοντέλα τα οποία πληρούν τα κριτήρια κατωφλίου για το ελάχιστο αποδεκτό Adjusted $R^2$ , το μέγιστο συντελεστή p-value, το μέγιστο VIF Value και το ελάχιστο αποδεκτό p-value Jarque-Bera. Στη περίπτωση που εντοπίσει ένα τέτοιο μοντέλο,

ελέγχει τη χωρική αυτοσυσχέτιση με τον ολικό δείκτη Moran's I. Στη συνέχεια επιλέγει τα τρία μοντέλα με το υψηλότερο Adjusted $R^2$ και μέγιστη τιμή Jarque-Bera.

Ο συντελεστής p-value αντιπροσωπεύει το επίπεδο εμπιστοσύνης και οι τιμές του κυμαίνονται από 1.0 έως 0.0. Η μικρότερη τιμή υποδηλώνει και ένα ισχυρότερο επίπεδο εμπιστοσύνης. Η προεπιλεγόμενη τιμή είναι 0.05, υποδεικνύοντας ότι τα μοντέλα τα οποία θα θεωρηθούν αποδεκτά, περιέχουν επεξηγηματικές μεταβλητές των οποίων οι συντελεστές βρίσκονται σε επίπεδο εμπιστοσύνης 95%. Η τιμή VIF αντικατοπτρίζει τη πλεοναστικότητα των επεξηγηματικών μεταβλητών. Όταν η τιμή VIF είναι μεγαλύτερη από 7.5 τότε το μοντέλο είναι ασταθές. Τέλος η τιμή Jarque-Bera υποδεικνύει αν τα αποτελέσματα κατανέμονται κανονικά. Μια μικρή τιμή σημαίνει ότι δεν ακολουθείται κανονική κατανομή και το μοντέλο θεωρείται προκατειλημμένο, συνεπώς σημαντικά είναι τα μοντέλα με τη μεγαλύτερη τιμή [33, 34].

### 2.2.2 Γεωγραφικά Σταθμισμένη Παλινδρόμηση

Η Γεωγραφικά Σταθμισμένη Παλινδρόμηση (Geographically Weighted Regression – GWR) αποτελεί μία σύγχρονη μέθοδο χωρικής ανάλυσης, επιτρέποντας την εξέταση των τοπικών διακυμάνσεων σε χωρικές διεργασίες. Όταν τα χωρικά δεδομένα παρουσιάζουν θετική χωρική αυτοσυσχέτιση δεν μπορεί να χρησιμοποιηθεί η Μέθοδος Ελαχίστων Τετραγώνων, διότι παραβιάζεται το κριτήριο της ανεξαρτησίας το οποίο απαιτεί η μέθοδος. Τα τοπικά μοντέλα επιτρέπουν στις σχέσεις μεταξύ εξαρτημένης και ανεξάρτητων μεταβλητών να μεταβάλλονται στο χώρο λαμβάνοντας υπόψη τους τη χωρική δομή των δεδομένων, ενώ τα κλασικά στατιστικά μοντέλα θεωρούν ότι οι σχέσεις είναι σταθερές σε όλη την εξεταζόμενη περιοχή.

Ουσιαστικά η γεωγραφικά σταθμισμένη παλινδρόμηση αποτελεί μία παραλλαγή της απλής ή πολλαπλής γραμμικής παλινδρόμησης, συμπεριλαμβάνοντας τη χωρική παράμετρο της θέσης. Έτσι, ο γενικός τύπος της πολλαπλής παλινδρόμησης πλέον παίρνει τη μορφή της εξίσωσης (2.14) [35]:

$$y_i = b_0 + \sum_{k=1}^{m} b_k x_{ki} + \varepsilon_i \qquad (2.14)$$

όπου $y_i$ είναι η παρατήρηση της εξαρτημένης μεταβλητής, $x_{ik}$ της ανεξάρτητης, το $\varepsilon_i$ είναι το σφάλμα και το διάνυσμα b αναπαριστά τις εκτιμώμενες παραμέτρους (εκτιμήσεις των $b_k$) του ολικού μοντέλου και δίνεται, σε μορφή πινάκων, από τη σχέση (2.15) :

$$b = (X^T X)^{-1} X^T y \quad (2.15)$$

Με την τεχνική GWR εκτιμώνται οι τοπικές παράμετροι για κάθε μεταβλητή ξεχωριστά γύρω από κάθε παρατήρηση βάσει των συντεταγμένων της ($u_i$ , $v_i$). Η λογική γύρω από την οποία κινείται και λειτουργεί είναι η λογική της γειτνίασης. Οι σχέσεις (2.14) , (2.15) παίρνουν τη μορφή που ακολουθεί, όπου για κάθε σημείο i του χώρου επιδρούν μόνο τα σημεία τα οποία βρίσκονται πλησίον [36]:

$$y_i = b_0(u_i, v_i) + \sum_{k=1}^{m} b_k (u_i, v_i)\, x_{ki} + \varepsilon_i \quad (2.16)$$

$$b(u_i, v_i) = (X^T W(u_i, v_i) X)^{-1} X^T W(u_i, v_i)\, y \quad (2.17)$$

όπου W(ui, vi) είναι ένα πίνακας (μήτρα), διάστασης nxn (για n παρατηρήσεις) του οποίου τα στοιχεία εκτός της διαγωνίου είναι 0 και τα στοιχεία της διαγωνίου είναι τα βάρη των παρατηρήσεων του μοντέλου για το σημείο i. Η μήτρα αυτή γράφεται ως [37]:

$$W(u_i, v_i) = \begin{pmatrix} w_{i1} & 0 & 0 & \cdots & 0 \\ 0 & w_{i2} & 0 & \cdots & 0 \\ 0 & 0 & w_{i3} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & w_{in} \end{pmatrix} \quad (2.18)$$

Η ιδιαιτερότητα του πίνακα (2.18) είναι ότι δεν είναι σταθερός αλλά μεταβάλλεται ανάλογα με τη θέση του σημείου i. Τα στοιχεία των παρατηρήσεων που βρίσκονται κοντύτερα στο i έχουν μεγαλύτερο βάρος από εκείνα που βρίσκονται σε μεγαλύτερη απόσταση. Η επιλογή βαρών είναι ιδιαίτερα σημαντική διότι εάν η ακτίνα επιρροής που θα επιλεγεί είναι μεγάλη συνεπάγεται ότι τα στοιχεία που θα συμμετέχουν σε κάθε εκτίμηση θα καλύπτουν σχεδόν όλη την περιοχή ενδιαφέροντος. Αντίστοιχα στη περίπτωση που η επιλεγόμενη ακτίνα είναι αρκετά μικρή συνεπάγεται ότι ο αριθμός των παρατηρήσεων θα είναι μικρός και οι τιμές που θα προκύψουν θα οδηγήσουν σε εκτιμήσεις με πολύ μεγάλο τυπικό σφάλμα [37].

Προκύπτει λοιπόν, η ανάγκη προσδιορισμού ενός χωρικού πυρήνα (spatial kernel) και ενός εύρους ζώνης (bandwidth), όπου οι τιμές που θα βρίσκονται εκτός του ορίου που θα οριστεί δεν θα λαμβάνουν βάρος στους υπολογισμούς [38]. Οι κύριες κατηγορίες χωρικών πυρήνων είναι ο σταθερός (fixed), όπου η απόσταση παραμένει σταθερή σε όλη την υπό εξέταση περιοχή και ο

προσαρμοστικός (adaptive), όπου η απόσταση μεταβάλλεται ανάλογα με τον καθορισμένο αριθμό των κοντινότερων γειτόνων (Near Neighbors). Το εύρος ζώνης καθορίζει την ακτίνα γύρω από το σημείο i, όπου οι παρατηρήσεις θα σταθμιστούν και θα συμπεριληφθούν στην παλινδρόμηση.

Για τον υπολογισμό βαρών με σταθερό χωρικό πυρήνα, ο πιο απλός τρόπος στάθμισης δίνεται από τη συνάρτηση [37]:

$$w_{ij} = \begin{cases} 1, & \alpha\nu\ d_{ij} < d \\ 0, & \alpha\nu\ d_{ij} > d \end{cases} \quad (2.19)$$

όπου, d είναι το εύρος ζώνης, δηλαδή η ορισμένη απόσταση που καθορίζει ποιες παρατηρήσεις γύρω από το i θα λάβουν μη μηδενικό βάρος και επομένως θα συμμετέχουν στον υπολογισμό του μοντέλου. Ωστόσο, η μέθοδος αυτή χαρακτηρίζεται από ασυνέχεια και δεν συνηθίζεται. Ο πιο συχνά χρησιμοποιούμενος υπολογισμός βαρών σταθερού πυρήνα είναι με τη χρήση της συνάρτησης Gauss και ορίζεται ως εξής:

$$w_{ij} = \begin{cases} e^{-1/2(\frac{d_{ij}}{h})^2}, & \alpha\nu\ d_{ij} < h \\ 0\ , & \alpha\nu\ d_{ij} > h \end{cases} \quad (2.20)$$

όπου, h είναι το σταθερό εύρος ζώνης και $d_{ij}$ η ευκλείδεια απόσταση από το σημείο παλινδρόμησης i προς το γειτονικό j. Στο σημείο παλινδρόμησης το βάρος ισούται με τη μονάδα, ενώ όσο αυξάνεται η απόσταση μειώνονται τα βάρη. Με τον τρόπο αυτό οι χωρικοί πυρήνες που προκύπτουν παραμένουν σταθεροί και αμετάβλητοι σε κάθε σημείο παλινδρόμησης. Το πρόβλημα που προκύπτει με τη μέθοδο αυτή εστιάζεται στο ότι δεν συνυπολογίζεται η συγκέντρωση σημείων και ακολούθως η εξέταση ύπαρξης μικρότερων πυρήνων. Έτσι, εισάγεται η έννοια των χωρικά μεταβαλλόμενων ή προσαρμοστικών πυρήνων.

Εναλλακτικά λοιπόν γίνεται η χρήση μιας διτετραγωνικής συνάρτησης (bi-square function) η οποία βασίζεται στους κοντινότερους γείτονες και για τα σημεία τα οποία απέχουν μεγαλύτερη απόσταση από το εύρος ζώνης h, τα βάρη μηδενίζονται:

$$w_{ij} = \begin{cases} \left[1 - \left(d_{ij}/h_i\right)^2\right]^2, & \alpha\nu\ d_{ij} < h_i \\ 0\ , & \alpha\nu\ d_{ij} > h_i \end{cases} \quad (2.21)$$

Τόσο στην περίπτωση του σταθερού πυρήνα όσο και στην περίπτωση του προσαρμοστικού πυρήνα είναι απαραίτητο να επιλεγεί το εύρος ζώνης, το οποίο στην πρώτη περίπτωση είναι μία συγκεκριμένη απόσταση, ενώ στη δεύτερη είναι ο αριθμός κοντινότερων γειτόνων. Στο λογισμικό ArcGIS είναι διαθέσιμο το εργαλείο GWR, το οποίο υπολογίζει αυτόματα το ιδανικό εύρος ζώνης, χρησιμοποιώντας ένα ευρετικό αλγόριθμο, ο οποίος δοκιμάζει επαναληπτικά διαφορετικές τιμές έως ότου συγκλίνει σε μία τιμή ελαχιστοποιώντας ένα στατιστικό μέτρο [4].

Το πιο διαδεδομένο στατιστικό μέτρο που βοηθά στην αξιολόγηση της απόδοσης του μοντέλου είναι το διορθωμένο κριτήριο πληροφόρησης του Akaike (Akaike Information Criterion – AICc), το οποίο λαμβάνει υπόψη του τους βαθμούς ελευθερίας. Η τιμή AICc είναι χρήσιμη για τη σύγκριση πολλαπλών μοντέλων και για τον προσδιορισμό της καλύτερης επίδοσης. Το μοντέλο με τη μικρότερη τιμή AICc παρέχει καλύτερη προσαρμογή στα παρατηρούμενα δεδομένα [39]. Εναλλακτικά μπορεί να χρησιμοποιηθεί η τεχνική Cross-Validation κατά την οποία στόχος είναι η ελαχιστοποίηση του CV score.

Το μήνυμα που εμφανίζεται στο λογισμικό ArcGIS περιλαμβάνει τα εξής αποτελέσματα [40]:

- Bandwidth/Neighbors: αποτελεί το βέλτιστο εκτιμώμενο εύρος ζώνης ή τον βέλτιστο αριθμό γειτόνων ανάλογα με τη μέθοδο χωρικού πυρήνα που έχει επιλεγεί.
- Residual Squares: είναι το άθροισμα των τετραγωνικών υπολειμμάτων στο μοντέλο, δηλαδή το άθροισμα των διαφορών μεταξύ της εκάστοτε παρατηρούμενης τιμής και της εκτιμώμενης τιμής που αποδίδει το μοντέλο. Όσο μικρότερη είναι η τιμή, τόσο πιο κοντά είναι η εφαρμογή του μοντέλου GWR στα παρατηρούμενα δεδομένα.
- Effective Number: η τιμή αυτή αντικατοπτρίζει μια ανταλλαγή μεταξύ της διακύμανσης των προσαρμοσμένων τιμών και της απόκλισης στις εκτιμήσεις των συντελεστών και σχετίζεται με το εύρος ζώνης. Καθώς το εύρος ζώνης προσεγγίζει το άπειρο, τα γεωγραφικά βάρη για κάθε παρατήρηση πλησιάζουν τη τιμή 1 και οι εκτιμήσεις των συντελεστών θα είναι αρκετά κοντά σε ένα ολικό μοντέλο OLS. Αντίθετα, καθώς το εύρος ζώνης προσεγγίζει το μηδέν, τα γεωγραφικά βάρη για κάθε παρατήρηση πλησιάζουν το μηδέν, με εξαίρεση το ίδιο το σημείο παλινδρόμησης.

- Sigma: Αυτή η τιμή χρησιμοποιείται για υπολογισμούς AICc και είναι η τετραγωνική ρίζα του κανονικοποιημένου εναπομένοντος ποσού τετραγώνων, όπου το υπολειπόμενο άθροισμα των τετραγώνων διαιρείται με τους πραγματικούς βαθμούς ελευθερίας του υπολείμματος. Αυτή είναι η κατ 'εκτίμηση τυπική απόκλιση για τα υπολείμματα. Προτιμότερες είναι οι μικρότερες τιμές αυτού του στατιστικού στοιχείου.
- AICc: αποτελεί ένα μέτρο απόδοσης του μοντέλου και είναι χρήσιμο για τη σύγκριση διαφορετικών μοντέλων παλινδρόμησης. Λαμβάνοντας υπόψη την πολυπλοκότητα του μοντέλου, το μοντέλο με την χαμηλότερη τιμή AICc παρέχει καλύτερη προσαρμογή στα παρατηρούμενα δεδομένα.
- R2 (R-Square): είναι ένα μέτρο που δείχνει την καλύτερη προσαρμογή. Η τιμή του κυμαίνεται από 0.0 έως 1.0, ενώ προτιμώνται οι υψηλότερες τιμές. Μπορεί να ερμηνευτεί ως το ποσοστό επεξήγησης της εξαρτημένης μεταβλητής από το μοντέλο παλινδρόμησης.
- R2 (R-Square) Adjusted: όπως και στη περίπτωση της παλινδρόμησης OLS, έτσι και εδώ υπολογίζεται το προσαρμοσμένο $R^2$, όπου με τη χρήση των βαθμών ελευθερίας αντισταθμίζεται ο αριθμός των μεταβλητών ενός μοντέλου. Η τιμή αυτή είναι πάντοτε μικρότερη του $R^2$.

Κατά την εκτέλεση της GWR δημιουργείται επίσης ένα νέο feature class, το οποίο περιλαμβάνει πληροφορία για τις εκτιμώμενες τιμές, όπως:

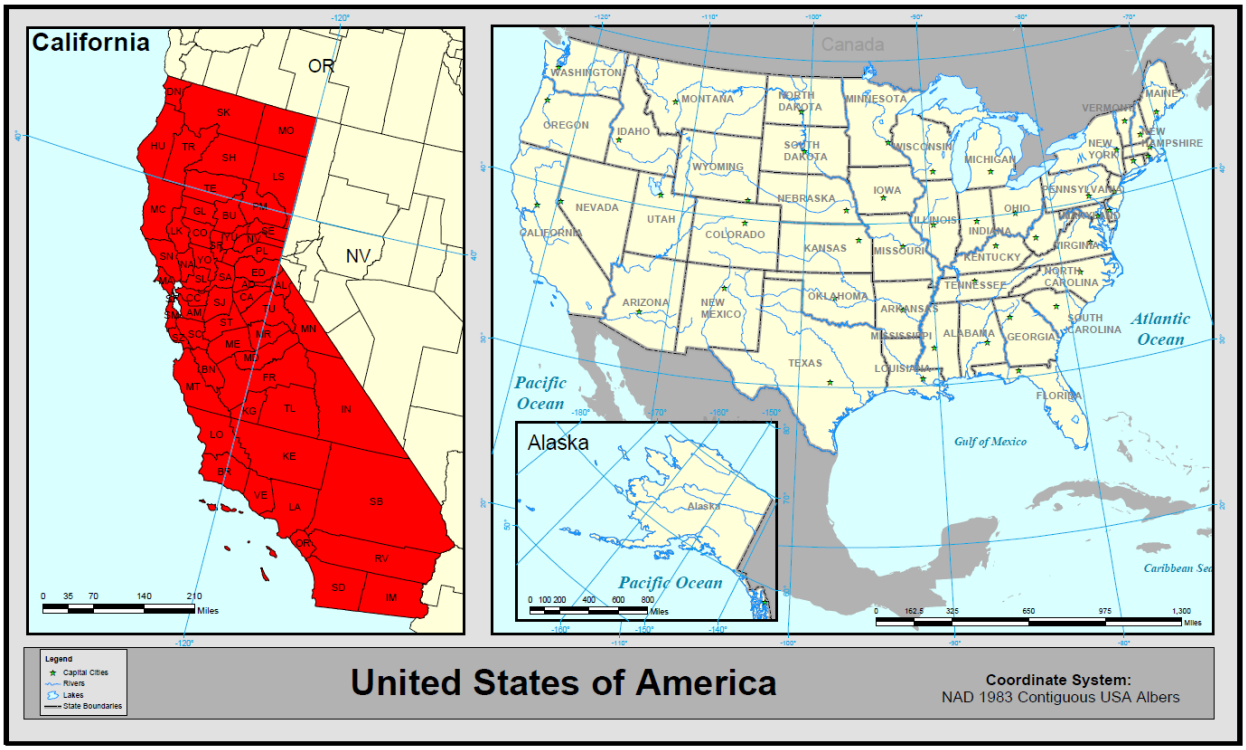
- Condition Number: Με την τιμή αυτή αξιολογείται η τοπική πολυκεντρικότητα. Τα αποτελέσματα με τιμή μεγαλύτερη από 30 αποδεικνύουν ισχυρή τοπική πολυκεντρικότητα και μπορεί να είναι αναξιόπιστα.
- Local R2: το τοπικό $R^2$ κυμαίνεται μεταξύ 0.0 και 1.0 και υποδεικνύει πόσο καλά προσαρμόζεται το τοπικό μοντέλο στις παρατηρούμενες τιμές.
- Predicted: είναι οι εκτιμώμενες τιμές που υπολογίζονται από τη GWR.
- Residuals: είναι οι υπολειμματικές τιμές που αφαιρέθηκαν από τις παρατηρούμενες τιμές. Τα υπόλοιπα έχουν μέση τιμή 0 και τυπική απόκλιση 1.

- Coefficient Standard Error: Αυτές οι τιμές μετρούν την αξιοπιστία κάθε εκτίμησης συντελεστών. Η εμπιστοσύνη στις εκτιμήσεις αυτές είναι μεγαλύτερη όταν τα τυπικά σφάλματα είναι μικρά σε σχέση με τις πραγματικές τιμές συντελεστών. Τα μεγάλα τυποποιημένα σφάλματα ενδέχεται να υποδεικνύουν προβλήματα με την τοπική πολυκεντρικότητα.

Το ArcMap δίνει τη δυνατότητα παρουσίασης αυτής της πληροφορίας, μέσω χάρτη όπου γίνεται κατηγοριοποίηση των τιμών από ψυχρά σε θερμά χρώματα, δίνοντας την δυνατότητα γρήγορης ερμηνείας των αποτελεσμάτων μέσω της οπτικοποίησής τους.

# Κεφάλαιο 3ο - Περιοχή Μελέτης

Η πολιτεία της Καλιφόρνια αποτελεί την μεγαλύτερη σε πληθυσμό πολιτεία των ΗΠΑ (Ηνωμένων Πολιτειών της Αμερικής) με 39.5 εκατομμύρια κατοίκους (σύμφωνα με την ετήσια έκθεση πληθυσμού που κυκλοφόρησε από το Υπουργείο Οικονομικών της Καλιφόρνια για την 01 Ιανουαρίου 2017 [41]). Έρχεται τρίτη κατά σειρά σε έκταση (423,970 $km^2$), ενώ προηγούνται η Αλάσκα και το Τέξας. Γεωγραφικά βρίσκεται στη δυτική Αμερική και συνορεύει βόρεια με την πολιτεία του Όρεγκον, ανατολικά και βορειοανατολικά με τη πολιτεία της Νεβάδα, νοτιοανατολικά με την πολιτεία της Αριζόνα, νότια με το Μεξικό και δυτικά με τον Ειρηνικό Ωκεανό.




*Εικόνα 4: Χάρτης της πολιτείας Καλιφόρνια των ΗΠΑ.*

Η Καλιφόρνια διαχωρίζεται σε 58 περιφέρειες (Εικόνα 5) και περιλαμβάνει 482 δήμους. Πρωτεύουσά της είναι το Σακραμέντο και οι μεγαλύτερες πόλεις της πολιτείας είναι το Λος Άντζελες, το Σαν Φρανσίσκο, το Σαν Ντιέγκο, το Φρέσνο και το Σακραμέντο.


| County | Abbr. | County | Abbr. |
|---|---|---|---|
| Alameda | AM | Orange | OR |
| Alpine | AL | Placer | PL |
| Amador | AD | Plumas | PM |
| Butte | BU | Riverside | RV |
| Calaveras | CA | Sacramento | SA |
| Colusa | CO | San Benito | BN |
| Contra Costa | CC | San Bernardino | SB |
| Del Norte | DN | San Diego | SD |
| El Dorado | ED | San Francisco | SF |
| Fresno | FR | San Joaquin | SJ |
| Glenn | GL | San Luis Obispo | LO |
| Humboldt | HU | San Mateo | SM |
| Imperial | IM | Santa Barbara | BR |
| Inyo | IN | Santa Clara | SC |
| Kern | KE | Santa Cruz | SZ |
| Kings | KG | Shasta | SH |
| Lake | LK | Sierra | SE |
| Lassen | LS | Siskiyou | SK |
| Los Angeles | LA | Solano | SL |
| Madera | MD | Sonoma | SN |
| Marin | MA | Stanislaus | ST |
| Mariposa | MR | Sutter | SR |
| Mendocino | MC | Tehama | TE |
| Merced | ME | Trinity | TR |
| Modoc | MO | Tulare | TL |
| Mono | MN | Tuolumne | TU |
| Monterey | MT | Ventura | VE |
| Napa | NA | Yolo | YO |
| Nevada | NV | Yuba | YU |

*Εικόνα 5: Χάρτης περιφερειών της Καλιφόρνια.*

## 3.1 Τοπογραφικά Χαρακτηριστικά

Η Καλιφόρνια εκτείνεται κατά μήκος της ακτής του Ειρηνικού Ωκεανού και η ακτογραμμή της είναι πάνω από 1,340 μίλια και αποτελεί σχεδόν τα ¾ της ακτογραμμής του Ειρηνικού των Ηνωμένων Πολιτειών.

Στον κεντρικό άξονα της πολιτείας βρίσκεται η ονομαζόμενη Κεντρική Κοιλάδα (Central Valley), που οριοθετείται δυτικά από την παράκτια οροσειρά και ανατολικά από την οροσειρά Σιέρα Νεβάδα, ενώ βόρεια της βρίσκεται η οροσειρά Κανσκέιντ και νότια τα όρη Τεχατσάπι. Η Σέντραλ Βάλεϊ διαχωρίζεται σε δύο τμήματα από το δέλτα των ποταμών Σακραμέντο και Σαν Τζοακίν.

Η τοπογραφία της πολιτείας ποικίλει και περιλαμβάνει την οροσειρά Σιέρα Νεβάδα, στην οποία βρίσκεται η υψηλότερη κορυφή των πολιτειών της Ηπειρωτικής Αμερικής (πλην της Αλάσκα) με υψόμετρο 4,421 μέτρα και μέρος της καλύπτεται από παγετώνες. Στα νοτιοανατολικά της, όμως, βρίσκεται η Κοιλάδα του Θανάτου (Death Valley) , με υψόμετρο 86 μέτρα υπό της στάθμης της θάλασσας [42]. Περίπου το 45% της συνολικής έκτασης της πολιτείας καλύπτεται από δάση και

το 2% από νερό, ενώ σημαντικό μέρος της καλύπτεται από γεωργική γη. Στα νότια και κεντρικά βρίσκεται η έρημος Μοχάβε.

## 3.2 Κλιματολογικά Στοιχεία

Το κλίμα της Καλιφόρνια ποικίλλει από πολικό έως υποτροπικό, αν και το μεγαλύτερο μέρος της έχει μεσογειακό κλίμα. Στην πολιτεία βρίσκονται περιοχές με μέτριες θερμοκρασίες αλλά και περιοχές που η θερμοκρασία φθάνει σε ακραίες τιμές ζέστης ή κρύου. Κατά κανόνα τα καλοκαίρια είναι ζεστά, ενώ οι χειμώνες είναι μέτριοι έως ψυχροί.

Η υψηλότερη θερμοκρασία που έχει καταγραφεί στη Γη ήταν 56.7°C και ήταν τον Ιούλιο του 1913 στην Κοιλάδα του Θανάτου. Η χαμηλότερη θερμοκρασία που έχει καταγραφεί στη πολιτεία ήταν -43°C τον Ιανουάριο του 1937 στην Μπόκα της Νεβάδα.

Η παράκτια Καλιφόρνια το καλοκαίρι τείνει να έχει δροσερό κλίμα, ενώ μόλις λίγα χιλιόμετρα στο εσωτερικό οι ακραίες θερμοκρασίες είναι σημαντικά υψηλότερες. Με αυξανόμενη την απόσταση από την ακτή, ανάλογα με την έκταση της θαλάσσιας επιρροής που παρατηρείται, σημειώνονται σημαντικά υψηλότερες θερμοκρασίες σε περιοχές όπως το Λος Άντζελες και ο κόλπος του Σαν Φρανσίσκο [43].

Το έτος 2015 βάσει στοιχείων του Αμερικανικού Εθνικού Κέντρου Περιβαλλοντικών Πληροφοριών η μέση θερμοκρασία που καταγράφηκε από τους 502 μετεωρολογικούς σταθμούς που διαθέτει ήταν 15.5°C, με μέγιστη τους 34°C στη Death Valley και χαμηλότερη τους -5.4°C στη Μπόντι της Σιερα Νεβάδα.

## 3.3 Δημογραφικά Στοιχεία

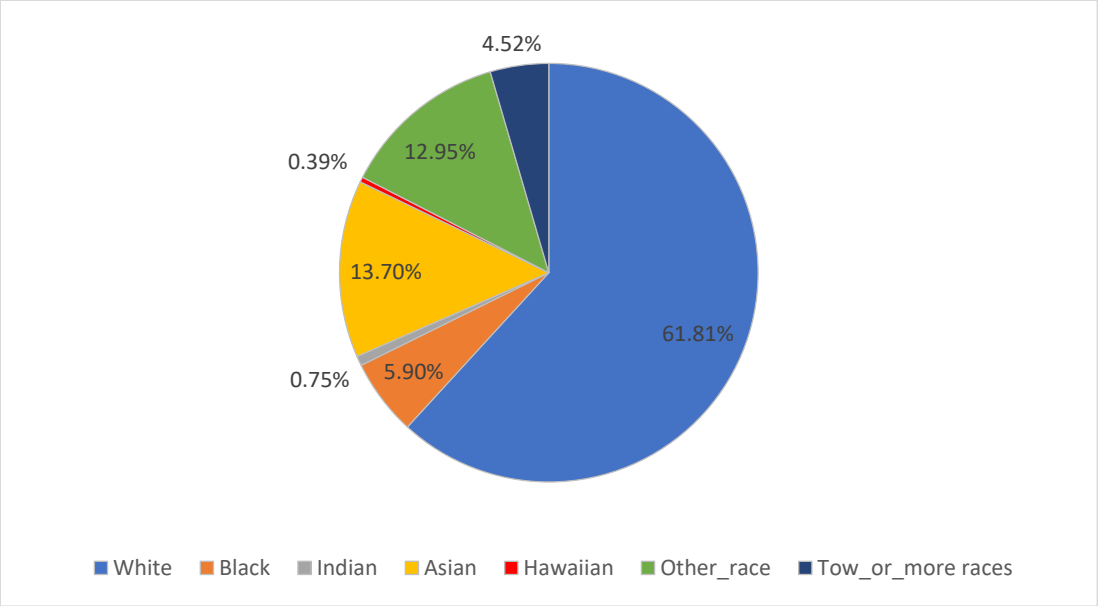
Ο πληθυσμός της πολιτείας της Καλιφόρνια το 2010 ανερχόταν σε 37.3 εκατομμύρια κατοίκους, ενώ σήμερα εκτιμάται ότι ανέρχεται σε 39.5 εκατομμύρια κατοίκους, δηλαδή έχει αυξηθεί κατά 5.94%. Ακολουθεί πίνακας με στατιστικά στοιχεία με την εκτιμώμενη πληθυσμιακή εξέλιξη της πολιτείας τα τελευταία χρόνια [44].

*Πίνακας 1: Εκτιμώμενος Πληθυσμός Καλιφόρνιας 2010 -2017.*

| Έτος | Πληθυσμός | Μεταβολή Πληθυσμού % | Αθροιστική Ποσοστιαία Αύξηση Πληθυσμού |
|---|---|---|---|
| **2010** | **37253956** | - | - |
| **2011** | 37536835 | 0.76% | 0.76% |
| **2012** | 37881357 | 0.92% | 1.68% |
| **2013** | 38238492 | 0.94% | 2.62% |
| **2014** | 38571211 | 0.87% | 3.49% |
| **2015** | 38915880 | 0.89% | 4.38% |
| **2016** | 39189035 | 0.70% | 5.09% |
| **2017** | 39523613 | 0.85% | 5.94% |

Αξίζει να σημειωθεί ότι το 1950 οι κάτοικοι ανέρχονταν μόλις σε 10.5 εκατομμύρια. Η ραγδαία αύξηση του πληθυσμού τα τελευταία χρόνια οφείλεται κυρίως στις εταιρίες που έχουν έδρα στις πόλεις της Καλιφόρνια, δημιουργώντας κέντρα παγκόσμιων τεχνολογιών και ψυχαγωγικών βιομηχανιών.

Για τη συγκεκριμένη μελέτη χρησιμοποιήθηκαν δεδομένα του έτους 2015. Στο γράφημα που ακολουθεί παρουσιάζονται τα φυλετικά ποσοστά στο σύνολο της πολιτείας.




*Εικόνα 6: Διάγραμμα φυλών στη Καλιφόρνια.*

# Κεφάλαιο 4ο - Παχυσαρκία

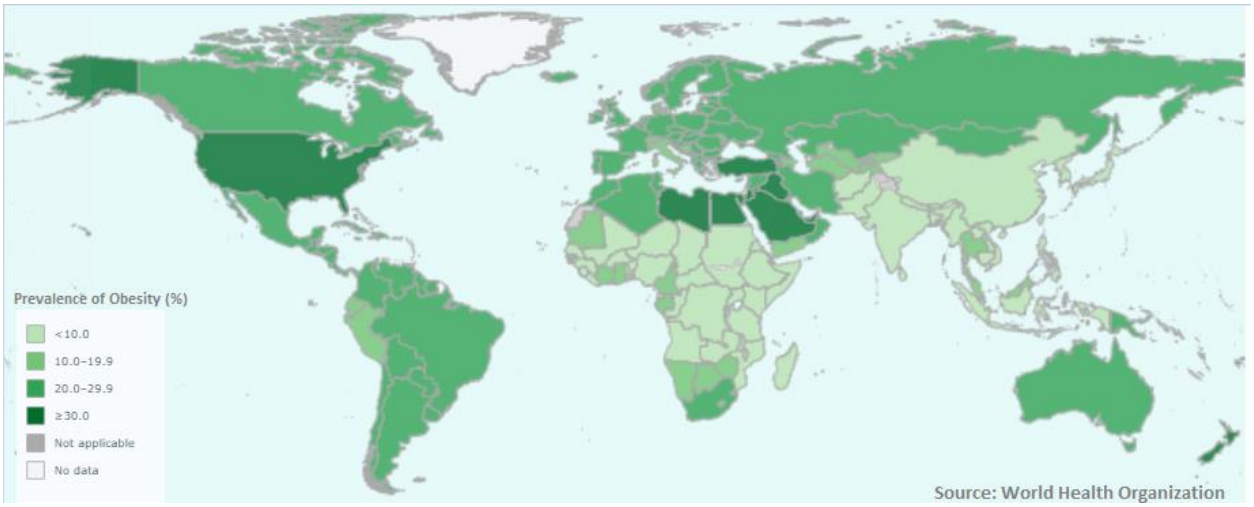
Το υπερβολικό σωματικό βάρος και η παχυσαρκία έχει αποδειχθεί ότι οδηγούν σε δυσμενείς μεταβολικές επιδράσεις στην αρτηριακή πίεση, στη χοληστερόλη, στα τριγλυκερίδια και σχετίζονται με διάφορες ασθένειες όπως καρδιαγγειακές παθήσεις, σακχαρώδης διαβήτης τύπου 2 αλλά και με ορισμένα είδη καρκίνου. Επακόλουθο αυτών των επιδράσεων είναι η μείωση του μέσου προσδόκιμου ζωής [45, 46].

Το 1832 ο Adolphe Quetelet δημιούργησε τον Δείκτη Μάζας Σώματος (Body Mass Index - BMI) [47], ώστε να υπολογίζεται ο βαθμός παχυσαρκίας ενός ατόμου. Η μέτρηση λαμβάνεται διαιρώντας το βάρος σε κιλά, με το τετράγωνο του ύψους σε μέτρα (BMI = kg / $m^2$).

Βάση του παραπάνω τύπου για τους ενήλικες, έχει γίνει παγκοσμίως αποδεκτή η εξής κατηγοριοποίηση [48]:

- BMI < 18.5 : Ελλιποβαρής
- 18.5 < BMI < 24.9 : Φυσιολογικού βάρους
- 25.0 < BMI < 29.9 : Υπέρβαρος
- BMI > 30 : Παχύσαρκος

Τόσο το υπερβολικό βάρος όσο και η παχυσαρκία δείχνουν σημαντική αύξηση τις τελευταίες δεκαετίες, με τα ποσοστά σχεδόν να είναι τριπλάσια το 2016 από το 1975. Βάσει του Παγκόσμιου Οργανισμού Υγείας (WHO), το 2016 το 39% των ανδρών και το 40% των γυναικών (39% στο σύνολο) ηλικίας άνω των 18 ετών, παγκοσμίως, ήταν υπέρβαροι (BMI ≥ 25 kg / m2) και 11% των ανδρών και το 15% των γυναικών (13% στο σύνολο) ήταν παχύσαρκοι (BMI ≥ 30 kg / m2). Τα ποσοστά αυτά αντιστοιχούν σε περίπου 2 δισεκατομμύρια ενήλικες να είναι υπέρβαροι, και από αυτούς, πάνω από μισό δισεκατομμύριο να είναι παχύσαρκοι [49].




*Εικόνα 7 : Ποσοστό παχυσαρκίας στο γενικό πληθυσμό (BMI > 30)*

Στις ΗΠΑ το 36.5% των ενηλίκων είναι παχύσαρκοι, ενώ στην πολιτεία της Καλιφόρνια το 25% βάσει έρευνας του έτους 2016 ήταν παχύσαρκοι [50]. Η Καλιφόρνια έχει το 5ο χαμηλότερο ποσοστό της χώρας, λόγω της καλύτερης πολιτικής που ακολουθεί. Ωστόσο το ποσοστό εξακολουθεί να είναι υψηλό και είναι σημαντικό να προσδιοριστούν περιβαλλοντικοί παράγοντες που το επηρεάζουν ώστε να οδηγήσουν στη μείωση του.

## 4.1 Περιγραφή Δεδομένων

Για τη χωρική και τη στατιστική ανάλυση της παχυσαρκίας για την πολιτεία της Καλιφόρνια των ΗΠΑ, χρησιμοποιήθηκαν τα παρακάτω δεδομένα, τα οποία διατίθενται ελεύθερα στο διαδίκτυο.

### 4.1.1 Δεδομένα Παχυσαρκίας

Πρόκειται για δεδομένα που διατίθενται σε μορφή shapefile της ESRI (http://synthpopviewer.rti.org/obesity/download.html), από τον μη κερδοσκοπικό οργανισμό RTI International, ο οποίος προσανατολίζεται στην έρευνα. Η περιγραφική πληροφορία που περιλαμβάνει το αρχείο, μεταξύ άλλων, αφορά τον πληθυσμό ανά κελί κανάβου 250μ , ο οποίος είναι 20 ετών ή μεγαλύτερος και το ποσοστό αυτού του πληθυσμού με BMI μεγαλύτερο ή ίσο με 30, δηλαδή το ποσοστό των ανθρώπων που κατατάσσονται στην κατηγορία της παχυσαρκίας.

### 4.1.2 Δημογραφικά Στοιχεία

Από την Υπηρεσία Απογραφής των Ηνωμένων Πολιτειών (United States Census Bureau) χρησιμοποιήθηκε η γεωβάση TIGER/Line® ACS_2015_5YR_ZCTA.gdb, η οποία περιλαμβάνει

δημογραφικά στοιχεία ανά ταχυδρομικό κώδικα. Συγκεκριμένα χρησιμοποιήθηκαν μεταβλητές που αφορούν στο επίπεδο εκπαίδευσης των ενηλίκων, στο εισόδημα τους και στο εάν ο τρόπος με τον οποίο μετακινούνται για την εργασία τους δεν περιλαμβάνει κάποιο αυτοκινούμενο μέσο. (https://www.census.gov/cgi-bin/geo/shapefiles/index.php).

### 4.1.3 Αθλητικές Εγκαταστάσεις

Οι αθλητικές εγκαταστάσεις για το σύνολο της πολιτείας της Καλιφόρνια, συλλέχθηκαν μέσω της διεπαφής προγραμματισμού (**A**pplication **P**rogramming **I**nterface - **API**) του Foursquare. Η πληροφορία αρχικά εγγράφηκε σε αρχείο csv, μέσω πηγαίου κώδικα Python και στη συνέχεια έγινε η μετατροπή σε shapefile μέσω του ArcCatalog της ESRI. Λεπτομερώς η διαδικασία προετοιμασίας των δεδομένων των αθλητικών εγκαταστάσεων περιγράφεται στο επόμενο κεφάλαιο.

#### *4.1.3.1 Προετοιμασία Δεδομένων Αθλητικών Εγκαταστάσεων*

Το Foursquare αποτελεί έναν δημοφιλή διαδικτυακό τόπο, ο οποίος βασίζεται στη χωρική πληροφορία και περιλαμβάνει τεράστιο εύρος κατηγοριών δραστηριότητας του πληθυσμού που το χρησιμοποιεί. Η χρήση του Foursquare και η άντληση πληροφορίας από αυτό μπορεί να εξασφαλίσει τα αρχικά δεδομένα ή μέρος των δεδομένων, μιας μελέτης ενός φαινομένου.

Στόχος αυτού του κεφαλαίου είναι η δημιουργία ενός αρχείου shp με σημειακές οντότητες, οι οποίες αντιστοιχούν στις αθλητικές εγκαταστάσεις της πολιτείας της Καλιφόρνια των ΗΠΑ, μέσω του Foursquare API.

Η υλοποίηση πραγματοποιήθηκε μέσω πηγαίου κώδικα Python, αφού πρωτίστως αποφασίστηκε η μεθοδολογία που θα ακολουθεί. Σημαντικό μέρος καταλαμβάνει, ωστόσο, η προετοιμασία του περιβάλλοντος που γράφτηκε ο πηγαίος κώδικας, η οποία περιλαμβάνει την εγκατάσταση της Python, του Jupyter Notebook και την εγκατάσταση βιβλιοθηκών της Python.

#### Jupyter Notebook

Το Jupyter Notebook είναι μία διαδικτυακή εφαρμογή ανοιχτού κώδικα, η οποία επιτρέπει στον χρήστη να δημιουργεί και να μοιράζεται έγγραφα που περιέχουν κώδικα, εξισώσεις, οπτικοποιήσεις και κείμενο. Οι χρήσεις περιλαμβάνουν τον καθαρισμό (data cleaning) και τον

μετασχηματισμό δεδομένων, την αριθμητική προσομοίωση, τη στατιστική μοντελοποίηση, την οπτικοποίηση δεδομένων, την μηχανική μάθηση καθώς και πολλά άλλα [51].

Το Notebook υποστηρίζει περισσότερες από 40 γλώσσες προγραμματισμού μεταξύ αυτών και την γλώσσα προγραμματισμού της Python. Ο κώδικας που γράφεται σε Jupyter Notebook μπορεί να παράγει πλούσια διαδραστική έξοδο όπως ιστοσελίδες HTML, εικόνες και βίντεο [52].

Notebook Document

Τα έγγραφα Notebook είναι έγγραφα που παράγονται από την εφαρμογή Jupyter Notebook, τα οποία περιέχουν τόσο πηγαίο κώδικα (π.χ. Python) όσο και στοιχεία εμπλουτισμένου κειμένου (παραγράφους, εξισώσεις, αριθμούς, links, κ.α.). Τα έγγραφα Notebook είναι ανθρώπινα αναγνώσιμα έγγραφα που περιέχουν την περιγραφή της ανάλυσης και τα αποτελέσματα, και ταυτόχρονα είναι εκτελέσιμα έγγραφα που μπορούν να χρησιμοποιηθούν για την ανάλυση των δεδομένων.

Jupyter Notebook App

Η εφαρμογή Jupyter Notebook είναι μια εφαρμογή διακομιστή-πελάτη που επιτρέπει την επεξεργασία και εκτέλεση εγγράφων Notebook μέσω ενός προγράμματος περιήγησης ιστού. Η εφαρμογή Jupyter Notebook μπορεί να εκτελεστεί σε έναν τοπικό υπολογιστή, που δεν απαιτεί πρόσβαση στο Internet ή μπορεί να εγκατασταθεί σε απομακρυσμένο διακομιστή και να έχει πρόσβαση μέσω του Διαδικτύου.

Εκτός από την εμφάνιση / επεξεργασία / λειτουργία σημειωματάριων, η εφαρμογή Jupyter Notebook App διαθέτει πίνακα ελέγχου, ο οποίος εμφανίζει τοπικά αρχεία και επιτρέπει την πρόσβαση σε έγγραφα Notebook ή τον τερματισμό των πυρήνων τους (kernel).

Kernel

Ο πυρήνας Notebook είναι ένας "υπολογιστικός μηχανισμός" που εκτελεί τον κώδικα που περιέχεται σε ένα έγγραφο Notebook. Ο πυρήνας ipython εκτελεί κώδικα Python. Υπάρχουν επίσημοι πυρήνες για πολλές άλλες γλώσσες προγραμματισμού.

Όταν ανοίγεται ένα έγγραφο σημειωματάριου, ο συνδεδεμένος πυρήνας εκκινείται αυτόματα. Όταν εκτελείται το Notebook (είτε cell-by-cell είτε με το μενού Cell -> Run All), ο πυρήνας εκτελεί τον υπολογισμό και παράγει τα αποτελέσματα. Ανάλογα με τον τύπο των υπολογισμών, ο πυρήνας μπορεί να καταναλώνει σημαντική CPU και RAM. Σημειώνεται ότι η μνήμη RAM δεν απελευθερώνεται μέχρι να τερματιστεί ο πυρήνας.

Notebook Dashboard

Ο πίνακας ελέγχου του Notebook είναι το στοιχείο που εμφανίζεται πρώτο όταν ξεκινάτε την εφαρμογή Jupyter Notebook. Ο πίνακας εργαλείων Notebook χρησιμοποιείται κυρίως για το άνοιγμα εγγράφων και για τη διαχείριση των πυρήνων που εκτελούνται (απεικόνιση και τερματισμός λειτουργίας).

Ο Πίνακας ελέγχου του Notebook διαθέτει άλλες λειτουργίες παρόμοιες με τις λειτουργίες διαχείρισης αρχείων, δηλαδή την πλοήγηση στους φακέλους και τη μετονομασία / διαγραφή αρχείων.

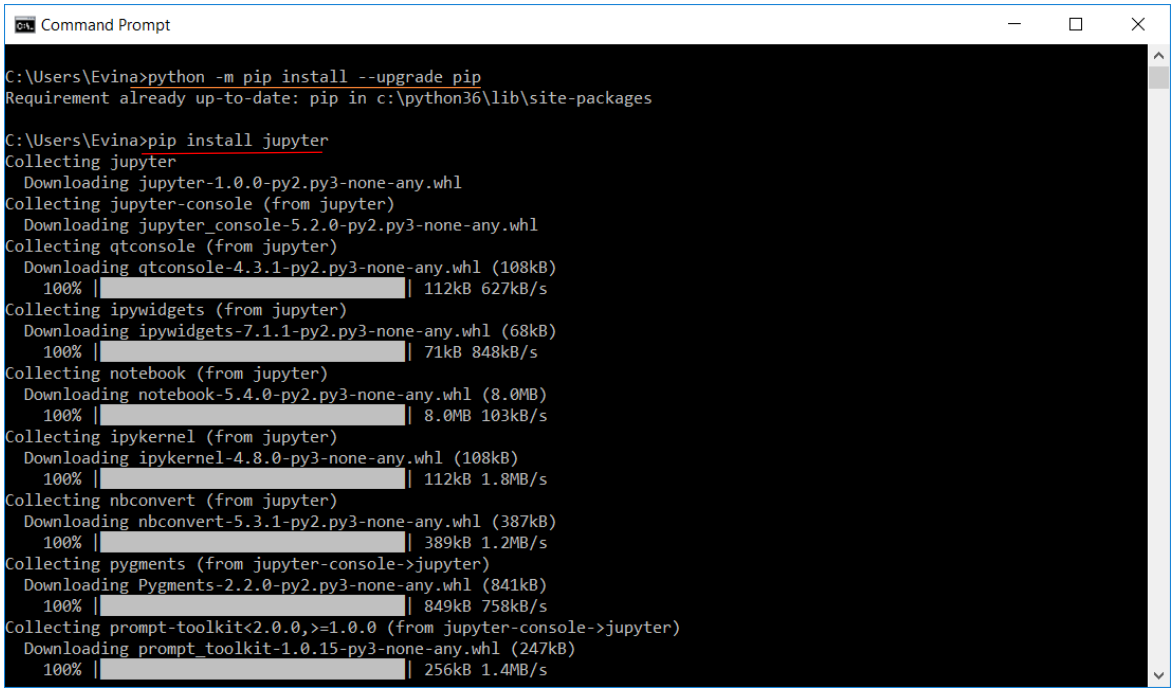
Εγκατάσταση Jupyter [53]

Βασική προϋπόθεση λειτουργίας του Jupyter είναι η εγκατάσταση της Python. Κατά την εγκατάσταση της Python πρέπει να δοθεί μεγάλη προσοχή στο να επιλεγεί το Add Python to PATH,έτσι ώστε η Python να είναι προσβάσιμη από την γραμμή εντολών.


*Εικόνα 8: Εγκατάσταση της Python.*

Η εγκατάσταση, η εκκίνηση και ο τερματισμός της εφαρμογής Jupyter Notebook πραγματοποιείται μέσω της γραμμής εντολών (Command Prompt για τα Windows, Terminal για Mac/Linux). Το Command Prompt χρησιμοποιείται για την εκτέλεση εντολών που έχουν εισαχθεί. Οι περισσότερες από αυτές τις εντολές χρησιμοποιούνται για την αυτοματοποίηση εργασιών μέσω δέσμης ενεργειών και αρχείων δέσμης, την εκτέλεση προηγμένων λειτουργιών διαχείρισης και την αντιμετώπιση προβλημάτων και την επίλυση συγκεκριμένων ζητημάτων των Windows [54].

Στη γραμμή εντολών χρησιμοποιείται η εντολή pip, η οποία βασίζεται σε ένα σύστημα διαχείρισης πακέτων που χρησιμοποιείται για την εγκατάσταση και διαχείρισης πακέτων λογισμικού γραμμένα σε Python. Η πρώτη μαρκαρισμένη εντολή είναι για την αναβάθμιση του pip ή για επαλήθευση ότι υπάρχει η τελευταία έκδοση του pip, ενώ η δεύτερη είναι για την εγκατάσταση του Jupyter.




*Εικόνα 9: Αναβάθμιση pip και εγκατάσταση Jupyter στο Command Prompt.*

### Python

Η Python είναι μία γλώσσα προγραμματισμού ανοιχτού κώδικα, πολύ υψηλού επιπέδου με δυναμική σημασιολογία. Κύριος στόχος της είναι η αυξημένη παραγωγικότητα του

προγραμματιστή και η αναγνωσιμότητα του κώδικα. Όσο πιο υψηλού επιπέδου είναι μία γλώσσα προγραμματισμού, τόσο πιο κοντά στην ανθρώπινη σκέψη βρίσκεται. Αυτό σημαίνει ότι είναι πιο εύκολο να γραφτούν προγράμματα και συνήθως λειτουργούν σε περισσότερες πλατφόρμες, θυσιάζοντας όμως μέρος της ταχύτητας των προς εκτέλεση προγραμμάτων. Η κύρια βιβλιοθήκη της περιλαμβάνει τα πάντα από ασύγχρονη επεξεργασία έως συμπιεσμένα αρχεία. Οι ευκολίες που παρέχει είναι πολύ σημαντικές καθώς καλύπτει ένα ευρύ φάσμα πιθανών προβλημάτων που μπορεί να αντιμετωπίσει κανείς.

Η ίδια η γλώσσα είναι επεκτάσιμη καθώς ένα βασικό σύνολο της γλώσσας αποτελεί τον πυρήνα της, ενώ όλα τα υπόλοιπα είναι αρθώματα (modules) που επεκτείνουν την λειτουργικότητας της, γεγονός που σε συνδυασμό με το ότι είναι ανοιχτού κώδικα την βοηθά να μην μένει στάσιμη [55].

### Βιβλιοθήκες Python

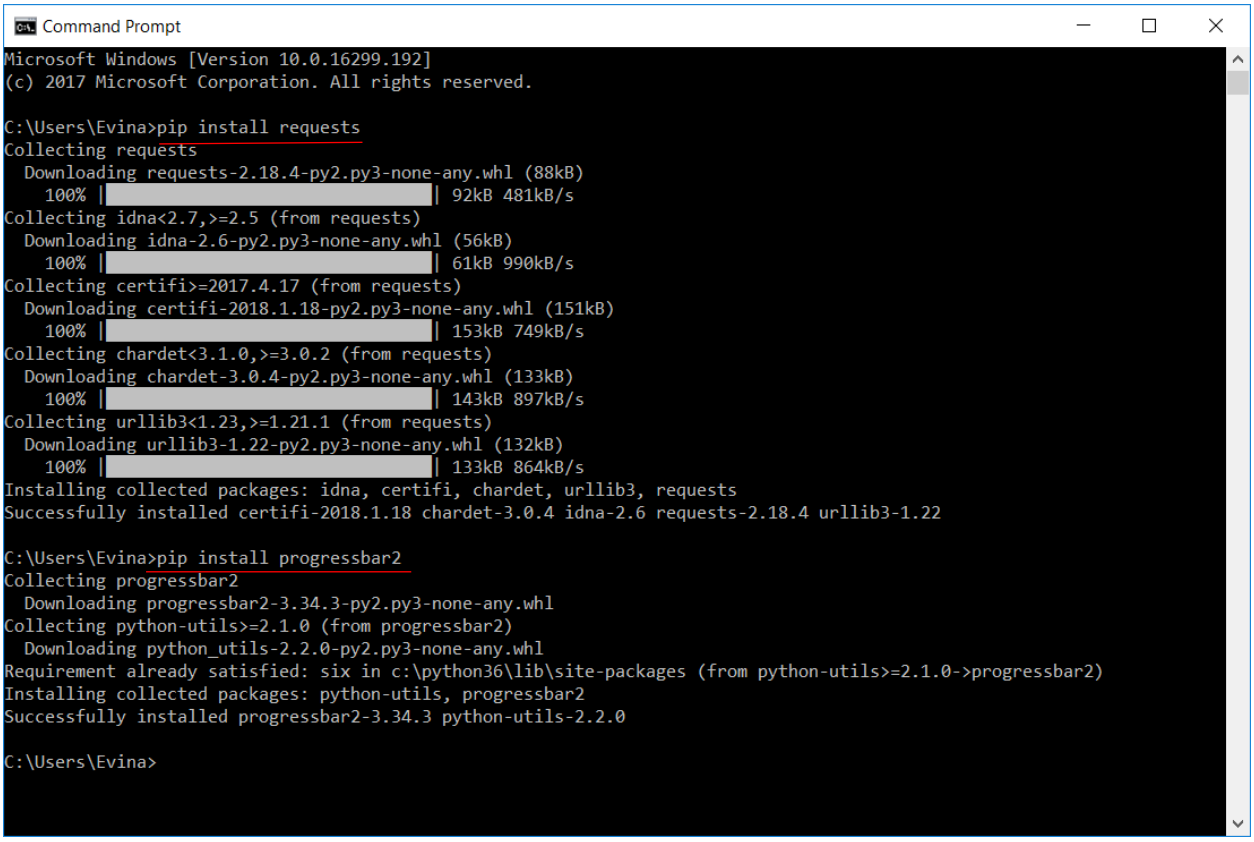
Όπως θα παρουσιαστεί στη συνέχεια στον πηγαίο κώδικα, αυτής της εργασίας, χρησιμοποιήθηκαν κάποιες έτοιμες βιβλιοθήκες και συγκεκριμένα οι παρακάτω :

- JSON (JavaScript Object Notation): Βιβλιοθήκη κωδικοποίησης και αποκωδικοποίησης JSON. Το JSON αποτελεί ένα ελαφρύ πρότυπο ανταλλαγής δεδομένων. Το JSON είναι μια μορφή αρχείου που χρησιμοποιεί κείμενο εύκολα αναγνώσιμο από τον άνθρωπο για τη ανταλλαγή και αποθήκευση δεδομένων που αποτελούνται από ζεύγη ονόματος/τιμής και από διατεταγμένη λίστα τιμών (πίνακες). Επίσης είναι απόλυτα ανεξάρτητη από τη γλώσσα προγραμματισμού αλλά χρησιμοποιεί συμβάσεις που είναι εξοικειωμένες με την οικογένεια γλωσσών C συμπεριλαμβανομένης και της Python [56, 57].
- Requests: Η βιβλιοθήκη requests επιτρέπει την αποστολή αιτημάτων HTTP μέσω της Python [58].
- Csv: Βιβλιοθήκη διαχείρισης αρχείων csv. Τα αρχεία csv χρησιμοποιούνται για την αποθήκευση μεγάλου αριθμού μεταβλητών. Το κείμενο μέσα σε ένα αρχείο csv είναι γραμμένο σε σειρές και κάθε μία από αυτές έχει στήλες διαχωρισμένες με κόμμα, το οποίο ορίζει την διαίρεση των κελιών. Χρησιμοποιώντας τη βιβλιοθήκη αυτή αυτοματοποιείται η διαδικασία διαχείρισης αρχείων csv, καθιστώντας πολύ πιο εύκολη

την επεξεργασία των αρχείων. Στα πλαίσια της εργασίας χρησιμοποιήθηκαν εντολές αυτόματης καταγραφής πληροφορίας σε αρχεία τέτοιου τύπου [59].

- Time: Το module time στην Python έχει μία εύχρηστη λειτουργία που ονομάζεται sleep (ύπνος). Όπως προκύπτει και από το όνομα είναι μία εντολή η οποία διακόπτει το πρόγραμμα της Python για ένα χρονικό διάστημα που έχει οριστεί. Η χρήση του στην παρούσα εργασία ήταν αναγκαία διότι το FourSquare επέτρεπε συγκεκριμένο αριθμό αναζητήσεων – ερωτημάτων ανά ώρα [60].
- Progressbar2: Η συγκεκριμένη βιβλιοθήκη προστέθηκε για την παροχή οπτικής προόδου. Παρέχει μήνυμα ότι η επεξεργασία βρίσκεται σε εξέλιξη και με τη βοήθεια γραφικών στοιχείων εμφανίζονται το στάδιο που βρίσκεται σε εξέλιξη η διαδικασία σε ποσοστό επί τοις εκατό, ο χρόνος που έχει περάσει από την έναρξη της διαδικασίας, ο χρόνος που απομένει για την ολοκλήρωση της διαδικασίας και ένα δυναμικό μήνυμα ( μετρητής [61].

Για την χρήση των βιβλιοθηκών requests και progressbar2 ήταν απαραίτητο να γίνει αρχικά εγκατάσταση μέσω του Command Prompt.




*Εικόνα 10: Εγκατάσταση βιβλιοθηκών της Python.*

### Foursquare API

Στον προγραμματισμό υπολογιστών, μια διεπαφή προγραμματισμού εφαρμογών (**A**pplication **P**rogramming **I**nterface - **API**) είναι ένα σύνολο υπορουτίνας για ορισμούς, πρωτόκολλα και εργαλεία για την κατασκευή λογισμικού εφαρμογών. Πρόκειται για ένα σύνολο σαφώς καθορισμένων μεθόδων επικοινωνίας μεταξύ διαφόρων στοιχείων λογισμικού. Ένα API διευκολύνει την ανάπτυξη ενός προγράμματος υπολογιστή, παρέχοντας όλα τα δομικά στοιχεία, τα οποία στη συνέχεια συνθέτουν οι προγραμματιστές. Ένα API μπορεί να είναι για ένα σύστημα διαδικτυακής βάσης, λειτουργικό σύστημα, σύστημα βάσης δεδομένων, υλικό υπολογιστή ή βιβλιοθήκη λογισμικού. Μια προδιαγραφή API μπορεί να λάβει πολλές μορφές, αλλά συχνά περιλαμβάνει προδιαγραφές για ρουτίνες, δομές δεδομένων, κλάσεις αντικειμένων, μεταβλητές ή απομακρυσμένα αιτήματα [62].

Το Foursquare API επιτρέπει στους χρήστες να προγραμματίζουν χρησιμοποιώντας δεδομένα από την εφαρμογή τους. Η διαδικασία πραγματοποιείται μέσω ¨αιτημάτων¨ που αντιστοιχούν σε ερωτήματα και ως απάντηση αυτών είναι η επιστρεφόμενη πληροφορία.

Ο μέγιστος αριθμός αιτημάτων είναι 100,000 ανά ημέρα και 5,000 ανά ώρα, όταν ο χρήστης έχει δημιουργήσει λογαριασμό καταχωρώντας τα προσωπικά του στοιχεία καθώς και έναν αριθμό πιστωτικής κάρτας, χωρίς όμως να υπάρχει χρέωση. Στη περίπτωση που δεν καταχωρηθούν τα στοιχεία του χρήστη ο αριθμός αιτημάτων περιορίζεται στις 1,000 ανά ημέρα, ενώ με συνδρομή δίνεται η δυνατότητα πολύ περισσότερων αιτημάτων [63].

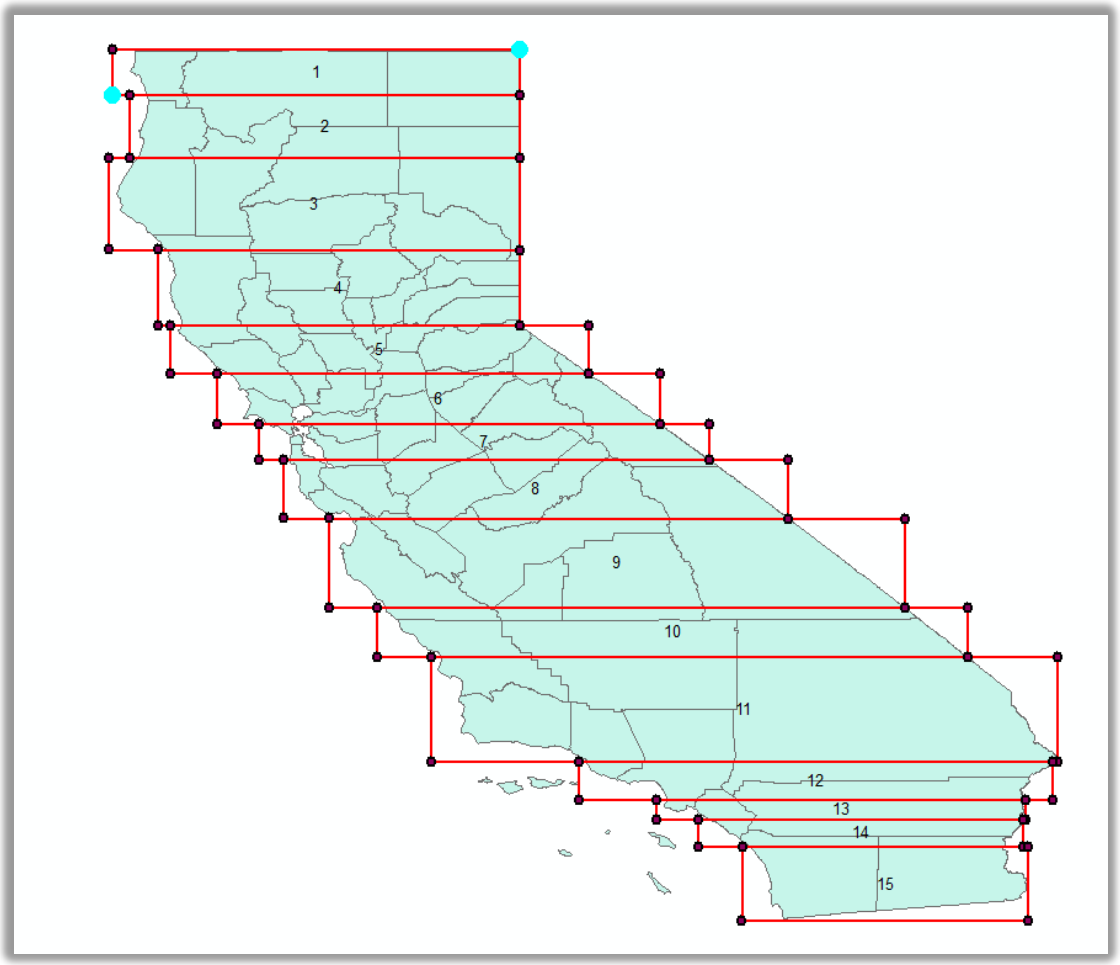
Η διαδικασία που ακολουθήθηκε βασίζεται στην αναζήτηση χώρων (Search for Venues), όπου κατά τη γραφή του κώδικα είναι απαραίτητο να προσδιορισθούν κάποιες παράμεροι (βλ. αναλυτικά Παράρτημα Β) , ώστε να επιστραφεί μία λίστα με τις τοποθεσίες στο εκάστοτε παράθυρο που έχει οριστεί. Αναλυτικά οι παράμετροι που χρησιμοποιήθηκαν είναι [64]:

- Client_id, client_secret: κωδικοί οι οποίοι δίνονται κατά την δημιουργία λογαριασμού και αντιστοιχούν στον κάθε χρήστη ως ταυτότητα.
- V (version): αντιστοιχεί στην ημερομηνία μέχρι την οποία υπάρχει συμβατότητα με το API. Σε περίπτωση αλλαγών από τον πάροχο  βάση πρωτοκόλλου δεν μπορεί να επηρεαστούν κώδικες που έχουν γραφτεί με παλαιότερη έκδοση.
- Sw, ne: ορισμός συντεταγμένων νοτιοδυτικού και βορειοανατολικού σημείου που ορίζει το παράθυρο αναζήτησης.
- Intent = ‘browse’: χρησιμοποιείται για την πραγματοποίηση αναζήτησης σε ένα συγκεκριμένο παράθυρο.
- Query: ορισμός της παραμέτρου αναζήτησης.
- categoryId: προσδιορίζεται η κατηγορία αναζήτησης (στην παρούσα η κατηγορία ¨Athletics & Sports¨, η οποία αντιστοιχεί στο ID:'4f4528bc4b90abdf24c9de85')

#### *4.1.3.2 Δημιουργία shapefile Αθλητικών Εγκαταστάσεων*

Ο αριθμός αναζητήσεων και αιτημάτων στο Foursquare API, όπως αναφέρθηκε παραπάνω, είναι περιορισμένος. Η αναζήτηση χώρων άθλησης σε εκτάσεις πέραν των ορίων της περιοχής μελέτης έχει ως αποτέλεσμα την αύξηση χρόνου της διαδικασίας καθώς και την εισαγωγή

μεγάλου όγκου δεδομένων, τα οποία στη συνέχεια δεν θα χρησιμοποιηθούν. Για την αντιμετώπιση των παραπάνω έγινε διαίρεση στην έκταση της Καλιφόρνια σε 15 πολύγωνα, όσο το δυνατόν καλύτερα προσαρμοσμένων στα όρια της πολιτείας, με τη βοήθεια του λογισμικού ArcMap 10.4.1.




*Εικόνα 11: Διαχωρισμός της πολιτείας της Καλιφόρνια σε 15 πολύγωνα.*

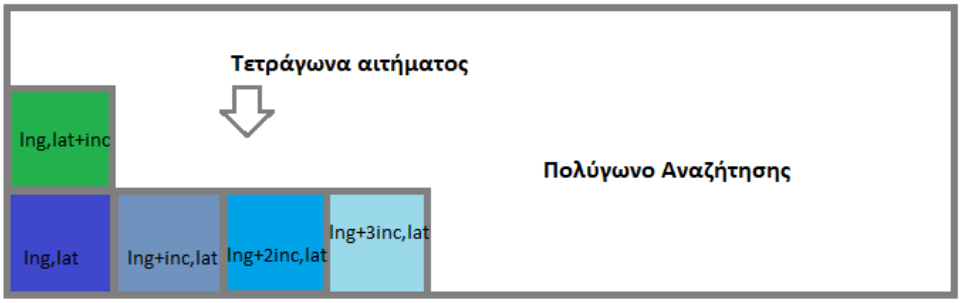
Για κάθε πολύγωνο προσδιορίστηκαν οι συντεταγμένες της Νοτιοδυτικής (sw) και της Βορειοανατολικής (ne) κορυφής του πολυγώνου, τα οποία χρησιμοποιήθηκαν για τον προσδιορισμό παραθύρου αναζήτησης.

Η αναζήτηση στο κάθε πολύγωνο γίνεται σταδιακά. Έχει ορισθεί ένα τετράγωνο περίπου ενός τετραγωνικού χιλιομέτρου, το οποίο στον κώδικα καθορίζεται ως βήμα ίσο με 0.01. Η εξίσωση υπολογισμού δίνεται προσεγγιστικά από τον τύπο:

$$\frac{inc \cdot \pi}{180°} \cdot R = \frac{0.01° \cdot \pi}{180°} \cdot 6371km = 1.11km \qquad (4.1)$$

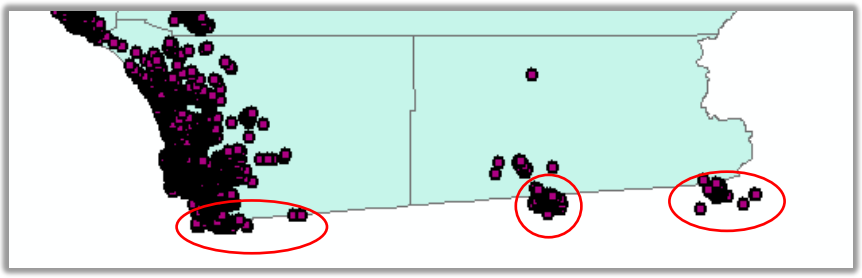
όπου inc: το βήμα και R: η ακτίνα της γης.

Το τετράγωνο σαρώνει την επιφάνεια του πολυγώνου, ξεκινώντας από το ΝΔ άκρο του πολυγώνου με φορά προς τα ανατολικά, αυξάνοντας δηλαδή κάθε φορά το lng κατά βήμα ≈1km. Κάθε θέση του τετραγώνου αντιστοιχεί σε ένα αίτημα. Όταν το τετράγωνο φτάσει στο ανατολικό όριο του πολυγώνου στη συνάρτηση αυξάνεται το lat κατά βήμα ≈1km από το αρχικό. Η διαδικασία επαναλαμβάνεται έως ότου το τετράγωνο φτάσει στη ΒΑ κορυφή του πολυγώνου που έχει ορισθεί.




*Εικόνα 12: Η κίνηση των τετραγώνων αιτημάτων.*

Όπως είναι φανερό δεν ήταν δυνατή η μη επιστροφή αποτελεσμάτων πέραν των ορίων της πολιτείας της Καλιφόρνια. Το πρόβλημα αυτό αντιμετωπίστηκε με τη βοήθεια του λογισμικού ArcGIS όπου προσδιορίστηκε το περίγραμμα της πολιτείας κ αφαιρέθηκαν οι γυμναστικές εγκαταστάσεις οι οποίες δεν ανήκουν γεωγραφικά σε αυτή.

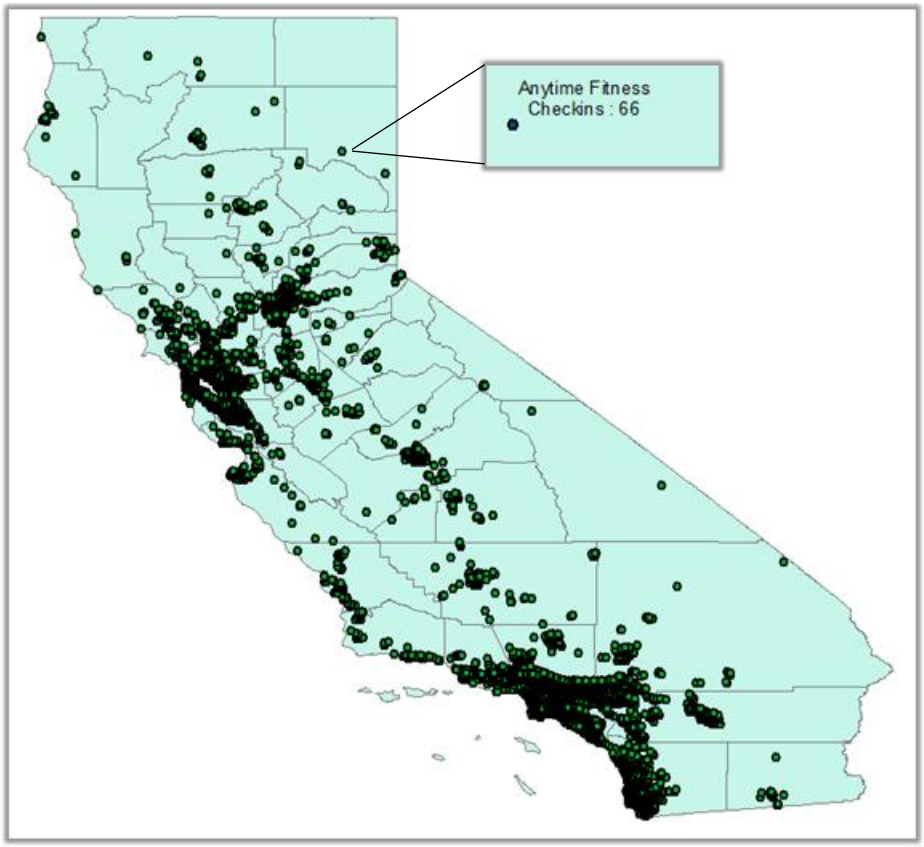


*Εικόνα 13: Επιστρεφόμενα σημεία πέραν των ορίων της περιοχής μελέτης.*

Ένα ακόμη πρόβλημα που έπρεπε να αντιμετωπιστεί ήταν η ορθότητα των καταχωρίσεων. Το Foursquare δίνει τη δυνατότητα σε κάθε χρήστη να καταχωρεί ένα μέρος χωρίς όμως να

ελέγχεται και να αποδεικνύεται άμεσα η ύπαρξη του. Υπήρχαν, για παράδειγμα, καταχωρήσεις όπου οι χρήστες δήλωναν ότι γυμνάζονται, αλλά η τοποθεσία που ήταν καταχωρημένη συνδεόταν με την οικεία τους ή με κάποιον εξωτερικό χώρο, ο οποίος δεν αποτελεί χώρο άθλησης. Για την αντιμετώπιση του προβλήματος ορίστηκε ένας αριθμός (=50) ως χαμηλότερο όριο των check-ins για την επιστροφή των αποτελεσμάτων. Με τον τρόπο αυτό δεν καταχωρήθηκαν χώροι, οι οποίο δεν αποτελούν αθλητικές εγκαταστάσεις καθώς και χώροι όπου δεν υπάρχει σημαντική συνάθροιση κοινού ή αθλητικά κέντρα τα οποία δεν βρίσκονται πλέον σε λειτουργία.

Τα τελικά αποτελέσματα περιλαμβάνουν 14,195 αθλητικές εγκαταστάσεις σε όλη την πολιτεία της, Καλιφόρνια. Η πληροφορία αποτελείται από τις συντεταγμένες, το όνομα των χώρων άθλησης και από το πλήθος των check-ins.




*Εικόνα 14: Χώροι άθλησης στην πολιτεία της Καλιφόρνια των Η.Π.Α.*

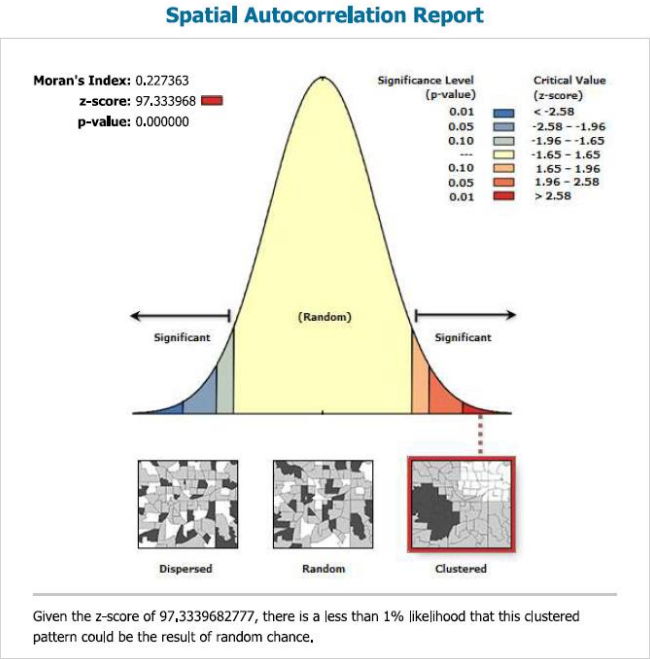
## 4.2 Μεθοδολογικό Πλαίσιο

Τα δεδομένα για την παχυσαρκία ήταν χωρισμένα σε κανάβους των 250 μέτρων. Αρχικά υπολογίστηκε το άθροισμα των ατόμων με παχυσαρκία ανά Ταχυδρομικό Κώδικα (ΤΚ) και στη συνέχεια το ποσοστό βάσει του συνολικού πληθυσμού κάθε ΤΚ. Αντίστοιχα υπολογίστηκε και ο αριθμός των συνολικών αθλητικών εγκαταστάσεων βάσει του εμβαδού κάθε πολυγώνου. Ο αριθμός των πολυγώνων των ΤΚ ανέρχεται στα 1755.

Σε δεύτερο στάδιο έγινε ανάλυση των δεδομένων ανά περιφέρεια της πολιτείας. Τα νέα ποσοστά προέκυψαν από το άθροισμα των ανθρώπων με παχυσαρκία προς το συνολικό πληθυσμό της κάθε περιφέρειας. Αντίστοιχα οι αθλητικές εγκαταστάσεις προστέθηκαν και διαιρέθηκαν με την έκταση της περιφέρειας.

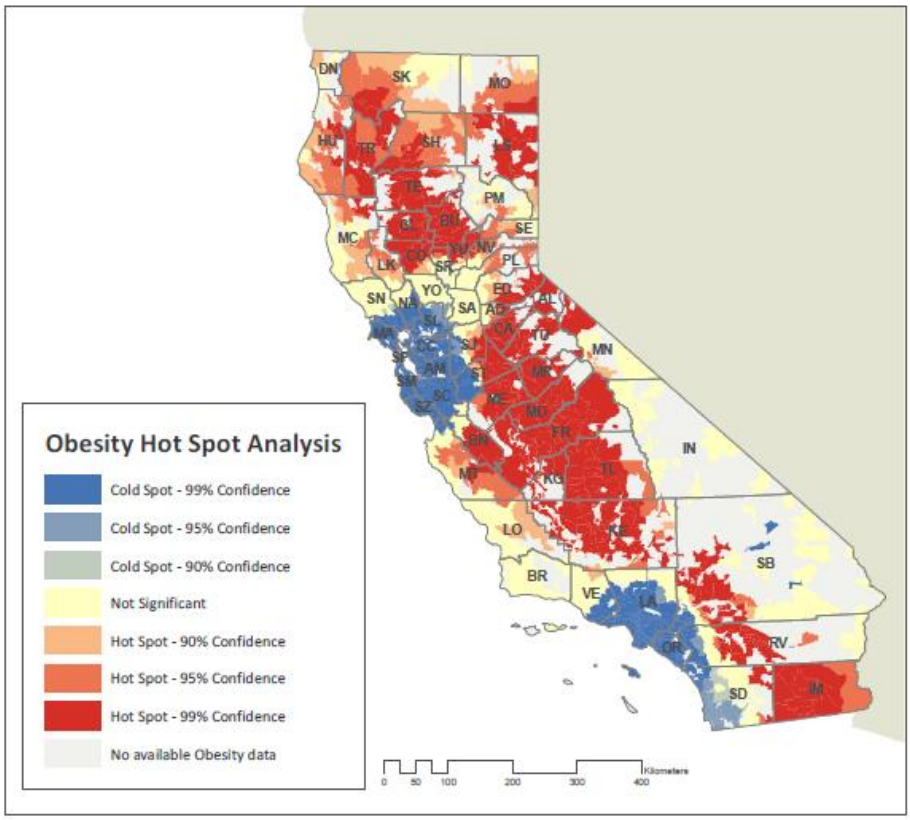
### 4.2.1 Hot Spot Ανάλυση

Αρχικά ελέγχθηκε η ύπαρξη χωρικής αυτοσυσχέτισης των τιμών της μεταβλητής της παχυσαρκίας με τον δείκτη Moran’s I. Το στατιστικό μέτρο Z που υπολογίστηκε είναι z-score:97.333968, δηλαδή μεγαλύτερο του 1,96 σε επίπεδο σημαντικότητας 95%, με P-value < 0.01 επομένως ο δείκτης Moran’s I θεωρείται στατιστικά σημαντικός και δείχνει χωρική ομαδοποίηση των υψηλών τιμών, σε συστάδες (clustered). Συγκεκριμένα, η πιθανότητα τα δεδομένα να είναι αποτέλεσμα τυχαιότητας είναι μικρότερη από 1%.




*Εικόνα 15: Αποτελέσματα ελέγχου χωρικής αυτοσυσχέτισης των τιμών της παχυσαρκίας.*

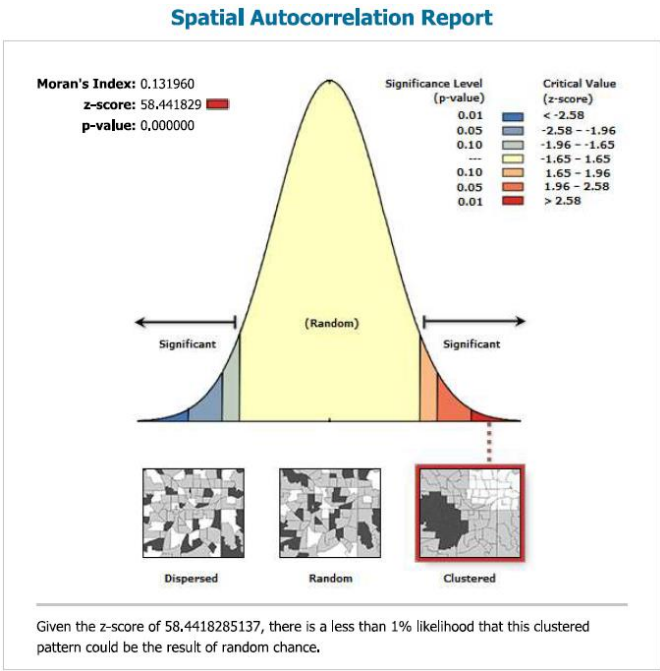
Στη συνέχεια πραγματοποιήθηκε η ανάλυση Hot Spot με χρήση του αλγορίθμου Getis-Ord Gi* του ArcMap 10.4.1. Στον χάρτη που ακολουθεί παρουσιάζονται τα Hot Spot της παχυσαρκίας στην πολιτεία της Καλιφόρνια των ΗΠΑ. Για την ανάλυση χρησιμοποιήθηκε η μέθοδος Fixed Distance μετρώντας Ευκλείδεια απόσταση , όπου κάθε χαρακτηριστικό αναλύεται στο πλαίσιο των γειτονικών του χαρακτηριστικών. Ως κρίσιμη απόσταση καθορίστηκαν τα 70058,3029 μέτρα, το οποίο συνεπάγεται ότι χαρακτηριστικά τα οποία βρίσκονται εντός του ορίου κατωφλίου ασκούν επιρροή στους υπολογισμούς. Τα γειτονικά χαρακτηριστικά που βρίσκονται εκτός της κρίσιμης απόστασης λαμβάνουν μηδενικό βάρος και δεν επηρεάζουν τους υπολογισμούς.




Εικόνα 16 : Ανάλυση Hot Spot της Παχυσαρκίας στην Καλιφόρνια των ΗΠΑ.

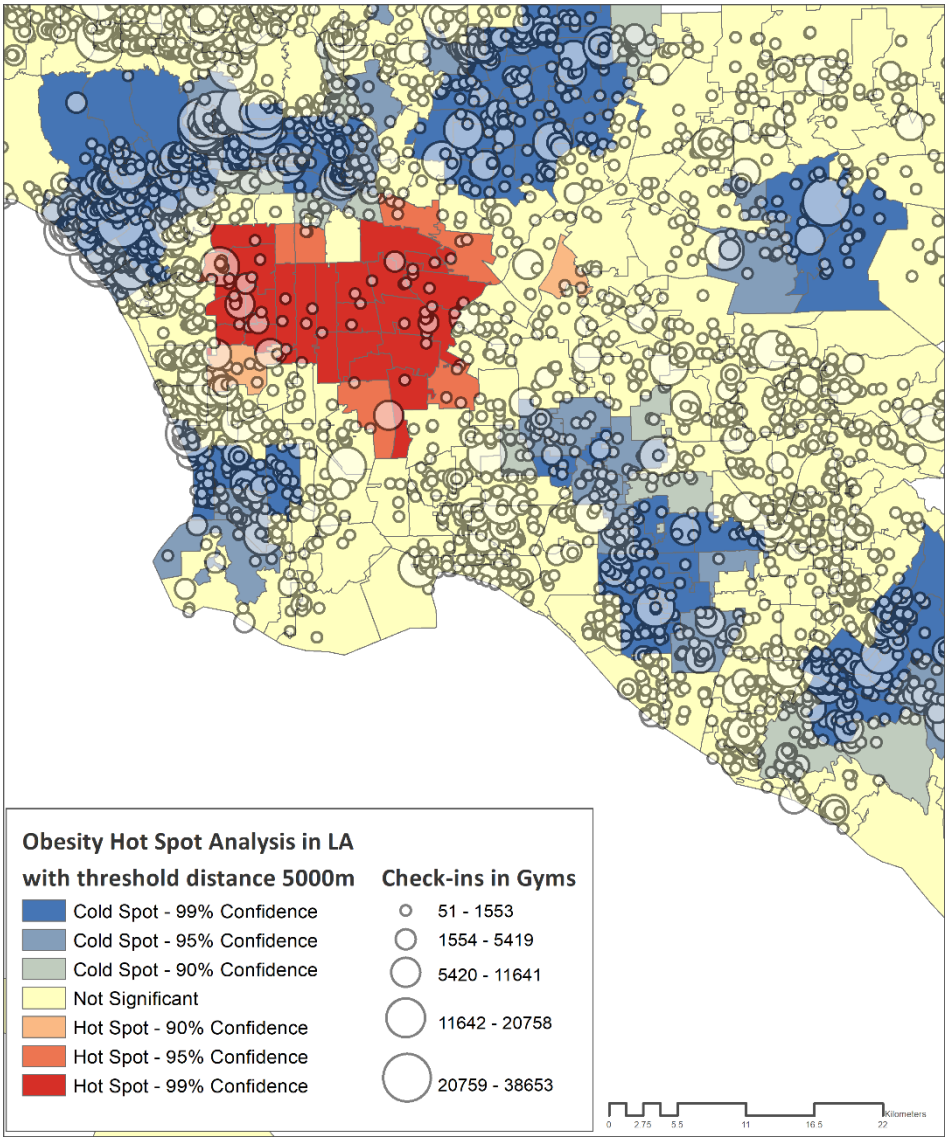
Όπως προκύπτει από την Hot Spot ανάλυση υπάρχει στατιστικά σημαντική συγκέντρωση μεγάλων τιμών παχυσαρκίας (Hot Spots) στην κεντρική Καλιφόρνια και κυρίως στις επαρχιακές περιοχές, ενώ αντίθετα στις μεγαλύτερες πόλεις και την ευρύτερη περιοχή τους, όπως την περιοχή των Σαν Φρανσίσκο – Σάντα Κλάρα και των Λος Άντζελες – Σαν Ντιέγκο, παρατηρούνται στατιστικά σημαντικά μικρές τιμές (Cold Spots).

Στη συνέχεια εξετάστηκε η χωρική αυτοσυσχέτιση των γυμναστηρίων για το σύνολο της πολιτείας. Όπως και προηγουμένως με τα κρούσματα παχυσαρκίας έτσι και με την χωρική κατανομή των γυμναστηρίων προκύπτει από την αναφορά (Report) πως είναι ομαδοποιημένα σε συστάδες. Το στατιστικό μέτρο Z που υπολογίστηκε είναι z-score:58.441829, με P-value < 0.01 επομένως ο δείκτης Moran’s I θεωρείται στατιστικά σημαντικός δείχνοντας τη χωρική ομαδοποίηση των υψηλών τιμών.




*Εικόνα 17: Αποτελέσματα ελέγχου χωρικής αυτοσυσχέτισης των τιμών των αθλητικών εγκαταστάσεων.*

Στον χάρτη που ακολουθεί παρουσιάζονται τα Cold Spots και τα Hot Spots για την περιοχή του Λος Άντζελες που αφορούν τη παχυσαρκία, καθώς και οι χώροι άθλησης. Οι χώροι άθλησης κατηγοριοποιήθηκαν βάσει επισκεψιμότητας, μέσω των check-ins, και απεικονίζονται με διαφορετικό μέγεθος. Η πληροφορία αυτή δείχνει έμμεσα και το μέγεθος μιας αθλητικής εγκατάστασης.




*Εικόνα 18 : Παρουσίαση αθλητικών εγκαταστάσεων σε σχέση με τα Hot Spot παχυσαρκίας.*

Η ανάλυση Hot Spot πραγματοποιήθηκε με χρήση του αλγορίθμου Getis-Ord Gi* με τη μέθοδο Fixed Distance και ορίστηκαν ως κατώφλι τα 5 χλμ. Παρατηρείται ότι στις περιοχές όπου υπάρχουν αρνητικές συγκεντρώσεις παχύσαρκων ατόμων, υπάρχει επίσης πληθώρα αθλητικών εγκαταστάσεων και μάλιστα με μεγάλη επισκεψιμότητα. Αντιθέτως, στο κέντρο του Λος Άντζελες, όπου φαίνονται μεγάλες τιμές παχυσαρκίας, παρατηρείται πως υπάρχει έλλειψη αθλητικών χώρων.

### 4.2.2 Χωρική και Στατιστική Ανάλυση Παχυσαρκίας

Από το προηγούμενο κεφάλαιο, προκύπτει ότι, τα ποσοστά παχυσαρκίας σε μία περιοχή, εκτός από παθολογικούς και γενετικούς παράγοντες, οι οποίοι δεν θα αναλυθούν στη παρούσα εργασία, επηρεάζονται και από περιβαλλοντικούς. Ενδεχομένως σημαντικός παράγοντας να είναι η έλλειψη αθλητικών εγκαταστάσεων, ή το επίπεδο εκπαίδευσης των ανθρώπων και το εισόδημά τους. Στο παρών κεφάλαιο, θα εξεταστούν και θα αναλυθούν μεταβλητές που επηρεάζουν την παχυσαρκία.

Exploratory Regression

Δεύτερο βήμα ήταν η αξιολόγηση των εξεταζόμενων μεταβλητών με τη χρήση του εργαλείου Exploratory Regression. Με εξαρτημένη μεταβλητή το ποσοστό παχυσαρκίας δοκιμάστηκαν ως ανεξάρτητες μεταβλητές η αναλογία αθλητικών εγκαταστάσεων, το ποσοστό πανεπιστημιακής εκπαίδευσης, το εισόδημα των ενηλίκων και το ποσοστό των ανθρώπων που πηγαίνουν στον χώρο εργασίας του με τα πόδια. Τα αποτελέσματα με κάθε μία μεταβλητή ξεχωριστά παρουσιάζονται στον Πίνακα 2 που ακολουθεί.

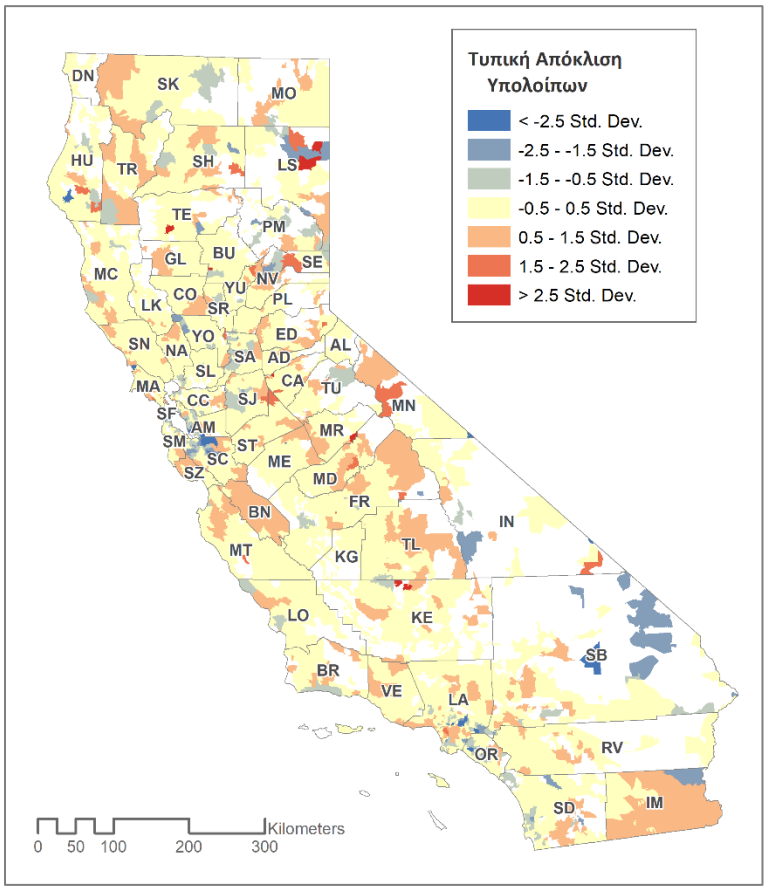
*Πίνακας 2: Προσαρμοσμένος συντελεστής προσδιορισμού με Exploratory Regression για τη παχυσαρκία.*

| **Ανεξάρτητη Μεταβλητή** | **Adjusted $R^2$** |
|---|---|
| Εκπαίδευση | 24 % |
| Εισόδημα | 16 % |
| Αθλητικές Εγκαταστάσεις | 10 % |
| Πεζοί | 1 % |

Το μοντέλο που εξηγεί καλύτερα την εξαρτημένη μεταβλητή , δηλαδή το ποσοστό της παχυσαρκίας, μετά από δοκιμές διαφόρων μοντέλων που χρησιμοποιεί η συγκεκριμένη μέθοδος, είναι το "Εκπαίδευση – Εισόδημα – Αθλητικές Εγκαταστάσεις", το οποίο εξηγεί το 29% της διακύμανσης σε επίπεδο εμπιστοσύνης 95%. Και οι τρεις ανεξάρτητες μεταβλητές έχουν αρνητικό πρόσημο, το οποίο συνεπάγεται μείωση της παχυσαρκίας με την αύξησή τους.

OLS Regression - Ανάλυση Δεδομένων ανά ΤΚ

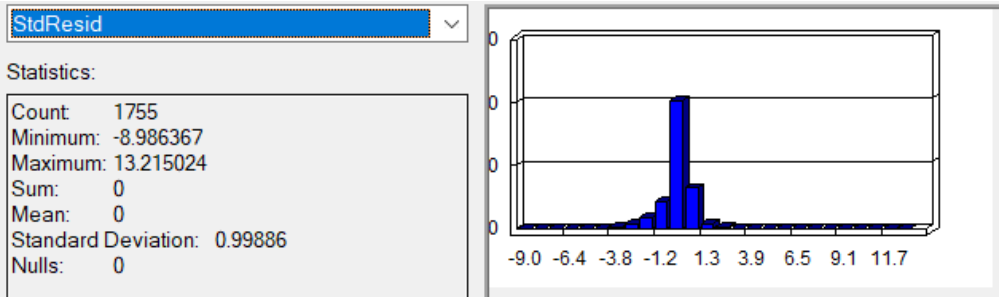
Στη συνέχεια πραγματοποιήθηκε στατιστική ανάλυση με τη μέθοδο γραμμικής παλινδρόμησης με την τεχνική Ordinary Least Squares (OLS). Η εξαρτημένη μεταβλητή του μοντέλου είναι το ποσοστό της παχυσαρκίας και οι ανεξάρτητες είναι το ποσοστό πανεπιστημιακής εκπαίδευσης, το εισόδημα σε δολάρια και το πλήθος των αθλητικών χώρων σε σχέση με το εμβαδό του κάθε ΤΚ.




*Εικόνα 19: Τυπική απόκλιση υπολοίπων παχυσαρκίας με τη μέθοδο OLS.*

Η ταξινόμηση των υπολοίπων που παρουσιάζεται στον χάρτη της Εικόνας 19 των OLS έγινε αυτοματοποιημένα από το ArcMap 10.4.1, το οποίο δημιούργησε επτά κλάσεις οι οποίες αντικατοπτρίζουν την τυπική απόκλιση των υπολοίπων. Από την παραπάνω ταξινόμηση παρατηρείται πως το σύνολο των τιμών βρίσκεται στο διάστημα -0.5 και 1.5, πέρα από ελάχιστες

εξαιρέσεις, κάτι που υποδεικνύει κανονική κατανομή και επομένως ένα μοντέλο το οποίο ταιριάζει ικανοποιητικά στα δεδομένα.




*Εικόνα 20: Στατιστικά στοιχεία υπολοίπων παχυσαρκίας με τη μέθοδο OLS.*

Όσον αφορά το εξαγόμενο μήνυμα αναφοράς (Report) το Coefficient αντικατοπτρίζει τον τύπο της σχέσης που έχει κάθε επεξηγηματική μεταβλητή με την εξαρτημένη. Όταν το πρόσημο είναι αρνητικό υποδηλώνει και αρνητική σχέση, δηλαδή με τη μείωση της επεξηγηματικής μεταβλητής συνεπάγεται αύξηση της εξαρτημένης και αντίστροφα. Όταν η τιμή είναι μηδενική υποδηλώνει ότι η συγκεκριμένη μεταβλητή δεν λαμβάνει μέρος στην παλινδρόμηση. Ο έλεγχος σημαντικότητας των ανεξάρτητων μεταβλητών γίνεται βάσει των τιμών της στήλης Probability [στήλη 1], όπου παρατηρείται πως και οι τρεις μεταβλητές έχουν την τιμή μηδέν.

Επίσης σημαντικό μέτρο είναι ο δείκτης VIF [στήλη 2], ο οποίος δείχνει την ύπαρξη πολυσυγγραμμικότητας μεταξύ των ανεξάρτητων μεταβλητών. Σύμφωνα με την ESRI η κρίσιμη τιμή είναι το 7.5, ενώ όπως παρατηρείται από το μήνυμα αναφοράς οι τιμές VIF του μοντέλου δεν ξεπερνούν την οριακή τιμή και επομένως δεν παρουσιάζεται κανένα πρόβλημα πολυσυγγραμμικότητας.

**Summary of OLS Results - Model Variables**

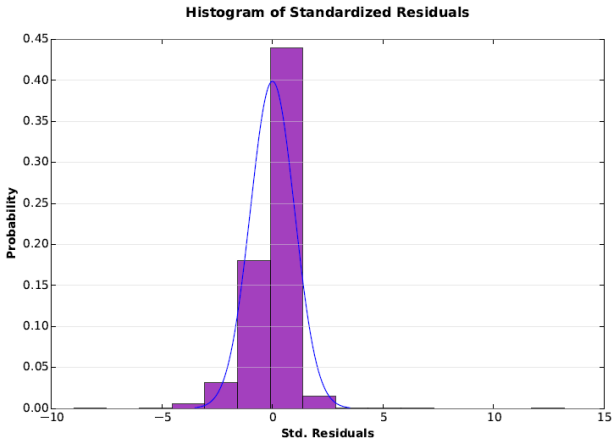
| Variable | Coefficient [a] | StdError | t-Statistic | 1 Probability [b] | Robust_SE | Robust_t | Robust_Pr [b] | 2 VIF [c] |
|---|---|---|---|---|---|---|---|---|
| Intercept | 0.418744 | 0.002467 | 169.753793 | 0.000000* | 0.003418 | 122.510176 | 0.000000* | -------- |
| HIGH_EDUCATI | -0.000873 | 0.000075 | -11.588136 | 0.000000* | 0.000087 | -10.065052 | 0.000000* | 1.946848 |
| B19019E1 | -0.000000 | 0.000000 | -5.853830 | 0.000000* | 0.000000 | -4.976408 | 0.000001* | 1.844090 |
| GYMS_AREA | -0.005313 | 0.000505 | -10.527742 | 0.000000* | 0.000835 | -6.361199 | 0.000000* | 1.076723 |

Σύμφωνα με τα στοιχεία που προκύπτουν από το μήνυμα αναφοράς της OLS, η τιμή του συντελεστή προσδιορισμού $R^2$ είναι 0.293439, ενώ η τιμή του προσαρμοσμένου $R^2$ είναι 0.292229.

**OLS Diagnostics**

| | | | |
|---|---|---|---|
| Input Features: | Obesity_zip_gyms | Dependent Variable: | OBES_RATE |
| Number of Observations: | 1755 | Akaike's Information Criterion (AICc) [d]: | -5775.875894 |
| Multiple R-Squared [d]: | 0.293439 | Adjusted R-Squared [d]: | 0.292229 |
| Joint F-Statistic [e]: | 242.400779 | Prob(>F), (3,1751) degrees of freedom: | 0.000000* |
| Joint Wald Statistic [e]: | 419.709260 | Prob(>chi-squared), (3) degrees of freedom: | 0.000000* |
| Koenker (BP) Statistic [f]: | 11.890862 | Prob(>chi-squared), (3) degrees of freedom: | 0.007767* |
| Jarque-Bera Statistic [g]: | 105059.304241 | Prob(>chi-squared), (2) degrees of freedom: | 0.000000* |

Ο επόμενος βασικός έλεγχος αφορά την κανονικότητα των υπολοίπων και ικανοποιείται με την στατιστική των Jarque και Bera. Στο σημείο αυτό εντοπίζεται πρόβλημα κανονικότητας αφού η τιμή πιθανότητας είναι 0 μικρότερη δηλαδή του 0.05 κάτι που καταδεικνύει πως το μοντέλο ενδέχεται να είναι προκατειλημμένο.




*Εικόνα 21: Ιστόγραμμα υπολοίπων ανάλυσης δεδομένων ανά ΤΚ.*

OLS Regression - Ανάλυση Δεδομένων ανά Περιφέρεια

Μια δεύτερη δοκιμή με την Μέθοδο OLS πραγματοποιήθηκε χωρίζοντας πλέον τα δεδομένα ανά περιφέρεια και όχι ανά ΤΚ. Τα αποτελέσματα του μηνύματος αναφοράς στη περίπτωση αυτή είναι εμφανώς διαφορετικά.

**Summary of OLS Results - Model Variables**

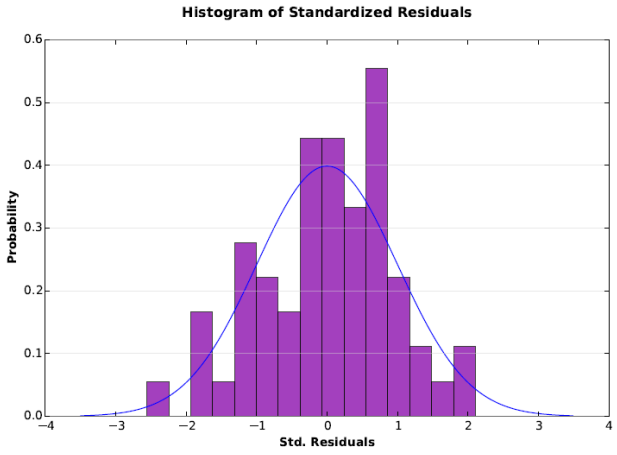
| Variable | Coefficient [a] | StdError | t-Statistic | Probability [b] | Robust_SE | Robust_t | Robust_Pr [b] | VIF [c] |
|---|---|---|---|---|---|---|---|---|
| Intercept | 0.445218 | 0.006083 | 73.189246 | 0.000000* | 0.007700 | 57.818933 | 0.000000* | -------- |
| GYMS_AREA | -0.000001 | 0.000000 | -4.972722 | 0.000007* | 0.000000 | -7.473851 | 0.000000* | 1.219314 |
| HIGH_EDUCATI | -0.111145 | 0.024958 | -4.453227 | 0.000044* | 0.029527 | -3.764127 | 0.000416* | 2.638666 |
| B19019E1_1 | -0.000001 | 0.000000 | -4.053509 | 0.000165* | 0.000000 | -3.276409 | 0.001842* | 2.612116 |

Παρατηρείται πως και οι τρεις επεξηγηματικές μεταβλητές λαμβάνουν μέρος στην παλινδρόμηση και μάλιστα με αρνητικό πρόσημο, ενώ οι τιμές p-value και VIF είναι αποδεκτές, καταδεικνύοντας τη στατιστική σημαντικότητα και τη μη ύπαρξη πολυσυγγραμικότητας.

**OLS Diagnostics**

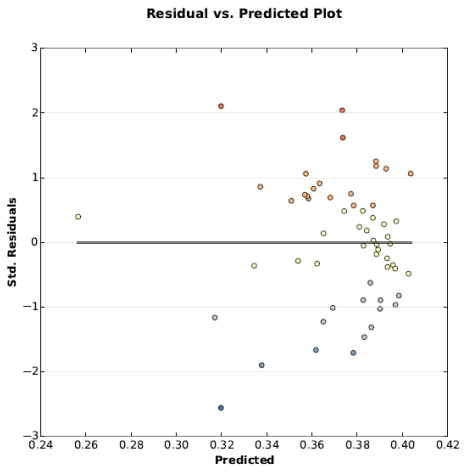
| | | | |
|---|---|---|---|
| Input Features: | Obesity_by_county | Dependent Variable: | OBES_RATE |
| Number of Observations: | 58 | Akaike's Information Criterion (AICc) [d]: | -338.474849 |
| Multiple R-Squared [d]: | 0.828026 | Adjusted R-Squared [d]: | 0.818471 |
| Joint F-Statistic [e]: | 86.666714 | Prob(>F), (3,54) degrees of freedom: | 0.000000* |
| Joint Wald Statistic [e]: | 663.143116 | Prob(>chi-squared), (3) degrees of freedom: | 0.000000* |
| Koenker (BP) Statistic [f]: | 17.129635 | Prob(>chi-squared), (3) degrees of freedom: | 0.000665* |
| Jarque-Bera Statistic [g]: | 0.648678 | Prob(>chi-squared), (2) degrees of freedom: | 0.723005 |

Η τιμή του συντελεστή προσδιορισμού $R^2$ είναι 0.828026, ενώ η τιμή του προσαρμοσμένου $R^2$ είναι 0.818471, σημαντικά μεγαλύτερη από την περίπτωση χωρισμού της περιοχής μελέτης ανά ΤΚ. Η p-value της τιμής Jarque και Bera είναι 0.723005 > 0.05, άρα δεν εντοπίζεται πρόβλημα κανονικότητας.




*Εικόνα 22: Ιστόγραμμα υπολοίπων ανάλυσης δεδομένων ανά περιφέρεια.*

Επίσης από το γράφημα των υπολοίπων προκύπτει ότι το μοντέλο είναι σωστά καθορισμένο αφού τα υπόλοιπα σε σχέση με την προβλεπόμενη εξαρτημένη μεταβλητή είναι τυχαία.




*Εικόνα 23: Κατανομή υπολοίπων ανάλυσης δεδομένων ανά περιφέρεια.*

## Geographically Weighted Regression - Ανάλυση Δεδομένων ανά ΤΚ

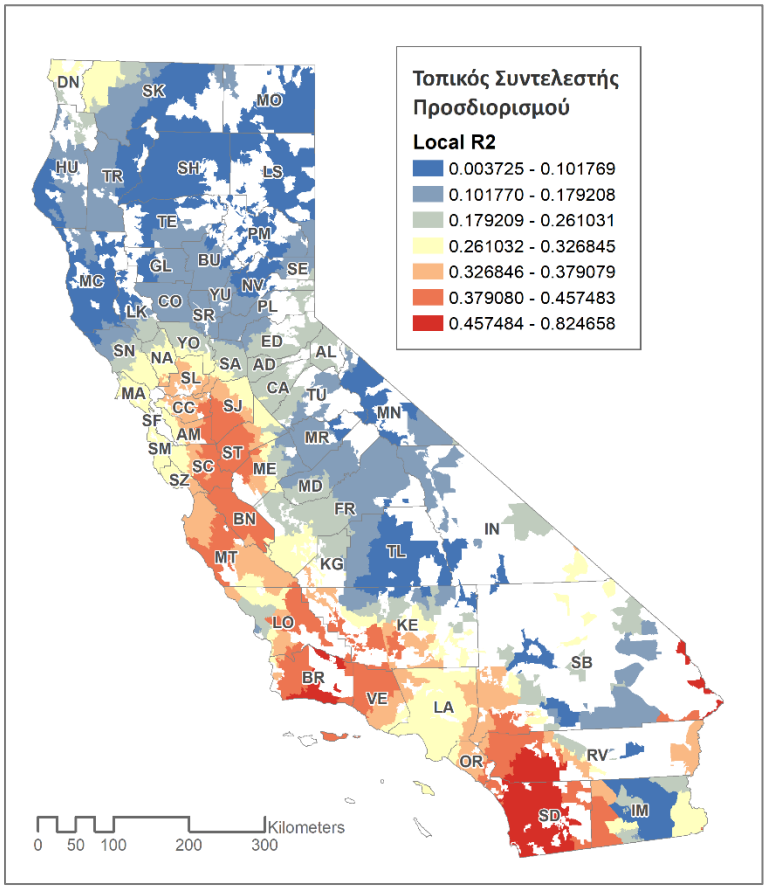
Η επόμενη δοκιμή για την ανάλυση των δεδομένων ήταν με τη μέθοδο της Γεωγραφικά Σταθμισμένης Παλινδρόμησης, με τη πληροφορία να είναι ανά Ταχυδρομικό Κώδικα. Η GWR πραγματοποιήθηκε με σταθερή απόσταση και το εύρος ζώνης με AICc. Ο πίνακας που ακολουθεί περιέχει τα αποτελέσματα της παλινδρόμησης.

*Πίνακας 3: Αποτελέσματα GWR για τη παχυσαρκία ανά ΤΚ με fixed kernel.*

| | OBJECTID * | VARNAME | VARIABLE | DEFINITION |
|---|---|---|---|---|
| ▸ | 1 | Bandwidth | 0.718259 | |
| | 2 | ResidualSquares | 2.962548 | |
| | 3 | EffectiveNumber | 112.263887 | |
| | 4 | Sigma | 0.042467 | |
| | 5 | AICc | -6050.464242 | |
| | 6 | R2 | 0.449445 | |
| | 7 | R2Adjusted | 0.412156 | |
| | 8 | Dependent Field | 0 | obes_rate |
| | 9 | Explanatory Field | 1 | gyms_area |
| | 10 | Explanatory Field | 2 | High_Education |
| | 11 | Explanatory Field | 3 | B19019e1 |

Η τιμή του προσαρμοσμένου συντελεστή προσδιορισμού $R^2$ είναι 0.412156. Οι χαμηλές τιμές των Residual Squares, Sigma και AICc υποδεικνύουν ένα μοντέλο καλά προσαρμοσμένο στα παρατηρούμενα δεδομένα με μικρή τυπική απόκλιση των υπολοίπων.

Από τον Attribute Table της feature class που δημιουργήθηκε από το ArcMap προκύπτει ότι η μεγαλύτερη Condition τιμή είναι 11.76341 < 30, από το οποίο συνεπάγεται ότι δεν εντοπίζεται πρόβλημα πολυκεντρικότητας και η μεγαλύτερη τιμή του τοπικού συντελεστή προσδιορισμού $R^2$ είναι 0.824658. Στον χάρτη της Εικόνας 24 παρουσιάζονται όλα τα τοπικά $R^2$ της περιοχής μελέτης.



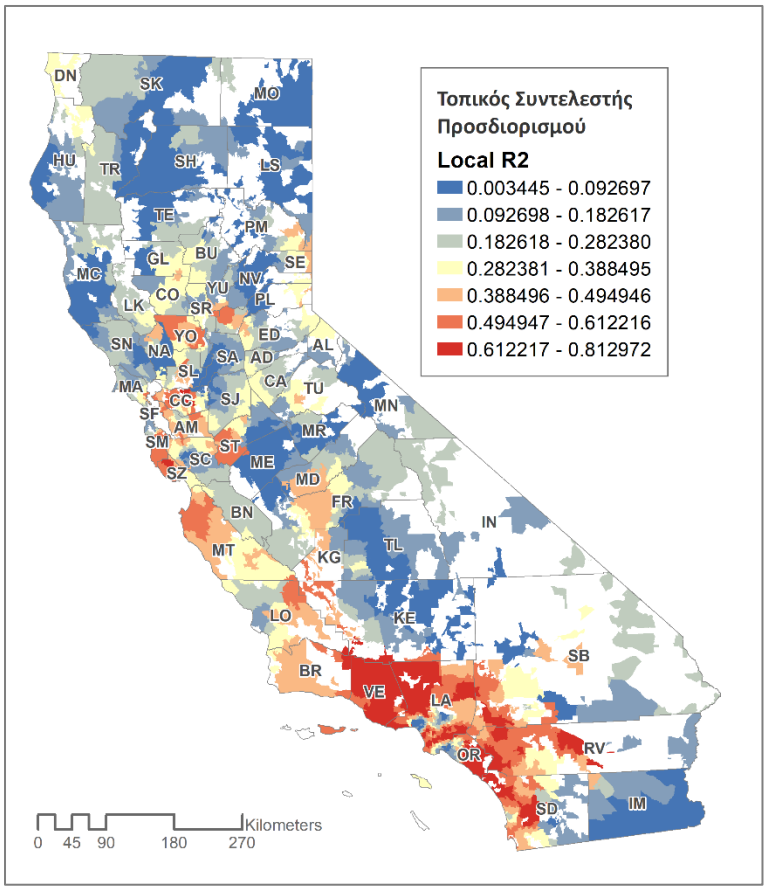

*Εικόνα 24: Τοπικός συντελεστής προσδιορισμού με τη μέθοδο GWR ανά ΤΚ.*

Τα αποτελέσματα με προσαρμοστικό πυρήνα (adaptive kernel) επηρεάζονται από 51 γείτονες και έχουν adjusted $R^2 = 0.558102$.

*Πίνακας 4: Αποτελέσματα GWR για τη παχυσαρκία ανά ΤΚ με adaptive kernel.*

| | OBJECTID * | VARNAME | VARIABLE | DEFINITION |
|---|---|---|---|---|
| ▸ | 1 | Neighbors | 51 | |
| | 2 | ResidualSquares | 1.848462 | |
| | 3 | EffectiveNumber | 391.506667 | |
| | 4 | Sigma | 0.03682 | |
| | 5 | AICc | -6318.884085 | |
| | 6 | R2 | 0.656485 | |
| | 7 | R2Adjusted | 0.558102 | |
| | 8 | Dependent Field | 0 | obes_rate |
| | 9 | Explanatory Field | 1 | gyms_area |
| | 10 | Explanatory Field | 2 | High_Education |
| | 11 | Explanatory Field | 3 | B19019e1 |

Οι τιμές του Condition Number αυξήθηκαν με την μεγαλύτερη να είναι πλέον 26.568137 < 30 και η μέγιστη τιμή του τοπικού συντελεστή προσδιορισμού $R^2$ είναι 0.812972.




*Εικόνα 25: Τοπικός συντελεστής προσδιορισμού με τη μέθοδο GWR ανά περιφέρεια.*

Geographically Weighted Regression - Ανάλυση Δεδομένων ανά Περιφέρεια

Η GWR δεν ενδείκνυται για τη χωρική ανάλυση των δεδομένων, όταν το πλήθος αυτών είναι μικρό. Πράγματι, παρατηρήθηκε ότι με χωρισμό της Καλιφόρνια ανά περιφέρεια (58 στο σύνολο), με τη μέθοδο του προσαρμοσμένου πυρήνα, στην παλινδρόμηση λαμβάνουν μέρος 58 γείτονες, επομένως για κάθε πυρήνα συμμετέχουν όλες οι περιφέρειες της πολιτείας.

*Πίνακας 5 : Αποτελέσματα GWR για τη παχυσαρκία ανά περιφέρεια με adaptive kernel.*

| | OBJECTID * | VARNAME | VARIABLE | DEFINITION |
|---|---|---|---|---|
| ▸ | 1 | Neighbors | 58 | |
| | 2 | ResidualSquares | 0.007903 | |
| | 3 | EffectiveNumber | 5.985296 | |
| | 4 | Sigma | 0.012326 | |
| | 5 | AICc | -337.518962 | |
| | 6 | R2 | 0.833936 | |
| | 7 | R2Adjusted | 0.81802 | |
| | 8 | Dependent Field | 0 | obes_rate |
| | 9 | Explanatory Field | 1 | gyms_area |
| | 10 | Explanatory Field | 2 | High_Education |
| | 11 | Explanatory Field | 3 | B19019e1_1 |

Ωστόσο, το μοντέλο παρουσιάζει προσαρμογή και εξηγείται το 82% της παχυσαρκίας από τις επεξηγηματικές μεταβλητές, με τους τοπικούς συντελεστές να κυμαίνονται από 0.802499 έως 0.840384.

## 4.3 Συμπεράσματα Ανάλυσης της Παχυσαρκίας

Από την στατιστική και χωρική ανάλυση των ποσοστών παχυσαρκίας στην πολιτεία της Καλιφόρνια των ΗΠΑ, προκύπτει ότι οι εξεταζόμενες μεταβλητές που επιλέχθηκαν επηρεάζουν με αρνητικό πρόσημο την εξαρτημένη μεταβλητή.

Είναι φανερό πως ο χώρος αποτελεί κύριο παράγοντα συμπεριφοράς και επιρροής σε θέματα που αφορούν την ιατρική. Με τις κλασσικές μεθόδους στατιστικής ανάλυσης ο παράγοντας αυτός δεν είναι δυνατό να συνυπολογιστεί. Αντιθέτως, με την ανάλυση της Γεωγραφικά Σταθμισμένης Παλινδρόμησης λαμβάνονται υπόψη τα ποσοστά των πλησιέστερων περιοχών, τα οποία είναι ικανά να διαφοροποιήσουν σε μεγάλο βαθμό τα αποτελέσματα μιας χωρικής ανάλυσης.

Στην περίπτωση της παχυσαρκίας τα δεδομένα που χρησιμοποιήθηκαν ήταν ανά ταχυδρομικό κώδικα, και επομένως η διαθέσιμη πληροφορία ήταν αρκετά λεπτομερείς. Καταλληλότερο

μοντέλο ανάλυσης προκύπτει με τη μέθοδο GWR με προσαρμοσμένο χωρικό πυρήνα, όπου το μοντέλο εξηγεί περίπου το 56% της διακύμανσης της εξαρτημένης μεταβλητής.

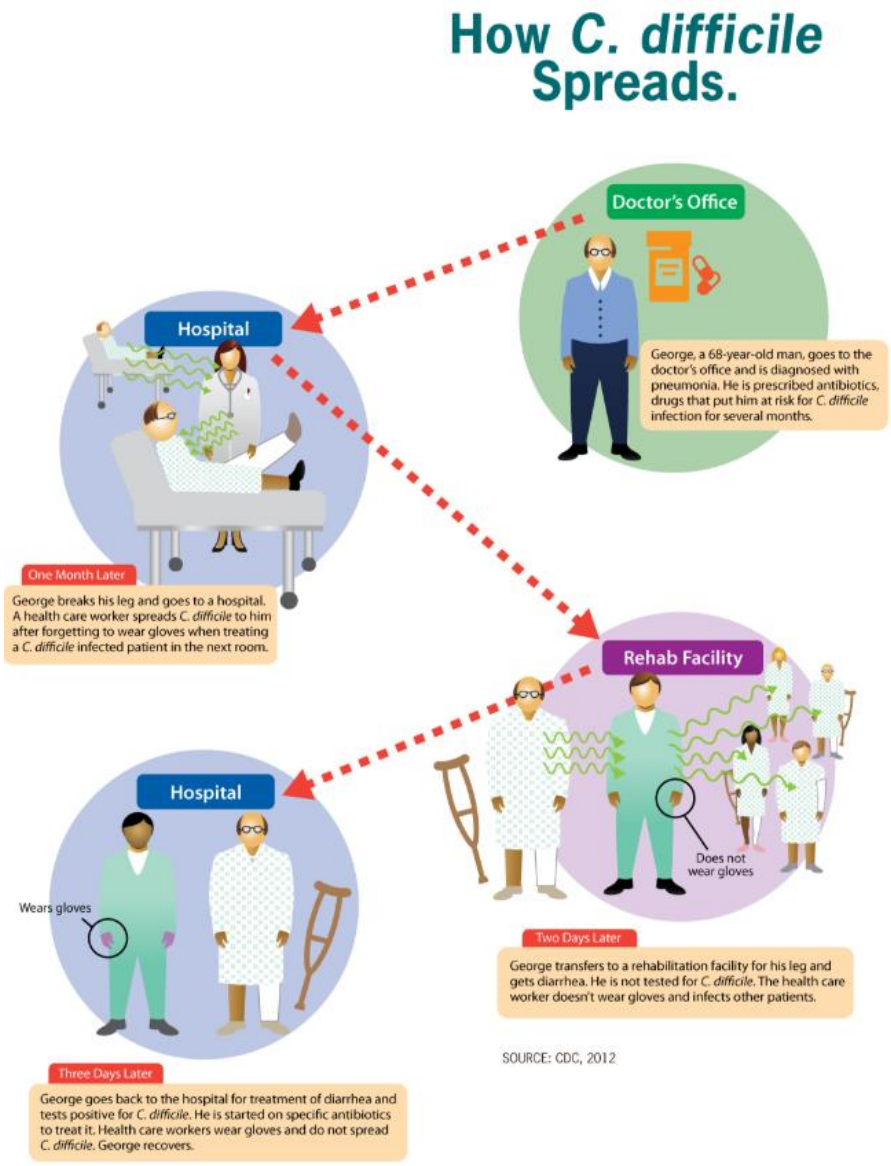
Είναι σαφές ότι με την αλλαγή κλίμακας των δεδομένων, από το πιο ειδικό στο πιο γενικό, δηλαδή αναγάγοντας τα δεδομένα από το ΤΚ στη περιφέρεια, χάνεται σημαντικό μέρος της ανάλυσης. Στη περίπτωση ανάλυσης ανά περιφέρεια καλύτερο μοντέλο αυτό που προέκυψε με τη μέθοδο OLS, ενώ η GWR παρατηρήθηκε ότι δεν ενδείκνυται για ανάλυση μικρού αριθμού δεδομένων.

# Κεφάλαιο 5ο - Clostridium Difficile Infection

Το βακτήριο *Clostridium Difficile* (C. diff) προέρχεται από την ελληνική λέξη “κλωστήρ” και τη λατινική “difficile” που σημαίνει δύσκολος. Το βακτήριο C. diff παράγει μεγάλο αριθμό τοξινών και περιγράφηκε για πρώτη φορά το 1935. Θεωρείται μία από τις πιο κοινές αιτίες μολύνσεως στο παχύ έντερο και μπορεί να προκαλέσει ακόμη και ψευδομεμβρανώδη κολίτιδα ή διάτρηση παχέος εντέρου. Τα συμπτώματα περιλαμβάνουν διάρροια, πυρετό, ναυτία και κοιλιακό άλγος [65].

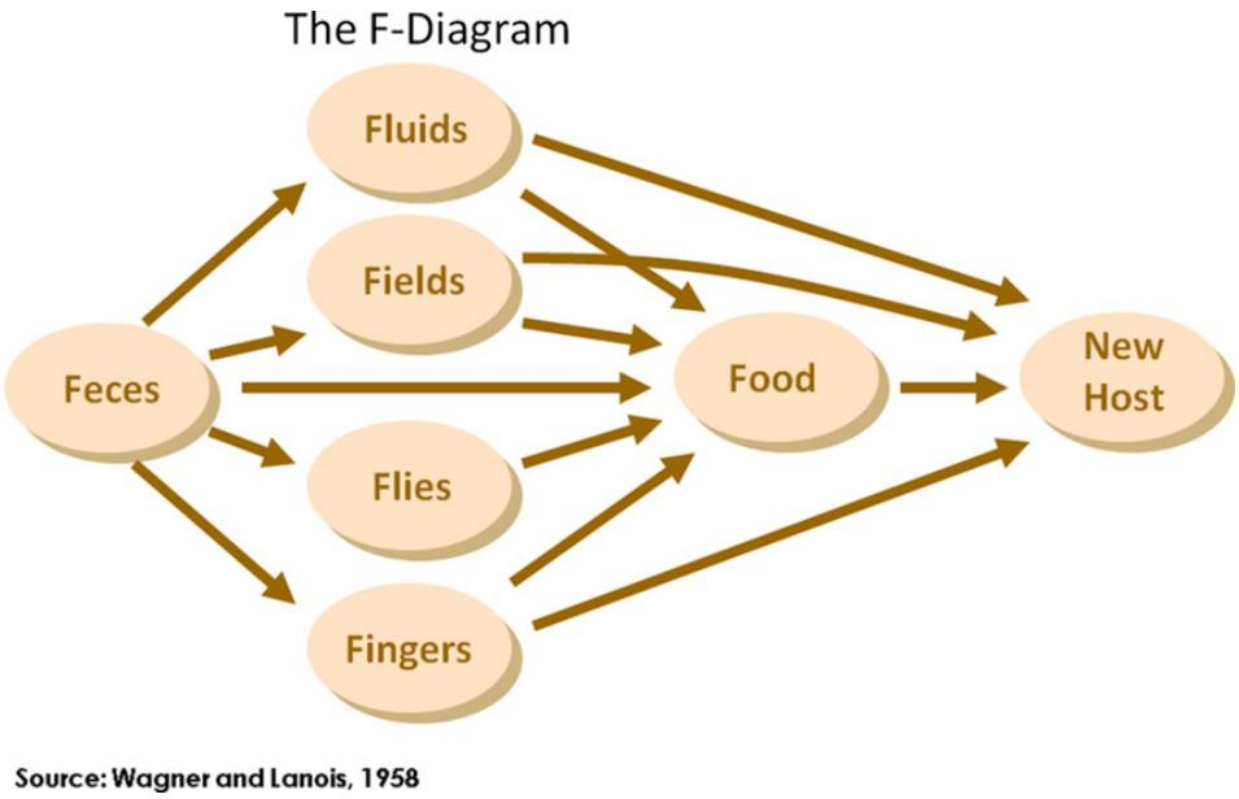
Η λοίμωξη από *Clostridium Difficile* είναι η πιο συνηθισμένη ενδονοσοκομειακή λοίμωξη στις ΗΠΑ. Το C.diff αποτελεί περίπου το 20% των περιπτώσεων διάρροιας που σχετίζονται με αντιβιοτικά, ενώ μπορεί να βρεθεί και σε άτομα που δεν έχουν νοσήσει. Ωστόσο, συμπληρωματικός με τη χρήση αντιβιοτικών παράγοντας κινδύνου είναι και το νοσοκομειακό περιβάλλον. Παρότι η μετάδοση του C.diff πραγματοποιείται πιο εύκολα στα νοσοκομεία, διότι οι υψηλότερες συγκεντρώσεις αυτών των βακτηρίων βρίσκονται σε νοσηλευόμενους ασθενείς που λαμβάνουν αντιβιοτικά, βακτήρια μεταδίδονται μέσω άμεσης ή έμμεσης επαφής με μολυσμένα αντικείμενα από άτομο σε άτομο [65].

Η συνύπαρξη σε δωμάτιο νοσοκομείου με ασθενή με θετική καλλιέργεια, ακόμη και αν είναι ασυμπτωματικός, θέτει τον συνοσηλευόμενο σε υψηλότερο κίνδυνο. Περισσότεροι από το 8% των εισαγόμενων ασθενών είναι φορείς του C.diff, και έχουν σχεδόν 6 φορές υψηλότερο κίνδυνο εμφάνισης λοίμωξης ακόμη και 3 μήνες μετά τη νοσηλεία τους [66]. Επίσης, εκτιμάται ότι το 0.9% του γενικού νοσοκομειακού πληθυσμού πάσχει από CDI (*Clostridium Difficile* Infection), ενώ στις Μονάδες Εντατικής Θεραπείας (ΜΕΘ) το ποσοστό εκτιμάται στο 2% [67]. Οι εργαζόμενοι στον τομέα της υγείας είναι δυνατό να συμβάλλουν στην εξάπλωση της νόσου, αφού τα χέρια αποτελούν κύρια εστία μόλυνσης. Είναι, επομένως, προφανές πως υπάρχουν περιβαλλοντικοί παράγοντες, οι οποίοι συμβάλλουν στην εξάπλωση του βακτηρίου C.diff.




*Εικόνα 26: Μετάδοση του βακτηρίου CDI [68].*

Τα ανθρώπινα και ζωικά περιττώματα είναι η κύρια πηγή διαρροϊκών παθογόνων. Το διάγραμμα F καθορίστηκε το 1958 από τον Παγκόσμιο Οργανισμό Υγείας ( World Health Organization – WHO) και πήρε το όνομά του από το αρχικό γράμμα "F" των λέξεων που το αποτελούν στην αγγλική γλώσσα (fluids, fingers, flies, fields, food, fomites) [69]. Το διάγραμμα F εξηγεί τις οδούς μετάδοσης της ασθένειας, και συγκεκριμένα με επίκεντρο τα κόπρανα πως μεταφέρονται τα βακτήρια, μέσω υγρών (κυρίως μέσω του νερού), επιφανειών, δακτύλων και μυγών στο φαγητό και με τελικούς αποδέκτες άλλους ανθρώπους [70].




*Εικόνα 27: Τρόποι Μεταφοράς Νόσων μέσω του Διαγράμματος F [71].*

Κάθε χρόνο περίπου μισό εκατομμύριο ασθένειες και 15.000 θάνατοι στις ΗΠΑ προκαλούνται από λοιμώξεις από *Clostridium Difficile [72]*. Στην πολιτεία της Καλιφόρνια, για το έτος 2015, καταγράφηκαν 10.762 κρούσματα CDI.

## 5.1 Περιγραφή Δεδομένων

Για τη χωρική ανάλυση των κρουσμάτων από το βακτήριο *Clostridium Difficile* χρησιμοποιήθηκαν δεδομένα, τα οποία διατίθενται ελεύθερα στο διαδίκτυο.

### 5.1.1 Κρούσματα CDI στα Νοσοκομεία

Πρόκειται για ένα σύνολο δεδομένων που διατίθεται ελεύθερα από τον Οργανισμό Υγείας και Ανθρώπινων Υπηρεσιών της Καλιφόρνια (California Health and Human Services Agency - CHHS) σε μορφή csv (https://data.chhs.ca.gov/dataset/clostridium-difficile-infections-cdi-in-healthcare) και βασίζεται σε δεδομένα του Εθνικού Κέντρου Ελέγχου και Πρόληψης Νοσημάτων (National Healthcare Safety Network - NHSN), που αφορούν περιπτώσεις εμφάνισης λοίμωξης

C. Diff που σημειώθηκαν σε νοσηλευόμενους μέσα σε νοσοκομεία της Καλιφόρνια για το έτος 2015, τρεις ήμερες μετά την εισαγωγή τους. Τα δεδομένα αφορούν 362 νοσοκομεία εκ των οποίων 27 είχαν ατελές report και δεν συμπεριλήφθηκαν στην ανάλυση.

### 5.1.2 Νοσοκομειακές Εγκαταστάσεις

Πρόκειται για ένα αρχείο που διατίθεται ελεύθερα από το CHHS σε μορφή csv (https://data.chhs.ca.gov/dataset/healthcare-facility-locations), το οποίο απαριθμεί 10,381 εγκαταστάσεις υγειονομικής περίθαλψης της Καλιφόρνια που λειτουργούν και διαθέτουν τρέχουσα άδεια που εκδίδεται από το Τμήμα Δημόσιας Υγείας της Καλιφόρνια (California Department of Public Health - CDPH). Από τις πληροφορίες που περιλαμβάνονται χρησιμοποιήθηκαν οι γεωγραφικές συντεταγμένες και η χωρητικότητα κάθε νοσοκομείου.

### 5.1.3 Πολύγωνα HSA

Τα HSA (Hospital Service Areas) πολύγωνα αντικατοπτρίζουν τις περιοχές με αυτοτελής νοσοκομειακή περίθαλψη, όπως ορίσθηκαν από το Κέντρο Ελέγχου και Πρόληψης Νοσημάτων (Centers for Disease Control and Prevention) και συγκεκριμένα από το Εθνικό Κέντρο Στατιστικών Υγείας (National Center for Health Statistics). Τα πολύγωνα HSA διατίθενται ελεύθερα από το Dartmouth Atlas Project (http://www.dartmouthatlas.org/tools/downloads.aspx?tab=39) και στην πολιτεία της Καλιφόρνια αντιστοιχούν 222 πολύγωνα.

### 5.1.4 Δημογραφικά Στοιχεία

Από την Υπηρεσία Απογραφής των Ηνωμένων Πολιτειών (United States Census Bureau) χρησιμοποιήθηκαν οι γεωβάσεις TIGER/Line® ACS_2015_5YR_BG_06_CALIFORNIA.gdb και TIGER/Line® ACS_2015_5YR_ZCTA.gdb, οι οποίες περιλαμβάνουν δημογραφικά στοιχεία ανά Block και ανά Zip code αντίστοιχα (https://www.census.gov/cgi-bin/geo/shapefiles/index.php).

### 5.1.5 Κλιματολογικά Στοιχεία

Πρόκειται για ένα σύνολο δεδομένων που διατίθεται ελεύθερα από το Αμερικάνικο Εθνικό Κέντρο Περιβαλλοντικών Πληροφοριών (National Centers for Environmental Information - NOAA) σε μορφή csv (https://www.ncdc.noaa.gov/cdo-web/datasets), και περιέχει ανεπεξέργαστη πληροφορία για περισσότερους από 25,000 μετεωρολογικούς σταθμούς της

Αμερικής. Για την παρούσα εργασία χρησιμοποιήθηκαν πληροφορίες από τους μετεωρολογικούς σταθμούς της Καλιφόρνια για το έτος 2015 (502 εγγραφές).

### 5.1.6 Μονάδες Νοσοκομειακής Περίθαλψης Ηλικιωμένων

Τα δεδομένα που αφορούν τα Nursing Homes, τα οποία επί της ουσίας πρόκεινται για μονάδες περίθαλψης ηλικιωμένων, διατίθενται από το Center for Medicare & Medicaid Services (https://data.medicare.gov/data/nursing-home-compare) σε μορφή csv. Η πληροφορία που περιλαμβάνει αφορά τη θέση και την χωρητικότητα των χώρων αυτών σε κρεβάτια. Το πλήθος των Nursing homes για την Καλιφόρνια είναι 1200.

## 5.2 Μεθοδολογικό Πλαίσιο

Για την ανάλυση της λοίμωξης από *Clostridium Difficile* ακολουθήθηκαν τέσσερις στρατηγικές. Στις δύο πρώτες η ανάλυση πραγματοποιήθηκε ως προς τη θέση των νοσοκομείων, δηλαδή λαμβάνοντας υπόψη σημειακές οντότητες, ενώ στις δύο επόμενες η ανάλυση έγινε με τη χρήση πολυγώνων. Τα πολύγωνα που αναλύθηκαν ήταν βάσει HSA στη πρώτη περίπτωση και στη δεύτερη δημιουργήθηκαν πολύγωνα Thiessen.

**Στρατηγική 1η**

Έχοντας το κάθε νοσοκομείο ως σημειακή οντότητα, υπολογίστηκε ο πληθυσμός, από τη βάση των δημογραφικών στοιχείων ανά Ταχυδρομικό Κώδικα, που βρίσκεται στο HSA στο οποίο ανήκει το κάθε νοσοκομείο. Το ποσοστό των κρουσμάτων CDI υπολογίστηκε ως το πλήθος των κρουσμάτων ανά νοσοκομείο προς τον συνολικό πληθυσμό του HSA στο οποίο ανήκει το νοσοκομείο.

Για τον έλεγχο επιρροής της θερμοκρασίας χρησιμοποιήθηκαν στοιχεία από τον κοντινότερο μετεωρολογικό σταθμό του κάθε νοσοκομείου.

Τα Nursing Homes υπολογίστηκαν με το εργαλείο Generate Near Table του ArcMap με ακτίνα επιρροής τα 10χλμ από τα νοσοκομεία. Με τον ορισμό αυτής της ακτίνας θεωρήθηκε επί της ουσίας ότι Nursing Homes τα οποία βρίσκονται πιο μακριά δεν ασκούν επιρροή. Από τον πίνακα αποτελεσμάτων προέκυψε ότι σε ακτίνα 10χλμ υπάρχουν νοσοκομεία που έχουν από κανένα έως 76 Nursing Homes. Για τον λόγο αυτό προστέθηκαν βάρη με αντίστροφη απόσταση ώστε να

εξεταστεί εάν η επιρροή σχετίζεται με το πόσο κοντά βρίσκονται στα νοσοκομεία. Οι σχέσεις που χρησιμοποιήθηκαν ήταν:

$$InvDist = 1 - \frac{dist}{10000} \quad (5.1)$$

$$weight = \sum_i (InvDist) \quad (5.2)$$

,για όλα τα $InvDist$ των Nursing Homes που βρίσκονται μέσα στην ακτίνα επιρροής κάθε νοσοκομείου.

Το μειονέκτημα αυτής της μεθόδου έγκειται στο ότι ο πληθυσμός που χρησιμοποιήθηκε για την ανάλυση δεν θεωρείται ότι θα διασπαστεί σε όλα τα νοσοκομεία, δηλαδή νοσοκομεία που ανήκουν στο ίδιο πολύγωνο HSA έχουν τον ίδιο πληθυσμό. Στην πραγματικότητα όμως ο πληθυσμός θα διαιρεθεί και αυτό συνεπάγεται ότι τα ποσοστά των κρουσμάτων CDI τελικά είναι μεγαλύτερα για τα νοσοκομεία που μοιράζονται ένα HSA. Το πρόβλημα αυτό δεν επηρεάζει τα πολύγωνα HSA τα οποία έχουν μόνο ένα νοσοκομείο.

**Στρατηγική 2η**

Επειδή παρατηρήθηκε η άνιση κατανομή πληθυσμού στην πρώτη δοκιμή, στη δεύτερη αποφασίστηκε να χωριστεί ο πληθυσμός ανάλογα με το μέγεθος του κάθε νοσοκομείου. Με τον τρόπο αυτό θεωρήθηκε ότι μεγαλύτερα νοσοκομεία θα εξυπηρετούν περισσότερους ασθενείς και αντίστοιχα μικρότερα νοσοκομεία λιγότερους ασθενείς. Ο πληθυσμός με βάρος τη χωρητικότητα και το μέγεθος του νοσοκομείου υπολογίστηκε από τη σχέση:

$$\mathrm{Hospital}_{\mathrm{population}} = \frac{\mathrm{populationHSA}}{\mathrm{capacityHSA}} * \mathrm{capacityHospital} \quad (5.3)$$

όπου, populationHSA ο συνολικός πληθυσμός που ανήκει στο κάθε πολύγωνο HSA, capacityHSA ο συνολικός αριθμός των κρεβατιών όλων των νοσοκομείων του HSA και capacityHospital ο διαθέσιμος αριθμός κλινών του κάθε νοσοκομείου. Το ποσοστό κρουσμάτων CDI ανά 10,000 άτομα, είναι ανάλογο πλέον της χωρητικότητας του νοσοκομείου και δίνεται από τον τύπο:

$$\mathrm{CDI}_{\mathrm{analog}} = \frac{O_{Hospital}}{Hospital_{population}} * 10000 \quad (5.4)$$

όπου, O_Hospital ο καταγεγραμμένος αριθμός κρουσμάτων CDI του κάθε νοσοκομείου.

Οι καιρικές συνθήκες κατά τη διάρκεια του έτους που χρησιμοποιήθηκαν ήταν και σε αυτή τη περίπτωση εκείνες από τον κοντινότερο μετεωρολογικό σταθμό.

Για τα Nursing Homes χρησιμοποιήθηκε ο ίδιος πίνακας που δημιουργήθηκε στην Στρατηγική 1 με βάρη ανάλογα της ανάστροφης απόστασης.

**Στρατηγική 3η**

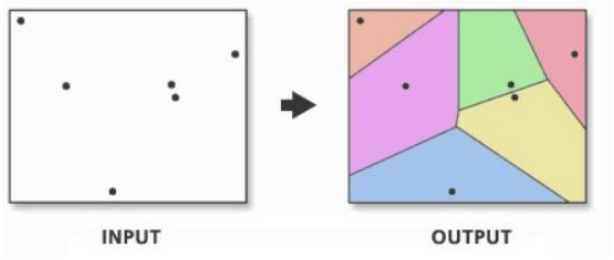
Στην παρούσα ανάλυση χρησιμοποιήθηκαν τα πολύγωνα HSA και αθροίστηκαν όλα τα περιστατικά λοίμωξης C. Diff που σημειώθηκαν σε νοσηλευόμενους των νοσοκομείων που ανήκουν στο κάθε HSA καθώς και οι κλίνες των νοσοκομείων. Για τον υπολογισμό του ποσοστού των κρουσμάτων ως προς τον πληθυσμό, το άθροισμα των περιστατικών διαιρέθηκε με τον συνολικό πληθυσμό που αντιστοιχεί στο κάθε πολύγωνο HSA.

Τα κλιματολογικά στοιχεία που χρησιμοποιήθηκαν προέκυψαν από τον μέσο όρο των στοιχειών που κατέγραψαν όλοι οι μετεωρολογικοί σταθμοί που βρίσκονται στο εκάστοτε HSA πολύγωνο.

Για την εξέταση επιρροής των Nursing Homes στο ποσοστό κρουσμάτων CDI, θεωρήθηκε ως μεταβλητή ο διαθέσιμος αριθμός κλινών των Μονάδων Νοσοκομειακής Περίθαλψης Ηλικιωμένων και συγκεκριμένα το άθροισμα όλων των κλινών που ανήκουν στο ίδιο πολύγωνο προς τον συνολικό πληθυσμό.

**Στρατηγική 4η**

Στην τέταρτη στρατηγική δημιουργήθηκαν πολύγωνα Thiessen. Σύμφωνα με τη μέθοδο αυτή η περιοχή μελέτης διασπάται σε υποπεριοχές, τα όρια των οποίων προκύπτουν από τις μεσοκαθέτους στις ευθείες που ενώνουν δύο παρακείμενα σημεία. Τα σημεία αυτά είναι τα νοσοκομεία και οι υποπεριοχές δημιουργούνται με τέτοιο τρόπο ώστε η κάθε μία να περιλαμβάνει ένα και μόνο νοσοκομείο. Χαρακτηριστικό των νέων πολυγώνων που δημιουργούνται είναι ότι το κάθε ένα περιλαμβάνει όλα τα σημεία που βρίσκονται πιο κοντά στο κάθε νοσοκομείο, παρά σε οποιοδήποτε άλλο νοσοκομείο [73].




*Εικόνα 28: Δημιουργία πολυγώνων Thiessen [73].*

Η πληροφορία που μέχρι τώρα χρησιμοποιούνταν για τον υπολογισμό του πληθυσμού αφορούσε δεδομένα ανά ταχυδρομικό κώδικα. Με τη δημιουργία των πολυγώνων Thiessen παρατηρείται πως κάποια πολύγωνα είναι μικρότερα από τα πολύγωνα του ΤΚ, με αποτέλεσμα κάποιες υποπεριοχές να μην περιέχουν την πληροφορία των δημογραφικών στοιχείων. Το πρόβλημα αυτό ξεπεράστηκε με τη χρήση της γεωβάσης που διατίθεται από την Υπηρεσία Απογραφής των Ηνωμένων Πολιτειών και περιλαμβάνει τα δημογραφικά στοιχεία ανά Οικοδομικό Τετράγωνο.

### 5.2.1 Χωρική και Στατιστική Ανάλυση CDI

Όμοια με το προηγούμενο κεφάλαιο που αφορούσε την παχυσαρκία, για την στατιστική και τη χωρική ανάλυση των κρουσμάτων λοίμωξης C. Diff χρησιμοποιήθηκαν τα εργαλεία παλινδρόμησης του ArcMap, Exploratory Regression, OLS Regression και GWR. Ως εξαρτημένη μεταβλητή θεωρήθηκε το ποσοστό κρουσμάτων CDI προς τον αντίστοιχο πληθυσμό στην κάθε περίπτωση, όπως περιγράφηκε, και ως ανεξάρτητες θεωρήθηκαν η χωρητικότητα της κάθε νοσοκομειακής εγκατάστασης, ο μέσος όρος της μέσης, ανώτερης ή χαμηλότερης θερμοκρασίας, η απόσταση ή το πλήθος των κλινών των Μονάδων Νοσοκομειακής Περίθαλψης Ηλικιωμένων.

**Στρατηγική 1η**

Αρχικά πραγματοποιήθηκε ανάλυση ως προς τις σημειακές οντότητες που αντιστοιχούν στις νοσοκομειακές εγκαταστάσεις και υπολογίστηκε ο πληθυσμός που εξυπηρετείται από αυτές ως το σύνολο του πληθυσμού που βρίσκεται στο HSA πολύγωνο.

Exploratory Regression

Με τη χρήση αυτού του εργαλείου έγινε η αναζήτηση εκείνων των μεταβλητών που επηρεάζουν περισσότερο στην αύξηση των κρουσμάτων CDI. Τα αποτελέσματα με κάθε μία μεταβλητή ξεχωριστά παρουσιάζονται στον Πίνακα 6 που ακολουθεί.

Πίνακας 6 : Προσαρμοσμένος συντελεστής προσδιορισμού με Exploratory Regression για CDI (Στρατηγική 1η).

| **Ανεξάρτητη Μεταβλητή** | **Adjusted $R^2$** |
|---|---|
| Νοσοκομειακές κλίνες ανά πληθυσμό | +20 % |
| Μέσος όρος μέσης θερμοκρασίας | +2 % |
| Ανάστροφη Απόσταση από Nursing Homes | 0 % |

Παρατηρείται ότι η χωρητικότητα των νοσοκομείων εξηγεί το 20% της διακύμανσης των κρουσμάτων και έχει θετικό πρόσημο, κάτι που όπως αναλύθηκε στην αρχή του παρόντος κεφαλαίου είναι αποτέλεσμα της μεγαλύτερης επισκεψημότητας του νοσοκομείου και του κινδύνου μετάδοσης της λοίμωξης από συνοσηλευόμενους ασθενείς. Θετικό πρόσημο έχει και ο μέσος όρος της μέσης θερμοκρασίας με τιμή 2%, το οποίο σημαίνει ότι η αύξηση της θερμοκρασίας τείνει να αυξάνει τα κρούσματα. Η μηδενική τιμή της μεταβλητής που αντιστοιχεί στην ανάστροφη απόσταση των Nursing Homes από τα νοσοκομεία, σημαίνει ότι η μεταβλητή δεν παίρνει μέρος στη παλινδρόμηση και επομένως τα κρούσματα δεν επηρεάζονται από το πόσο κοντά βρίσκονται οι εγκαταστάσεις που φιλοξενούν και περιθάλπουν ηλικιωμένους.

**Στρατηγική 2η**

Στη παρούσα θεωρήθηκε ότι ο πληθυσμός θα χωριστεί ανάλογα με το μέγεθος των νοσοκομείων. Τα μεγαλύτερα νοσοκομεία θα εξυπηρετήσουν περισσότερους ασθενείς από τα μικρότερα.

Exploratory Regression

Τα ποσοστά που προέκυψαν, για κάθε μία μεταβλητή ξεχωριστά, με τη χρήση του διερευνητικού εργαλείου παλινδρόμησης με εξαρτημένη μεταβλητή το αναλογικό ποσοστό κρουσμάτων CDI σε σχέση με το μέγεθος του νοσοκομείου παρουσιάζονται στον Πίνακα 7 που ακολουθεί.

*Πίνακας 7: Προσαρμοσμένος συντελεστής προσδιορισμού με Exploratory Regression για CDI (Στρατηγική 2η ).*

| **Ανεξάρτητη Μεταβλητή** | **Adjusted $R^2$** |
|---|---|
| Νοσοκομειακές κλίνες ανά πληθυσμό | +4 % |
| Ανάστροφη Απόσταση από Nursing Homes | +4 % |
| Μέσος όρος χαμηλότερης θερμοκρασίας | +5 % |
| Μέσος όρος μέσης θερμοκρασίας | +4% |
| Πλήθος Nursing Homes σε ακτίνα 10χλμ | +4% |

Το μοντέλο που εξηγεί καλύτερα την εξαρτημένη μεταβλητή, δηλαδή το ποσοστό κρουσμάτων λοίμωξης C. diff είναι το "Νοσοκομειακές κλίνες ανά πληθυσμό - Μέσος όρος χαμηλότερης θερμοκρασίας - Ανάστροφη Απόσταση από Nursing Homes", το οποίο εξηγεί το 11% της διακύμανσης σε επίπεδο εμπιστοσύνης 95%. Και οι τρεις ανεξάρτητες μεταβλητές έχουν θετικό πρόσημο, το οποίο συνεπάγεται αύξηση του CDI με την αύξησή τους. Σε σχέση με τις Μονάδες Νοσοκομειακής Περίθαλψης Ηλικιωμένων προκύπτει ότι όσο πιο κοντά βρίσκεται μία τέτοια εγκατάσταση σε ένα νοσοκομείο τόσο μεγαλύτερο ποσοστό κρουσμάτων τείνει να έχει το νοσοκομείο.

OLS Regression

Στη συνέχεια πραγματοποιήθηκε στατιστική ανάλυση με τη μέθοδο γραμμικής παλινδρόμησης με την τεχνική Ordinary Least Squares (OLS). Με εξαρτημένη μεταβλητή το αναλογικό ποσοστό των κρουσμάτων σε σχέση με το μέγεθος των νοσοκομείων και ανεξάρτητες τις μεταβλητές που προέκυψαν ως καλύτερο μοντέλο από την Exploratory Regression.

Από το εξαγόμενο μήνυμα αναφοράς η θετική σχέση μεταξύ της εξαρτημένης μεταβλητής και των επεξηγηματικών μεταβλητών υποδηλώνεται από το Coefficient, ενώ η σημαντικότητα των ανεξάρτητων μεταβλητών από τη στήλη Probability και η μη ύπαρξη πολυσυγγραμμικότητας από την τιμή VIF, η οποία είναι κάτω από 7.5.

**Summary of OLS Results - Model Variables**

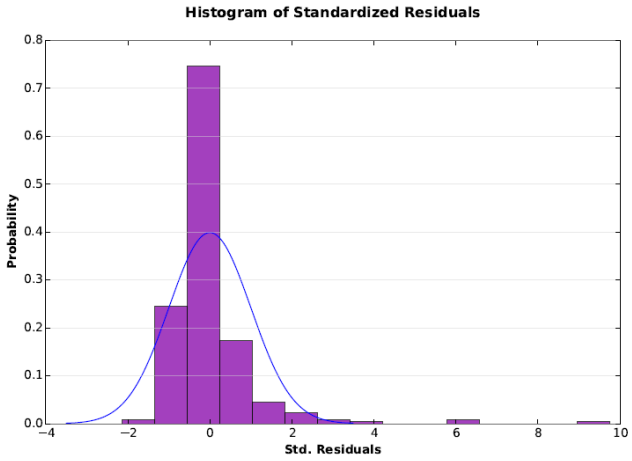
| Variable | Coefficient [a] | StdError | t-Statistic | Probability [b] | Robust_SE | Robust_t | Robust_Pr [b] | VIF [c] |
|---|---|---|---|---|---|---|---|---|
| Intercept | -0.538624 | 0.834311 | -0.645592 | 0.518991 | 0.617583 | -0.872148 | 0.383747 | -------- |
| CLIMA_TMIN | 0.216700 | 0.074698 | 2.901011 | 0.003974* | 0.055626 | 3.895663 | 0.000127* | 1.248473 |
| CAPACITY_POP | 0.027882 | 0.006247 | 4.462919 | 0.000014* | 0.013457 | 2.071926 | 0.039037* | 1.019527 |
| WEIGHTS | 0.090864 | 0.033931 | 2.677900 | 0.007774* | 0.028722 | 3.163628 | 0.001714* | 1.267534 |

Ο συντελεστής προσδιορισμού $R^2$ είναι 0.120405, ενώ η τιμή του προσαρμοσμένου $R^2$ είναι 0.112433.

**OLS Diagnostics**

| | | | |
|---|---|---|---|
| Input Features: | cdi_capa_pop | Dependent Variable: | CDI_ANALOG |
| Number of Observations: | 335 | Akaike's Information Criterion (AICc) [d]: | 1801.907694 |
| Multiple R-Squared [d]: | 0.120405 | Adjusted R-Squared [d]: | 0.112433 |
| Joint F-Statistic [e]: | 15.103182 | Prob(>F), (3,331) degrees of freedom: | 0.000000* |
| Joint Wald Statistic [e]: | 32.785596 | Prob(>chi-squared), (3) degrees of freedom: | 0.000000* |
| Koenker (BP) Statistic [f]: | 8.657230 | Prob(>chi-squared), (3) degrees of freedom: | 0.034213* |
| Jarque-Bera Statistic [g]: | 19004.107667 | Prob(>chi-squared), (2) degrees of freedom: | 0.000000* |

Ωστόσο ο συντελεστής Jarque-Bera καταδεικνύει πρόβλημα κανονικότητας των υπολοίπων, με το μοντέλο να είναι πιθανώς προκατειλημμένο.




*Εικόνα 29: Ιστόγραμμα υπολοίπων.*

Geographically Weighted Regression

Η επόμενη δοκιμή για την ανάλυση των δεδομένων της 2ης στρατηγικής, ήταν με τη μέθοδο της Γεωγραφικά Σταθμισμένης Παλινδρόμησης, με σταθερό πυρήνα και εύρος ζώνης με τη AICc μέθοδο. Τα αποτελέσματα της GWR δίνονται στον Πίνακα 8.

*Πίνακας 8: Αποτελέσματα GWR για CDI με fixed kernel (Στρατηγική 2η ).*

| OBJECTID * | VARNAME | VARIABLE | DEFINITION |
|---|---|---|---|
| 1 | Bandwidth | 1.356921 | |
| 2 | ResidualSquares | 3289.803686 | |
| 3 | EffectiveNumber | 32.293165 | |
| 4 | Sigma | 3.296658 | |
| 5 | AICc | 1771.660657 | |
| 6 | R2 | 0.298404 | |
| 7 | R2Adjusted | 0.225875 | |
| 8 | Dependent Field | 0 | CDI_analog |
| 9 | Explanatory Field | 1 | capacity_pop |
| 10 | Explanatory Field | 2 | weights |
| 11 | Explanatory Field | 3 | clima_TMIN |

Η τιμή του προσαρμοσμένου συντελεστή προσδιορισμού $R^2$ είναι 0.2255875 δηλαδή περίπου 23%. Παρατηρούνται μεγάλες τιμές των Residual Squares, Sigma και AICc, επομένως το μοντέλο δεν είναι καλά προσαρμοσμένο ως προς τα δεδομένα. Από τον Attribute Table της feature class που δημιουργήθηκε από το ArcMap προκύπτει ότι η μεγαλύτερη Condition τιμή είναι 27.78466 < 30, από το οποίο συνεπάγεται ότι δεν εντοπίζεται πρόβλημα πολυκεντρικότητας και η μέση τιμή του τοπικού συντελεστή προσδιορισμού είναι 0.20091 ενώ η μεγαλύτερη είναι 0.583904.

**Στρατηγική 3η**

Η παρούσα ανάλυση πραγματοποιήθηκε βάσει των πολυγώνων των HSAs και όχι βάσει των νοσοκομείων. Έγινε χωρική σύνδεση (Spatial Join) των νοσοκομείων που ανήκουν στο ίδιο πολύγωνο εξυπηρέτησης.

Exploratory Regression

Με εξαρτημένη μεταβλητή το ποσοστό των κρουσμάτων CDI έγινε έλεγχος των εξαρτημένων μεταβλητών που αφορούν τη χωρητικότητα των Νοσοκομείων και των Μονάδων Νοσοκομειακής Περίθαλψης Ηλικιωμένων καθώς και ο μέσος όρος της θερμοκρασίας των πολυγώνων.

*Πίνακας 9: Προσαρμοσμένος συντελεστής προσδιορισμού με Exploratory Regression για CDI (Στρατηγική 3η ).*

| **Ανεξάρτητη Μεταβλητή** | **Adjusted $R^2$** |
|---|---|
| Νοσοκομειακές κλίνες ανά πληθυσμό | +17 % |
| Κλίνες των Nursing Homes ανά πληθυσμό | +7 % |
| Μέσος όρος χαμηλότερης θερμοκρασίας | +8 % |
| Μέσος όρος μέσης θερμοκρασίας | +8% |

Το μοντέλο με το μεγαλύτερο προσαρμοσμένο συντελεστή προσδιορισμού (22%) είναι το "Νοσοκομειακές κλίνες ανά πληθυσμό – Κλίνες των Nursing Homes ανά πληθυσμό – Μέσος όρος χαμηλότερης θερμοκρασίας ", χωρίς όμως τα αποτελέσματα να θεωρούνται σημαντικά. Το στατιστικά σημαντικό μοντέλο με το μεγαλύτερο $R^2$ είναι το μονομεταβλητό μοντέλο με ανεξάρτητη μεταβλητή τις Νοσοκομειακές κλίνες ανά πληθυσμό.

OLS Regression

Τη μη σημαντικότητα των αποτελεσμάτων έρχεται να επιβεβαιώσει και η ανάλυση OLS όπου παρατηρείται πως η τιμή Probability της μεταβλητής που αντιστοιχεί στις κλίνες των Nursing Homes ανά τον πληθυσμό του πολυγώνου HSA είναι μεγαλύτερη από 0.05 και η τιμή Coefficient είναι αρνητική.

**Summary of OLS Results - Model Variables**

| Variable | Coefficient [a] | StdError | t-Statistic | Probability [b] | Robust_SE | Robust_t | Robust_Pr [b] | VIF [c] |
|---|---|---|---|---|---|---|---|---|
| Intercept | -0.005177 | 0.007282 | -0.711023 | 0.477951 | 0.007502 | -0.690168 | 0.490941 | -------- |
| CLIMA_TMIN | 0.002317 | 0.000612 | 3.786614 | 0.000214* | 0.000565 | 4.098614 | 0.000067* | 1.012255 |
| CAPA_POP | 0.033663 | 0.007394 | 4.552739 | 0.000011* | 0.014747 | 2.282730 | 0.023560* | 2.241529 |
| NURSBEDS_POP | -0.005030 | 0.007029 | -0.715611 | 0.475119 | 0.009112 | -0.552031 | 0.581592 | 2.228107 |

Αντίστοιχα και η τιμή p-value του δείκτη Jarque-Bera υποδηλώνει πρόβλημα κανονικότητας των υπολοίπων.

**OLS Diagnostics**

| | | | |
|---|---|---|---|
| Input Features: | HSA_CDI_nurs | Dependent Variable: | CDI_POP |
| Number of Observations: | 191 | Akaike's Information Criterion (AICc) [d]: | -822.753907 |
| Multiple R-Squared [d]: | 0.231644 | Adjusted R-Squared [d]: | 0.219318 |
| Joint F-Statistic [e]: | 18.792308 | Prob(>F), (3,187) degrees of freedom: | 0.000000* |
| Joint Wald Statistic [e]: | 23.459344 | Prob(>chi-squared), (3) degrees of freedom: | 0.000032* |
| Koenker (BP) Statistic [f]: | 15.469439 | Prob(>chi-squared), (3) degrees of freedom: | 0.001456* |
| Jarque-Bera Statistic [g]: | 9380.886446 | Prob(>chi-squared), (2) degrees of freedom: | 0.000000* |

Geographically Weighted Regression

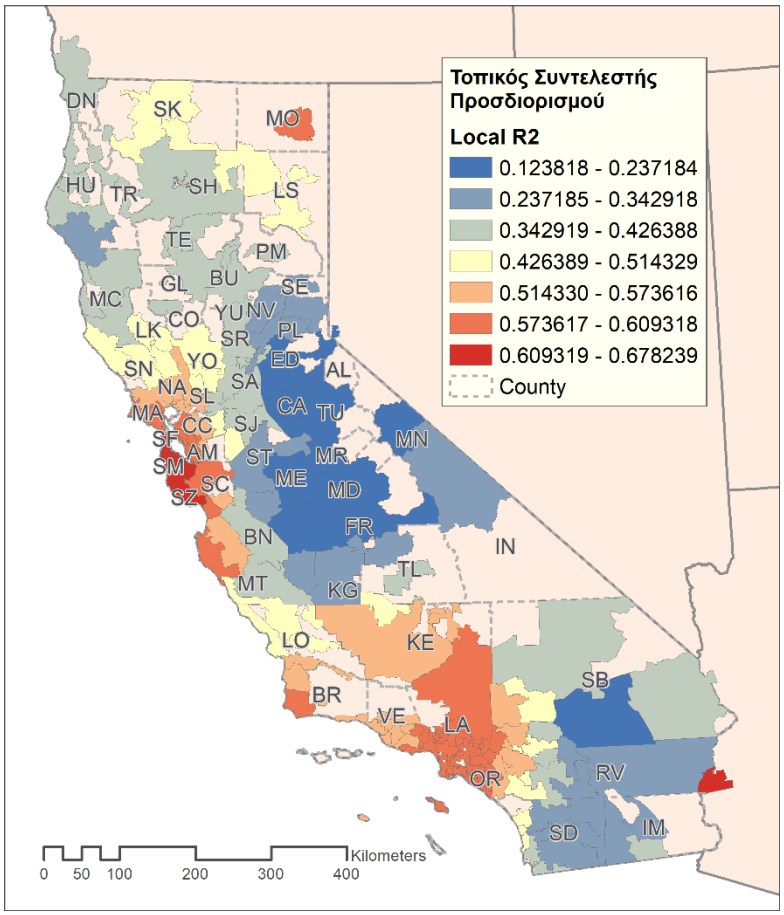
Με την μέθοδο της Γεωγραφικά Σταθμισμένης Παλινδρόμησης τα αποτελέσματα διαφοροποιούνται, κάτι που επισημαίνει την σημαντικότητα της συγκεκριμένης μεθόδου.

Με σταθερό πυρήνα για την ανάλυση και υπολογισμό του εύρους ζώνης σύμφωνα με τη μέθοδο του κριτηρίου Akaike, ο προσαρμοσμένος συντελεστής προσδιορισμού $R^2$ είναι 0.594714.

*Πίνακας 10: Αποτελέσματα GWR για CDI με fixed kernel (Στρατηγική 3η ).*

| OBJECTID * | VARNAME | VARIABLE | DEFINITION |
|---|---|---|---|
| 1 | Bandwidth | 1.545582 | |
| 2 | ResidualSquares | 0.064089 | |
| 3 | EffectiveNumber | 29.181377 | |
| 4 | Sigma | 0.019901 | |
| 5 | AICc | -932.60037 | |
| 6 | R2 | 0.654828 | |
| 7 | R2Adjusted | 0.594714 | |
| 8 | Dependent Field | 0 | CDI_pop |
| 9 | Explanatory Field | 1 | Capa_pop |
| 10 | Explanatory Field | 2 | clima_TMIN |
| 11 | Explanatory Field | 3 | NursBeds_pop |

Οι χαμηλές τιμές των Residual Squares, Sigma και AICc υποδεικνύουν ένα μοντέλο καλά προσαρμοσμένο στα παρατηρούμενα δεδομένα με μικρή τυπική απόκλιση των υπολοίπων. Επίσης, από τον Attribute Table της feature class που δημιουργήθηκε από το ArcMap προκύπτει ότι η μεγαλύτερη Condition τιμή είναι 29.969666 < 30, από το οποίο συνεπάγεται ότι δεν εντοπίζεται πρόβλημα πολυκεντρικότητας και η μεγαλύτερη τιμή του τοπικού συντελεστή προσδιορισμού $R^2$ είναι 0.678239. Στον χάρτη της Εικόνας 30 παρουσιάζονται όλα τα τοπικά $R^2$ της περιοχής μελέτης.




*Εικόνα 30: Τοπικός συντελεστής προσδιορισμού με τη μέθοδο GWR.*

Τα αποτελέσματα με προσαρμοστικό πυρήνα (adaptive kernel) επηρεάζονται από 81 γείτονες και έχουν adjusted $R^2 = 0.618876$.

*Πίνακας 11: Αποτελέσματα GWR για CDI με adaptive kernel (Στρατηγική 3η ).*

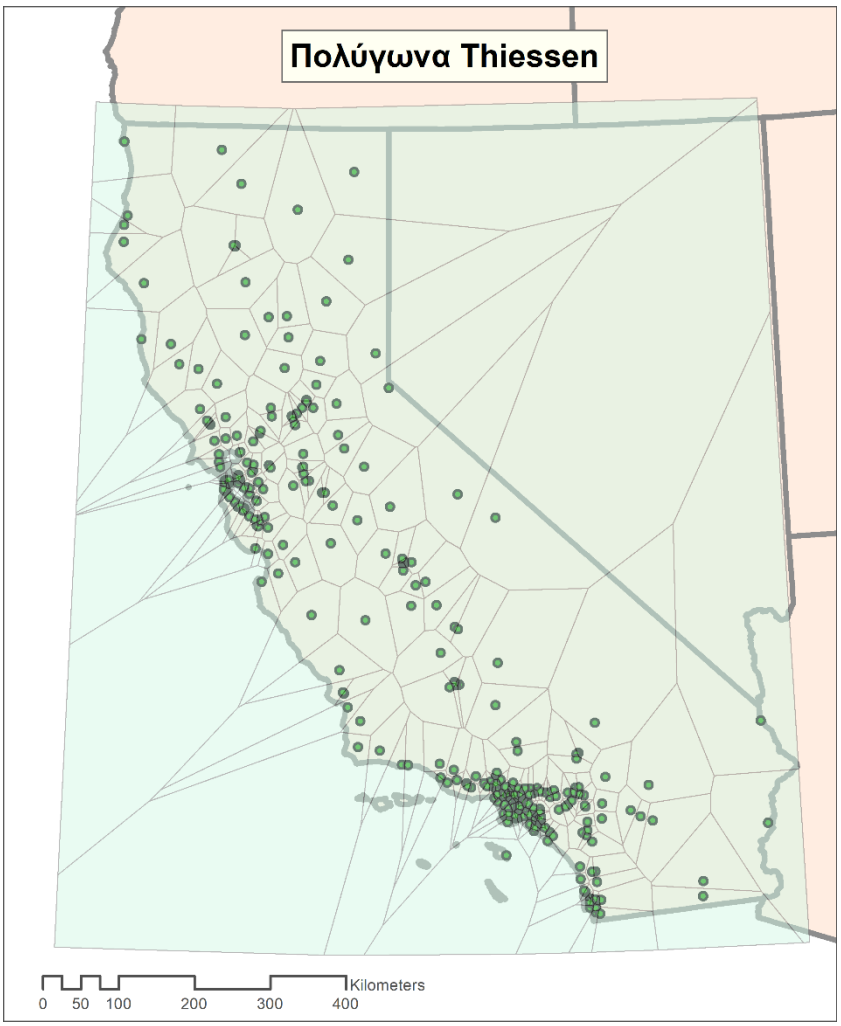
| | OBJECTID * | VARNAME | VARIABLE | DEFINITION |
|---|---|---|---|---|
| ▸ | 1 | Neighbors | 81 | |
| | 2 | ResidualSquares | 0.061861 | |
| | 3 | EffectiveNumber | 24.905982 | |
| | 4 | Sigma | 0.019299 | |
| | 5 | AICc | -945.917043 | |
| | 6 | R2 | 0.66683 | |
| | 7 | R2Adjusted | 0.618876 | |
| | 8 | Dependent Field | 0 | CDI_pop |
| | 9 | Explanatory Field | 1 | Capa_pop |
| | 10 | Explanatory Field | 2 | clima_TMIN |
| | 11 | Explanatory Field | 3 | NursBeds_pop |

Οι τιμές του Condition μειώθηκαν με τη μεγαλύτερη να είναι 24.336031 < 30 και τη μέγιστη τιμή του τοπικού συντελεστή προσδιορισμού $R^2$ είναι 0.782431. Ωστόσο, ο αριθμός των γειτόνων

που παίρνουν μέρος στην παλινδρόμηση είναι αρκετά μεγάλος καθώς αποτελεί το ήμισυ του συνολικού αριθμού των πολυγώνων HSA.

**Στρατηγική 4η**

Η τέταρτη στρατηγική αφορά των επιμερισμό της περιοχής μελέτης σε πολύγωνα Thiessen, με κέντρο τα νοσοκομεία, στα οποία σημειώθηκαν κρούσματα CDI, όπως παρουσιάζεται στον χάρτη της Εικόνας 31.




*Εικόνα 31: Διαχωρισμός περιοχής μελέτης με πολύγωνα Thiessen.*

Exploratory Regression

Με τη χρήση αυτού του εργαλείου που διατίθεται από ArcMap εξετάστηκαν, ομοίως με παραπάνω, οι μεταβλητές που αφορούν τη χωρητικότητα των νοσοκομείων και των Nursing

Homes, καθώς και ο μέσος όρος χαμηλότερης θερμοκρασίας που έχει καταγραφεί από τον κοντινότερο μετεωρολογικό σταθμό των νοσοκομείων. Όπως αναφέρθηκε και προηγουμένως κατά την ανάλυση της μεθόδου, ο πληθυσμός υπολογίστηκε βάσει Οικοδομικών Τετραγώνων και όχι βάσει ΤΚ. Τα αποτελέσματα της Exploratory Regression με κάθε μία μεταβλητή ξεχωριστά είναι:

*Πίνακας 12 : Προσαρμοσμένος συντελεστής προσδιορισμού με Exploratory Regression για CDI (Στρατηγική 4η ).*

| **Ανεξάρτητη Μεταβλητή** | **Adjusted $R^2$** |
|---|---|
| Νοσοκομειακές κλίνες ανά πληθυσμό | +90% |
| Κλίνες των Nursing Homes ανά πληθυσμό | +6 % |
| Μέσος όρος χαμηλότερης θερμοκρασίας | +1 % |

Είναι εμφανής η σημαντική επιρροή που αποτελεί το μέγεθος του νοσοκομείο στα ποσοστά κρουσμάτων της λοίμωξης από το βακτήριο C.diff. Το πιο σημαντικό μοντέλο είναι το μονομεταβλητό με ανεξάρτητη μεταβλητή τις Νοσοκομειακές κλίνες ανά πληθυσμό.

OLS Regression

Όπως προκύπτει από το Report της παλινδρόμησης OLS, τα αποτελέσματα που αφορούν την χωρητικότητα των νοσοκομείων είναι στατιστικά σημαντικά.

**Summary of OLS Results - Model Variables**

| Variable | Coefficient [a] | StdError | t-Statistic | Probability [b] | Robust_SE | Robust_t | Robust_Pr [b] |
|---|---|---|---|---|---|---|---|
| Intercept | -0.316320 | 0.204667 | -1.545532 | 0.123328 | 0.142307 | -2.222794 | 0.026993* |
| CAPACITY_POP | 0.162589 | 0.003116 | 52.184131 | 0.000000* | 0.005898 | 27.567250 | 0.000000* |

Ο προσαρμοσμένος συντελεστής προσδιορισμού $R^2$ είναι 0.904024, με το p-value του μέτρου Jarque-Bera να επισημαίνει πρόβλημα κανονικότητας των υπολοίπων.

**OLS Diagnostics**

| | | | |
|---|---|---|---|
| Input Features: | cdi_clima_demoBG_Nurs_T | Dependent Variable: | CDI_POPULATION |
| Number of Observations: | 290 | Akaike's Information Criterion (AICc) [d]: | 1505.289113 |
| Multiple R-Squared [d]: | 0.904357 | Adjusted R-Squared [d]: | 0.904024 |
| Joint F-Statistic [e]: | 2723.183575 | Prob(>F), (1,288) degrees of freedom: | 0.000000* |
| Joint Wald Statistic [e]: | 759.953275 | Prob(>chi-squared), (1) degrees of freedom: | 0.000000* |
| Koenker (BP) Statistic [f]: | 29.828559 | Prob(>chi-squared), (1) degrees of freedom: | 0.000000* |
| Jarque-Bera Statistic [g]: | 2313.357002 | Prob(>chi-squared), (2) degrees of freedom: | 0.000000* |

Geographically Weighted Regression

Τα αποτελέσματα της GWR, με εξαρτημένη μεταβλητή τα κρούσματα CDI προς τον πληθυσμό που αναλογεί στα Οικοδομικά Τετράγωνα εμφανίζονται στον πίνακα που ακολουθεί.

*Πίνακας 13: Αποτελέσματα GWR για CDI με fixed kernel (Στρατηγική 4η ).*

| OBJECTID * | VARNAME | VARIABLE | DEFINITION |
|---|---|---|---|
| 1 | Bandwidth | 2.579913 | |
| 2 | ResidualSquares | 2670.463181 | |
| 3 | EffectiveNumber | 7.284766 | |
| 4 | Sigma | 3.073397 | |
| 5 | AICc | 1480.517149 | |
| 6 | R2 | 0.914462 | |
| 7 | R2Adjusted | 0.912561 | |
| 8 | Dependent Field | 0 | cdi_population |
| 9 | Explanatory Field | 1 | capacity_pop |

Η τιμή του προσαρμοσμένου συντελεστή προσδιορισμού $R^2$ είναι περίπου 91%. Ωστόσο, παρατηρούνται μεγάλες τιμές των Residual Squares, Sigma και AICc, επομένως το μοντέλο δεν είναι καλά προσαρμοσμένο ως προς τα δεδομένα.

## 5.3 Συμπεράσματα Ανάλυσης της λοίμωξης από *Clostridium Difficile*

Στο κεφάλαιο αυτό παρουσιάστηκαν τέσσερις στρατηγικές επεξεργασίας των δεδομένων, που αφορούν τα κρούσματα λοίμωξης από *Clostridium Difficile* που έχουν σημειωθεί σε νοσοκομεία της πολιτείας Καλιφόρνια των ΗΠΑ το έτος 2015. Για κάθε μία από τις στρατηγικές αυτές πραγματοποιήθηκε στατιστική και χωρική ανάλυση των δεδομένων με τις μεθόδους των ελαχίστων τετραγώνων (OLS) και της Γεωγραφικά Σταθμισμένης παλινδρόμησης.

Κατά την πρώτη στρατηγική δεν ήταν δυνατή η ύπαρξη κάποιου μοντέλου που να εξηγεί τα ποσοστά κρουσμάτων. Το πρόβλημα που προέκυψε αφορούσε τον τρόπο διαχωρισμού του

πληθυσμού που εξυπηρετείται από τα νοσοκομεία. Ο πληθυσμός ενός HSA πολυγώνου είναι αδύνατο να εξυπηρετηθεί από ένα μόνο νοσοκομείο, όταν μέσα στο πολύγωνο βρίσκονται και άλλα.

Το πρόβλημα αυτό επιχειρήθηκε να λυθεί με την εισαγωγή βαρών που αφορούν το μέγεθος του κάθε νοσοκομείου. Με τη μέθοδο αυτή, επιτυγχάνεται η πιο ρεαλιστική διάσπαση των ασθενών στα νοσοκομεία. Τα αποτελέσματα όμως της στατιστικής αλλά και της χωρικής ανάλυσης, υπέδειξαν ένα μοντέλο το οποίο δεν προσαρμόζεται στα δεδομένα.

Η τρίτη ανάλυση πραγματοποιήθηκε βάση των πολυγώνων HSA και όχι πλέον ανά νοσοκομείο. Η μέθοδος ελαχίστων τετραγώνων κατέδειξε ένα μοντέλο που τα αποτελέσματά του δεν χαρακτηρίζονται σημαντικά. Ωστόσο, παρατηρήθηκε ότι με τη Γεωγραφικά Σταθμισμένη Παλινδρόμηση και συγκεκριμένα με τη μέθοδο σταθερού πυρήνα, οι ανεξάρτητες μεταβλητές εξηγούν το 60% του μοντέλου, το οποίο προσαρμόζεται ικανοποιητικά στα δεδομένα. Η χωρική ανάλυση με προσαρμοσμένο πυρήνα δεν κρίνεται σημαντική, διότι το κάθε πολύγωνο HSA δέχεται επιρροή από 81 γείτονες από τους συνολικά 191.

Τέλος, χωρίζοντας την περιοχή της Καλιφόρνια σε πολύγωνα Thiessen, προέκυψε πρόβλημα ακραίων τιμών. Στις μεγάλες πόλεις της πολιτείας, εκεί όπου ο αριθμός των νοσοκομείων είναι μεγάλος, δημιουργούνται πολύγωνα μικρά σε εμβαδό και επομένως με μικρό πληθυσμό. Τα νοσοκομεία αυτά όμως, στην πραγματικότητα, δεν έχουν τόσο μικρή ακτίνα εξυπηρέτησης. Επομένως, η χρήση των πολυγώνων Thiessen για την ανάλυση δεδομένων που αφορούν κρούσματα σε νοσοκομειακές εγκαταστάσεις κρίνεται ακατάλληλη.

# 6. Κεφάλαιο 6ο - Συμπεράσματα και Προτάσεις

Στο τελευταίο κεφάλαιο παρουσιάζονται συγκεντρωτικά τα συμπεράσματα που εξάγονται από τη διπλωματική εργασία εστιάζοντας κυρίως στα αποτελέσματα που προκύπτουν από την εφαρμογή της στατιστικής και χωρικής ανάλυσης των δεδομένων, στα προβλήματα που παρουσιάστηκαν κατά την εκπόνηση της μελέτης, καθώς και στη σημασία του τρόπου αντιμετώπισης και μεθοδολογίας των διαθέσιμων δεδομένων. Παράλληλα, βασικό τμήμα του κεφαλαίου αποτελούν οι μελλοντικές προτάσεις που προτείνονται για την συνέχιση του συγκεκριμένου επιστημονικού πεδίου, ώστε με το επιπλέον έργο να υπάρξει εξέλιξη που αφορά τις μεθόδους ανάλυσης δεδομένων στον κλάδο της Ιατρικής.

## 6.1 Συμπεράσματα

Με την ολοκλήρωση της παρούσας εργασίας, αποδεικνύεται ότι περιβαλλοντικοί παράγοντες επιδρούν στην αύξηση ποσοστών ιατρικών κρουσμάτων. Ανάλογα με τα διαθέσιμα δεδομένα δίνεται και η επιλογή της κατάλληλης μεθοδολογίας που ακολουθείται για την ανάλυση αυτών των δεδομένων.

Η γραμμική παλινδρόμηση αποτελεί σημαντικό εργαλείο για τη στατιστική ανάλυση των διαθέσιμων δεδομένων, ωστόσο στην έρευνα της Ιατρικής απαραίτητο παράγοντα ανάλυσης αποτελεί ο χώρος. Η Γεωγραφικά Σταθμισμένη Παλινδρόμηση έρχεται να δώσει ένα πιο ολοκληρωμένο μοντέλο και να εξηγήσει καλύτερα την επιρροή των περιβαλλοντικών παραγόντων σε προβλήματα ιατρικής.

Βασικό εμπόδιο αποτελεί η συλλογή των κατάλληλων δεδομένων, δηλαδή η εύρεση εκείνων των στοιχείων που αποδεδειγμένα επηρεάζουν μία νόσο ή εκείνων που θεωρούνται ότι συμβάλλουν στην αύξηση ποσοστών της. Χαρακτηριστικό παράδειγμα είναι η ανωφελής αναζήτηση χώρων πρασίνου που χρησιμοποιούν χλοοτάπητα, ο οποίος ευθύνεται για τη μετάδοση του βακτηρίου C.Diff ή οι δείκτες ποιότητας των νοσοκομείων, πληροφορία η οποία δεν διατίθεται στο διαδίκτυο. Παρόλα αυτά, δεν πρέπει να ξεχνά κανείς την ύπαρξη πολλών πηγών, οι οποίες μπορούν να χρησιμοποιηθούν για τη δημιουργία δεδομένων. Τέτοια πηγή στη παρούσα αποτέλεσε ο διαδικτυακός τόπος Foursquare, από όπου εξήχθη η πληροφορία των χώρων άθλησης στην εξέταση της παχυσαρκίας.

Βασικό και κύριο μέλημα στην ανάλυση δεδομένων συνιστά ο τρόπος διαχωρισμού της περιοχής προς μελέτη. Ανάλογα με το πλήθος των διαθέσιμων δεδομένων αλλά και με την κλίμακα ανάλυσης που επιθυμείται, η μεθοδολογία που ακολουθείται διαφέρει. Επομένως, ένας αναλυτής θα πρέπει να προσαρμόζει την ανάλυση σε σχέση με τα διαθέσιμα δεδομένα, χωρίς όμως να τα παραποιεί.

## 6.2 Μελλοντικές Προτάσεις

Μελετώντας τα αποτελέσματα της διπλωματικής εργασίας γίνεται αντιληπτό ότι η προσπάθεια ανάλυσης ιατρικών προβλημάτων είναι ιδιαίτερα δύσκολη διαδικασία. Η προσπάθεια διαχείρισης μεγάλων δεδομένων (big data) αποτελεί μία πρόταση για τη συνέχιση της παρούσας έρευνας. Περισσότερα δεδομένα μπορούν να φανούν ικανά για μία πιο λεπτομερής ανάλυση, όπως για παράδειγμα τα στοιχεία ασθενών που έχουν προσβληθεί από μία νόσο και αφορούν δημογραφικές μεταβλητές όπως την τοποθεσία διαμονής του ασθενούς, το εισόδημά του, την ηλικία του, την φυλή του κ.ά., αλλά και στοιχεία από τον ιατρικό του φάκελο.

Εισάγοντας μεταβλητές οι οποίες δεν αφορούν μόνο το περιβάλλον, αλλά εστιάζονται στην ιατρική και γενετική θα ήταν ενδιαφέρον να μελετηθούν, ώστε να προκύψει κάποιο πιο ολοκληρωμένο μοντέλο, καθώς αυτοί οι παράγοντες αποδεδειγμένα επηρεάζουν τα ποσοστά κρουσμάτων και δεν συμπεριλήφθηκαν στην παρούσα εργασία.

Τέλος, αφού έχουν προηγηθεί οι δύο προαναφερθείσες προτάσεις, θα ήταν σκόπιμο και ενδιαφέρον να δημιουργηθεί ένα μοντέλο πρόβλεψης που να τις περιλαμβάνει. Το μοντέλο αυτό θα μπορούσε να χρησιμοποιηθεί για να προβλέπει τα αναμενόμενα ποσοστά κρουσμάτων μίας νόσου ανάλογα με τα υπόλοιπα χαρακτηριστικά, περιβαλλοντικά και μη, που έχει κάποια περιοχή.

# 7. Βιβλιογραφία


[1] Μ. Κάβουρας, *Αρχές γεωπληροφορικής και συστημάτων γεωγραφικών πληροφοριών*. Αθήνα: Εθνικό Μετσόβιο Πολυτεχνείο, 2007.

[2] Κ. Χ. Κουτσόπουλος, *Γεωγραφικά συστήματα πληροφοριών και ανάλυση χώρου*. Αθήνα: Παπασωτηρίου, 2005.

[3] P. A. Burrough, "Development of intelligent geographical information systems," *International Journal of Geographical Information Systems,* vol. 6, no. 1, pp. 1-11, 1992/01/01 1992.

[4] Σ. Καλογήρου, Χωρική ανάλυση, Αθήνα: Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών, 2015. [Online]. Available: http://hdl.handle.net/11419/5029.

[5] C. Yiakoumettis, Doulamis, N., Miaoulis, G., & Ghazanfarpour, D., "Active learning of user's preferences estimation towards a personalized 3D navigation of geo-referenced scenes," *Geoinformatica,* 18(1), pp. 27-62, 2014.

[6] N. Doulamis, Yiakoumettis, C., & Miaoulis, G., "Personalised 3D navigation and understanding of Geo-referenced Scenes. In World of Wireless, Mobile and Multimedia Networks (WoWMoM)," in *IEEE 14th International Symposium and Workshops*, 2013, pp. 1-6: IEEE.

[7] D. Unwin, *Introductory spatial analysis*. London ; New York: Methuen, 1981, pp. xii, 212 p.

[8] T. C. Bailey, "GIS and simple systems for visual, interactive, spatial analysis," *The Cartographic Journal,* vol. 27, no. 2, pp. 79-84, 1990/12/01 1990.

[9] T. C. Bailey and A. C. Gatrell, *Interactive spatial data analysis*. Harlow Essex, England

New York, NY: Longman Scientific & Technical ;

J. Wiley, 1995, pp. xiv, 413 p.

[10] Κ. Κουτσόπουλος, *Γεωγραφία: Μεθοδολογία και Μέθοδοι Ανάλυσης Χώρου*. Αθήνα: Εκδόσεις Συμμετρία, 1990.

[11] Γ. Ν. Φώτης, *Ποσοτική χωρική ανάλυση*. Αθήνα: Εκδόσεις Γκοβόστη, 2009.

[12] A. S. Fotheringham, C. Brunsdon, and M. Charlton, *Quantitative geography : perspectives on spatial data analysis*. London ; Thousand Oaks, Calif.: Sage Publications, 2000, pp. xii, 270 p.

[13] G. J. Musa *et al.*, "Use of GIS Mapping as a Public Health Tool-From Cholera to Cancer," *Health Serv Insights,* vol. 6, pp. 111-6, 2013.

[14] D. Z. Sui, "Geographic Information Systems and Medical Geography: Toward a New Synergy," *Geography Compass,* vol. 1, no. 3, pp. 556-582, 2007.

[15] E. K. Cromley and S. L. McLafferty, *GIS and public health*, 2nd ed. New York: The Guilford Press, 2012, p. 503 p.

[16] Esri. (March 20). *Public Health*. Available: https://www.esri.com/en-us/industries/health/segments/public-health

[17] M. o. H. a. S. Department of Public Health, Myanmar. (March 20). *Geo-enabling the Health Information System in Myanmar*. Available: https://doph.maps.arcgis.com/apps/MapJournal/index.html?appid=d0349cc76cb749fd8e6e2d4e3e756131

[18] L. Anselin, "Local Indicators of Spatial Association—LISA," *Geographical Analysis,* Article vol. 27, no. 2, pp. 93-115, 1995.

[19] D. A. Griffith, D. W. S. Wong, and T. Whitfield, "Exploring Relationships Between the Global and Regional Measures of Spatial Autocorrelation," *Journal of Regional Science,* vol. 43, no. 4, pp. 683-710, 2003.

[20] Esri. (March 22). *Spatial Autocorrelation (Global Moran's I)*. Available: http://pro.arcgis.com/en/pro-app/tool-reference/spatial-statistics/spatial-autocorrelation.htm

[21] M. Sawada. (March 22). *Global Spatial Autocorrelation Indices - Moran's I, Geary's C and the General Cross-Product Statistic*. Available: http://www.lpc.uottawa.ca/publications/moransi/moran.htm

[22] P. A. P. Moran, "The Interpretation of Statistical Maps," *Journal of the Royal Statistical Society. Series B (Methodological),* vol. 10, no. 2, pp. 243-251, 1948.

[23] J. K. Ord and A. D. Cliff, *Spatial autocorrelation* (Monographs in spatial and environmental systems analysis, no. 5). London: Pion, 1973, p. 178 S.

[24] A. D. Cliff and J. K. Ord, *Spatial processes models and applications*. London: Pion, 1981, p. 266 S.

[25] Esri. (March 23). *How Cluster and Outlier Analysis (Anselin Local Moran's I) works*. Available: http://desktop.arcgis.com/en/arcmap/10.3/tools/spatial-statistics-toolbox/h-how-cluster-and-outlier-analysis-anselin-local-m.htm

[26] A. Getis and J. K. Ord, *The analysis of spatial association by use of distance statistics*. pp. S. 189-206.

[27] J. K. Ord and A. Getis, "Local Spatial Autocorrelation Statistics: Distributional Issues and an Application," *Geographical Analysis,* vol. 27, no. 4, pp. 286-306, 1995/10/01 1995.

[28] S. M. Kim and Y. Choi, "Assessing Statistically Significant Heavy-Metal Concentrations in Abandoned Mine Areas via Hot Spot Analysis of Portable XRF Data," *Int J Environ Res Public Health,* vol. 14, no. 6, Jun 18 2017.

[29] Esri. (March 28). *Hot Spot Analysis (Getis-Ord Gi*)*. Available: http://pro.arcgis.com/en/pro-app/tool-reference/spatial-statistics/hot-spot-analysis.htm

[30] A. D. Doulamis, Doulamis, N. D., & Kollias, S. D. , "Recursive non linear models for on line traffic prediction of VBR MPEG coded video sources," in *Neural Networks*, IJCNN 2000, 2000, vol. 6, pp. 114-119: Proceedings of the IEEE-INNS-ENNS International Joint Conference.

[31] J. F. Kenney and E. S. Keeping, *Mathematics of statistics*, 3d ed. New York,: Van Nostrand company, 1954, p. v.

[32] Δ. Πετρίδης, Ανάλυση πολυμεταβλητών τεχνικών, Αθήνα: Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών, 2015. [Online]. Available: https://repository.kallipos.gr/pdfviewer/web/viewer.html?file=/bitstream/11419/2126/7/petridis%28whole%29_KOY.pdf.

[33] Esri. (April 12). *Exploratory Regression*. Available: http://desktop.arcgis.com/en/arcmap/10.3/tools/spatial-statistics-toolbox/exploratory-regression.htm

[34] Esri. (April 12). *Ordinary Least Squares (OLS)*. Available: http://desktop.arcgis.com/en/arcmap/10.3/tools/spatial-statistics-toolbox/ordinary-least-squares.htm

[35] C. Brunsdon, A. S. Fotheringham, and E. Charlton Martin, "Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity," *Geographical Analysis,* vol. 28, no. 4, pp. 281-298, 1996/10/01 1996.

[36] C. Brunsdon, S. Fotheringham, and M. Charlton, "Geographically Weighted Regression as a Statistical Model," Spatial Analysis Research Group Department of Geography, UK, Working Paper 2000, Available: http://eprints.maynoothuniversity.ie/5975/.

[37] C. Brunsdon, S. Fotheringham, and M. Charlton, "Geographically Weighted Regression," *Journal of the Royal Statistical Society: Series D (The Statistician),* vol. 47, no. 3, pp. 431-443, 1998/09/01 2002.

[38] A. D. Doulamis, Doulamis, N. D., & Kollias, S. D., "An adaptable neural-network model for recursive nonlinear traffic prediction and modeling of MPEG video sources," in *IEEE Transactions on Neural Networks*, 2003,14(1), pp. 150-166.

[39] Esri. (April 27). *What they don't tell you about regression analysis*. Available: http://pro.arcgis.com/en/pro-app/tool-reference/spatial-statistics/what-they-don-t-tell-you-about-regression-analysis.htm

[40] Esri. (April 28). *Interpreting GWR results*. Available: http://desktop.arcgis.com/en/arcmap/10.3/tools/spatial-statistics-toolbox/interpreting-gwr-results.htm

[41] D. o. Finance. (February 26). *NEW STATE POPULATION REPORT*. Available: http://www.dof.ca.gov/Forecasting/Demographics/Estimates/E-1/documents/E-1_2017PressRelease.pdf

[42] USGS. (June 17). *Elevations and Distances in the United States*. Available: https://web.archive.org/web/20111015012701/http:/egsc.usgs.gov/isb/pubs/booklets/elvadist/elvadist.html

[43] W. R. C. Center. *CLIMATE OF CALIFORNIA*. Available: https://wrcc.dri.edu/narratives/CALIFORNIA.htm

[44] C. D. o. Finance. (March 1). *Population Estimates for Cities, Counties, and the State, 2011-2017, with 2010 Benchmark*. Available: http://www.dof.ca.gov/Forecasting/Demographics/Estimates/E-4/2010-17/

[45] D. W. Haslam and W. P. James, "Obesity," *Lancet,* vol. 366, no. 9492, pp. 1197-209, Oct 1 2005.

[46] A. Angeli and N. Doulamis, "Atherosclerosis Risk Factors and Degrees of Stenosis in three Arterial Sites," in *Live 2018 Conference*, Patras, Greece, May 2018.

[47] G. Eknoyan, "Adolphe Quetelet (1796-1874)--the average man and indices of obesity," *Nephrol Dial Transplant,* vol. 23, no. 1, pp. 47-51, Jan 2008.

[48] N. Choices. (May 2). *Obesity*. Available: https://www.nhs.uk/conditions/obesity/

[49] WHO. (May 2). *Obesity and overweight*. Available: http://www.who.int/en/news-room/fact-sheets/detail/obesity-and-overweight

[50] T. S. o. Obesity. (May 3). *Adult Obesity in the United States*. Available: https://stateofobesity.org/adult-obesity/

[51] Jupyter. Available: http://jupyter.org/

[52] (December 14). *What is the Jupyter Notebook?* Available: http://jupyter-notebook-beginner-guide.readthedocs.io/en/latest/what_is_jupyter.html

[53] Jupyter. (December 12). *Installing Jupyter Notebook*. Available: http://jupyter.readthedocs.io/en/latest/install.html

[54] lifewire. (December 12). *Command Prompt: What It Is and How to Use It*. Available: https://www.lifewire.com/command-prompt-2625840

[55] Δ. Λεβεντέας, Οδηγός Python Μέσω Παραδειγμάτων. [Online]. Available: http://python.org.gr/phocadownload/Tutorials/tutorial_by_example.pdf.

[56] (December 14). *Εισαγωγή στο JSON*. Available: https://www.json.org/json-el.html

[57] (December 14). *JSON encoder and decoder*. Available: https://docs.python.org/2/library/json.html

[58] (December 15). *Using the Requests Library in Python*. Available: http://www.pythonforbeginners.com/requests/using-requests-in-python

[59] (December 15). *CSV File Reading and Writing*. Available: https://docs.python.org/2/library/csv.html

[60] (December 15). *Python's time.sleep()*. Available: https://www.pythoncentral.io/pythons-time-sleep-pause-wait-sleep-stop-your-code/

[61] (December 15). *progressbar2*. Available: https://pypi.python.org/pypi/progressbar2

[62] (December 17). *Application programming interface*. Available: https://en.wikipedia.org/wiki/Application_programming_interface

[63] (January 4). *Places API*. Available: https://developer.foursquare.com/places-api

[64] (January 4). *Search for Venues*. Available: https://developer.foursquare.com/docs/api/venues/search

[65] C. P. Davis. (April 20). *Is C. diff (Clostridium difficile) Contagious?* Available: https://www.medicinenet.com/is_c_diff_clostridium_difficile_contagious/article.htm#what_is_c_diff_clostridium_difficile

[66] I. M. Zacharioudakis, F. N. Zervou, E. E. Pliakos, P. D. Ziakas, and E. Mylonakis, "Colonization with toxinogenic C. difficile upon hospital admission, and risk of infection: a systematic review and meta-analysis," *Am J Gastroenterol,* vol. 110, no. 3, pp. 381-90; quiz 391, Mar 2015.

[67] S. Karanika, S. Paudel, F. N. Zervou, C. Grigoras, I. M. Zacharioudakis, and E. Mylonakis, "Prevalence and Clinical Outcomes of Clostridium difficile Infection in the Intensive Care Unit: A Systematic Review and Meta-Analysis," *Open Forum Infect Dis,* vol. 3, no. 1, p. ofv186, Jan 2016.

[68] (April 20). *File: How C. difficile spreads*. Available: https://commons.wikimedia.org/wiki/File:How_C._difficile_spreads.png

[69] (April 20). *Fecal–oral route*. Available: https://en.wikipedia.org/wiki/Fecal%E2%80%93oral_route

[70] (April 20). *PATHS OF DISEASE TRANSMISSION*. Available: https://water1st.org/problem/f-diagram/

[71] (April 20). *The F-Diagram*. Available: https://www.researchgate.net/figure/Figure1-The-F-diagram-showing-the-different-faecal-oral-transmission-routes-and_fig2_237200300

[72] CDC, "Antibiotic Use in the United States," Atlanta2017, Available: https://www.cdc.gov/antibiotic-use/stewardship-report/pdf/stewardship-report.pdf.

[73] Esri. (May 29). *Create Thiessen Polygons*. Available: http://pro.arcgis.com/en/pro-app/tool-reference/analysis/create-thiessen-polygons.htm

# ΠΑΡΑΡΤΗΜΑ: Κώδικας Python

# Foursquare to ShapeFile

## Importing libraries and defining variables

In [ ]:
```
#Εισαγωγή βιβλιοθηκών
import json, requests, csv
from time import sleep
import progressbar

#καθορισμός μεταβλητής μετρητή των αθλητικών χώρων που καταγράφονται στη progressbar
count=0

#URL κλήσης του API
url = 'https://api.foursquare.com/v2/venues/search'

#καθορισμός βήματος και συντεταγμένων κορυφής του πολυγώνου
inc=0.01
latSW=41.515537579700002
lngSW=-124.445528932000000
latNE=42.006392332499999
lngNE=-119.999167761000000

#υπολογισμός της τιμής που αντιστοιχεί στο 100% στη progressbar
maxVal=int((latNE-latSW)/inc+1)*int((lngNE-lngSW)/inc+1)
#σχεδίαση της progressbar
bar = progressbar.ProgressBar(widgets=[progressbar.Percentage(),
    ' [', progressbar.Timer(), '] ',
    progressbar.Bar(),
    ' (', progressbar.ETA(), ') ', progressbar.DynamicMessage('count')
],redirect_stdout=True, max_value=maxVal)
```

## Defining function that calls the foursquare API

In [ ]:
```
#συνάρτηση αιτήματος στο foursquare API
def query(latSW, lngSW):
    global count
    #αυτόματη δημιουργία και άνοιγμα αρχείου csv για την εγγραφή των αποτελεσμάτων, μέσω της βιβλιοθήκη
ς csv
    with open('gym01.csv', 'a', encoding="utf-8") as csvfile:
        #δημιουργία εγγραφέα και καθορισμός μορφοποίησης και διαχείρισης του αρχείου csv
        gymwriter = csv.writer(csvfile, delimiter=';', quotechar='|', quot
ing=csv.QUOTE_MINIMAL)
        #υπολογισμός του παραθυρου αναζητησης
        latNE=latSW+inc
        lngNE=lngSW+inc
        #προσδιορισμός παραμέτρων του αιτήματος search for venues
        params = dict(
          client_id='0NIVG5WGW4JUF1L5PB5U4X3R4YTGHJIYYJXHQKS2OIIYRCDO',
          client_secret='TNFKNOZ5FIGLK1J1RPBNA1QEI1NQPJHMPRSL5WNFAQCVPLTV'
,
          v='20171101',
          sw="%0.7f,%0.7f" % (latSW, lngSW),
          ne="%0.7f,%0.7f" % (latNE, lngNE),
          intent='browse',
          query='',
          categoryId=['4f4528bc4b90abdf24c9de85']

        )
        #Αίτημα στο foursquare API και απάντηση
        resp = requests.get(url=url, params=params)
        data = json.loads(resp.text)
        code=data['meta']['code']
        #code=200 σημαίνει ότι δεν υπάρχει κάποιο σφάλμα, αν δεν είναι 200 τύπωσε τον κωδικό error
        #και το μήνυμα σφάλματος,αλλίως καταχώρησε την απάντηση
        if (code!=200):
            print(data['meta']['code'])
            print(data['meta']['errorDetail'])
            csvfile.close()
            return -1
        venues=data['response']['venues']
        #για διάστημα από 0 έως όσα είναι τα γυμναστήρια, αν τα checkins είναι περισσότερα από 50 γράψε
        #στο αρχείο lat,lng παραθύρου, όνομα, lat,lng γυμναστηρίου και αριθμό checkins
        for i in range(0,len(venues)):
           if (venues[i]['stats']['checkinsCount']>50):
               count=count+1
               gymwriter.writerow([latSW,lngSW,venues[i]['name'],venues[i]
['location']['lat'],
                                   venues[i]['location']['lng'],venues[i][
'stats']['checkinsCount']])
        csvfile.close()
        return 0
```

## Main loop

In [ ]:
```
i=0
#όσο το ΝΑ σημείο είναι μικρότερο από το ΒΔ σε κάθε επανάληψη αποθήκευσε αν το αποτέλεσμα είναι 0 ή -1
while (latSW<latNE):
    lngSWtemp=lngSW
    while (lngSWtemp<lngNE):
        result=query(latSW,lngSWtemp)
        #αν το αποτελεσμα είναι 0 τότε προχώρα το παράθυρο αναζήτησης κατά βήμα, αυξησε το i,
        #ενημέρωσε την progressbar και κοιμήσου 0.8sec
        if (result>=0):
            lngSWtemp=lngSWtemp+inc
            #το i θα αυξανεται κατά 1 κάθε φορά που καλείται η query και θα ενημερώνεται η progressbar
            i=i+1
            bar.update(i,count=count)
            sleep(0.8)
        else:
            #αν το αποτέλεσμα είναι -1 κοιμήσου 4sec
            sleep(4)
    #το ΝΔ lat αυξάνεται πλεόν κατά βήμα
    latSW=latSW+inc
```