Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Πολιτικών Μηχανικών
Τομέας Υδατικών Πόρων και Περιβάλλοντος

Διερεύνηση της επίδρασης του δείκτη ENSO στον πλημμυρικό κίνδυνο και τις αιτήσεις αποζημίωσης στις ΗΠΑ

Κ.Χ.Τσολακίδης

Επιβλέπουσα: Θεανώ Ηλιοπούλου, Διδάκτωρ ΕΜΠ, ΕΔΙΠ

ΠΙΝΑΚΑΣ ΠΕΡΙΕΧΟΜΕΝΩΝ

Κατάλογος Εικόνων

Κατάλογος Πινάκων

Κατάλογος Διαγραμμάτων

Περίληψη

Η παρούσα πτυχιακή εργασία πραγματεύεται τη διερεύνηση της επίδρασης των φαινομένων ENSO (El Niño–Southern Oscillation) σε σχέση με τα πλημμυρικά φαινομένα στις Ηνωμένες Πολιτείες, καθώς και τη πίθανή σύνδεσή τους με τα αιτήματα αποζημιώσεων του Εθνικού Προγράμματος Ασφάλισης Πλημμυρών (NFIP). Δεδομένης της αυξανόμενης συχνότητας και σφοδρότητας των καιρικών φαινομένων εξαιτίας των κλιματικών μεταβολών, η εργασία επιχειρεί να ποσοτικοποιήσει το βαθμό συσχέτισης μεταξύ δεικτών των μελετώμενων δεικτών ENSO και καταγεγραμμένων οικονομικών μεγεθών τόσο σε επίπεδο Πολιτειών όσο Κομητειών στις Η.Π.Α. Ιδιαίτερη έμφαση δίνεται στη ανάλυση της Πολιτείας της Καλιφόρνια, η οποία παρουσιάζει αυξημένη ευαισθησία στα φαινόμενα El Niño.

Η μεθοδολογία, που αναπτύχθηκε βασίστηκε στην αξιοποίηση πολλαπλών βάσεων δεδομένων (ENSO indices από NOAA, US-CAMELS streamflow data, COBE SST, DEM, NHD, OSM, US Census), με σκοπό την εξαγωγή γεωχωρικών και φυσικών χαρακτηριστικών όπως η υδρογραφική και οδική πυκνότητα, το μέσο υψόμετρο, η απόσταση από τη θάλασσα και οι συντεταγμένες κεντροειδών Κομητειών και ο πληθυσμός. Ακόμα έλαβε χώρα ανάλυση τους με τη χρήση στατιστικών εργαλείων όπως ο συντελεστής συσχέτισης Pearson και η μέθοδος της Υπέρβασης Κατωφλιού (threshold overcome analysis) για διαφορετικές τιμές κατώφλιων (90–99%), καθώς και την ανάπτυξη ενός μοντέλου μηχανικής εκμάθησης τύπου, με προβλεπόμενη τιμή τα αιτήματα αποζημίωσεων ανά 100.000 κατοίκους.

Τα αποτελέσματα της ανάλυσης υποδεικνύουν ότι η συσχέτιση δεικτών ENSO με τις παροχές υδάτων (streamflow extremes) είναι πολύ μεγαλύτερη από τη συσχέτιση με τα Ασφαλιστικά Αιτήματα Αποζημιώσεων, γεγονός που εξηγείται από την επιρροή κοινωνικοοικονομικών παραγόντων στην διαδικασία καταγραφής αιτημάτων αποζημίωσης. Η Πολιτεία της Καλιφόρνια σημειώνει τη μεγαλύτερη θετική συσχέτιση μεταξύ του δείκτη μέγιστης ετήσιας τιμής ENSO και αιτημάτων αποζημίωσης (MAX_ENSO–claims) μεταξύ όλων των Πολιτειών ($r \approx 0.35$), ενώ το αναπτυχθέν μοντέλο CatBoost πέτυχε ικανοποιητική ακρίβεια ($R^2 = 0.638$) μέσω της χρήσης

στατικών και δυναμικών χαρακτηριστικών. Η μελέτη συμπεραίνει ότι οι δείκτες ENSO είναι καλό να ενσωματώνονται σε εργαλεία πρόβλεψης κινδύνου.

Η εργασία προτείνει τη μελλοντική εφαρμογή της μεθοδολογίας σε περισσότερες Πολιτείες ή και στο σύνολο των Η.Π.Α, καθώς και την ενσωμάτωση επιπλέον δεδομένων εισαγωγής, με στόχο τη βελτιστοποίηση της ακρίβειας των μοντέλων πρόβλεψης.

Abstract

This thesis investigates the influence of ENSO (El Niño–Southern Oscillation) phenomena in accorance to extreme flood events in the United States, as well as their potential connection to flood insurance claims (National Flood Insurance Program (NFIP)). Recently an increasing frequency of extreme weather events has been observed, so this study aims to quantify the correlation between ENSO indicators and recorded economic data, at State and County level across the U.S.A. The thesis emphasizes on the analysis of the state of California, which is the most sensitive to El Niño events.

The methodology is based on the use of multiple datasets (ENSO indices from NOAA, US-CAMELS streamflow data, COBE SST, DEM, NHD, OSM, US Census), with the aim of extracting geospatial and physical features such as hydrographic and transport (road) density, mean elevation, distance to the sea, county centroid coordinates, and population. These features were analyzed by statistical tools ,such as the Pearson correlation coefficient and the Threshold Overcome Analysis method, applied across various thresholds (90–99%). Also, this thesis includes the development of a machine learning model, aiming to predict insurance claim counts per 100,000 residents.

The analysis reults show that the correlation between ENSO indices and streamflow data is significantly stronger than the correlation with insurance claim records. This is explained by the big impact of social and economic factors on the claims filling procedure. The state of California has the highest positive correlation between the maximum annual ENSO index (MAX_ENSO) and insurance claims ($r \approx 0.35$). The CatBoost model, we developed, achieved good accuracy ($R^2 = 0.638$) with the use of static and dynamic features, as well. The study concludes that ENSO indices can contribute in the flood risk prediction process.

Some good future proposals are the appliance of the same methodology to additional states or the whole USA and the addition of new input features, in order to maximize performance of the model.

1. Εισαγωγή

1.1 Σκοπιά Ερευνητή

Κατά τις τελευταίες δεκαετίες, τα περιβαλλοντικά θέματα ανέρχονται προοδευτικά στη λίστα ενδιαφέροντος της παγκόσμιας επιστημονικής κοινότητας, εξαιτίας της εμφάνισης περιβαλλοντικών καταστροφών με μεγάλες κοινωνικές επιπτώσεις. Έναν από τους τρόπους εμφάνισης των φαινομένων αυτών αποτελούν τα ακραία καιρικά φαινόμενα με κύριο και συχνότερο εκφραστή τις πλημμύρες. Ο κίνδυνος από τα ακραία γεγονότα αποτελεί ανασταλτικό παράγοντα για την ανάπτυξη ποικίλων τομέων παραγωγής με κυριότερο τον πρωτογενή, της γεωργίας δηλαδή και, φαίνεται, ότι η ασφάλιση είναι στρατηγικά σημαντική για την αντιμετώπιση αυτού του κινδύνου. Ειδικότερα, η ασφάλιση απόδοσης καλλιεργειών αγοράζεται από τους γεωργούς και συχνά επιδοτείται από τις κυβερνήσεις για να τους προστατεύσει από την απώλεια των καλλιεργειών τους λόγω φυσικών καταστροφών, όπως οι ακραίες πλημμύρες. (Jonkman, S. N. (2005)) Οι κίνδυνοι όμως, που προκαλούνται από τις φυσικές καταστροφές δεν περιορίζονται στη γεωργία. Αντίθετα συμπεριλαμβάνουν καταστροφή κτιρίων (συμπεριλαμβανομένων κατοικιών) και το κυριότερο απώλειες ανθρωπίνων ζωών. Αναλύοντας τα φαινόμενα των φυσικών καταστροφών γενικότερα, παρατηρείται ότι η αύξηση του πληθυσμού, η οικονομική ανάπτυξη και η αμέλεια για τους κινδύνους που ελλοχεύουν λόγω έντονης αστικοποίησης συχνά αυξάνουν την έκθεση σε φυσικές καταστροφές, συμπεριλαμβανομένωνυ των πλημμυρών. Οι αστικές πλημμύρες είναι πιο δύσκολο να διαχειριστούν, καθώς η υψηλότερη πυκνότητα πληθυσμού και περιουσιακών στοιχείων στο αστικό περιβάλλον αυξάνει τις περιβαλλοντικές και κοινωνικές επιπτώσεις των πλημμυρών και καθιστά τις πιθανές ζημιές από τέτοια φαινόμενα πιο κοστοβόρες. Ως εκ τούτου, η ανάγκη για μια ολοκληρωμένη πολιτική ασφάλισης πλημμυρών και ακριβής αξιολόγηση του κινδύνου πλημμύρας είναι έκδηλη, προκειμένου να μειωθούν οι οικονομικές συνέπειες από τα ακραία πλημμυρικά φαινόμενα, που όπως προαναφέρθηκε σε πολλές περιπτώσεις απειλούν την περιβαλλοντική, κοινωνική και οικονομική ισορροπία.

Αρκετές πρόσφατες μελέτες υποδεικνύουν ένα συντελεστή βαρύνουσας σημασίας, που επηρεάζει και πολλαπλασιάζει την ένταση των φυσικών καταστροφών. Πρόκειται για τα φαινόμενα ENSO (El Niño-Southern Oscillation) τα οποία πλήττουν τα υπεράκτια νερά στον κεντρικό και ανατολικό Ειρηνικό, προκαλώντας διαταραχές στα ατμοσφαιρικά ρεύματα και επηρεάζοντας τα μοτίβα των βροχοπτώσεων και των θερμοκρασιών παγκοσμίως. Τα φαινόμενα ENSO τείνουν να προκαλούν επιπτώσεις στον πρωτογενή τομέα παραγωγής, και τη δημόσια υγεία γενικότερα, καθώς επηρεάζουν τον καιρό και το κλίμα σε παγκόσμια κλίμακα.

Σχετικά με τον κεντρικό ρόλο της ασφάλισης πλημμυρών για τις κοινωνίες και τα άτομα ως εργαλείο για την αντιστάθμιση του κινδύνου οικονομικών απωλειών λόγω φυσικών καταστροφών, αυτή η μελέτη διερευνά την επίδραση που έχουν τέτοια φαινόμενα μεταξύ άλλων σε ασφαλιστικά μεγέθη, συνδυάζοντας δεδομένα κλιματικών δεικτών και εξετάζοντας παράλληλα συμπεράσματα που προκύπτουν από μετά από εφαρμογή τεχνικών μηχανικής μάθησης.

Με δεδομένα όλα τα παραπάνω , ο στόχος αυτής της έρευνας είναι η εφαρμογή μιας στοχαστικής προσέγγισης για υδρολογικά ακραία φαινόμενα με έμφαση στις πλημμύρες, εξετάζοντας την επιρροή των φαινομένων ENSO τόσο στις ακραίες παροχές όσο και στα τελικά αιτήματα προς αποζημίωση έναντι πλημμύρας. Στόχος είναι η διερεύνηση και κατανόηση της επίδρασης τόσο στον πλημμυρικό κίνδυνο (hazard) όσο και στην τελική διακινδύνευση (risk) όπως εκφράζεται μέσα από τα αιτήματα για πλημμυρικές αποζημιώσεις.

1.2 Δομή Εργασίας

Η εργασία διαρθρώνεται σε 8 διακριτά κεφάλαια, τα οποία έχουν ταξινομηθεί με τρόπο τέτοιο, ώστε να ευνοείται η βαθύτερη κατανόηση του περιεχομένου και η επίτευξη του στόχου της.

Στο πρώτο κεφάλαιο παρουσιάζουμε έναν πρόλογο στο θέμα, το γενικό πλαίσιο στο οποίο εισάγεται αυτή η μελέτη, καθώς και τη δομή της.

Στο δεύτερο κεφάλαιο πραγματοποιούμε εισαγωγή στον τρόπο ανάπτυξης των φαινομένων ENSO. Παρουσιάζονται οι μηχανισμοί με τους οποίους αυτά τα φαινόμενα επηρεάζουν τα υδρολογικά στοιχεία μιας περιοχής μελέτης και τα οποία σχετίζονται με την πρόκληση φυσικών καταστροφών. Ακόμη αναφέρονται ιστορικά τα εντονότερα εμφανιζόμενα φαινόμενα και οι συσχετιζόμενες καταστροφές με σκοπό την κατανόηση της σημασίας τους.

Στο τρίτο κεφάλαιο παραθέτουμε τις βάσεις δεδομένων, στις οποίες στηριχτήκαμε κατά τη διάρκεια της πτυχιακής εργασίας και αντλήσαμε τα πρωτογενή δεδομένα. Αναλυτικότερα αναπτύχθηκαν πολλοί κώδικες κάνοντας χρήση διαφορετικών εφαρμογών (QGIS , προγραμματισμός σε γλώσσα Python, κ.α.)

Στο τέταρτο κεφάλαιο περιγράφεται η Μεθοδολογία και οι τρόποι υπολογισμού των διαφορετικών μεταβλητών της ανάλυσης. Στα διαφορετικά στάδια της Εργασίας, έγινε επεξεργασία μεγάλης ποικιλίας δεδομένων με σκοπό την εξαγωγή σχετικών συμπερασμάτων. Το σύνολο των θεωρητικών και στατιστικών εργαλείων, που χρησιμοποιήθηκαν περιγράφονται σε αυτή την ενότητα.

Στο πέμπτο κεφάλαιο γίνεται η ανάλυση των μεταβλητών – δεδομένων, τα οποία χειριζόμαστε και επεξεργαζόμαστε για να διαμορφωθούν ενιαίες βάσεις δεδομένων, οι οποίες εισήχθησαν στο μοντέλο μηχανικής μάθησης.

Στο έκτο κεφάλαιο αναλύεται η διαδικασία διαμόρφωσης του μοντέλου μηχανικής μάθησης, με στόχο την πρόβλεψη των Αιτημάτων Αποζημίωσης Λόγω Πλυμηρρικών Καταστροφών. Περιγράφεται η επιλογή μεταβλητών, τα στάδια εκπαίδευσης, και θεωρητικά στοιχεία για τους διαφορετικούς τύπους μοντέλων που εξετάστηκαν (βαθμός απόδοσης κ.α).

Στο έβδομο κεφάλαιο παρουσιάζονται τα αποτελέσματα της ανάλυσης τόσο της επιρροής του δείκτη ENSO στις ακραίες πλημμύρες όσο και στα αιτήματα προς αποζημίωση έναντι πλημμυρών ενώ παρουσιάζονται και τα αποτελέσματα της διαμόρφωσης του μοντέλου μηχανικής εκμάθησης.

Στο όγδοο κεφάλαιο καταλήγουμε στα συμπεράσματα της παρούσας ερευνητικής εργασίας. Ακόμα περιλαμβάνει προτάσεις για βελτιώσεις, μελλοντική έρευνα και επέκτασης της μεθοδολογίας που εφαρμόστηκε.

2. Φαινόμενα ENSO (El Niño-Southern Oscillation)

2.1 Επισκόπηση φαινομένων ENSO

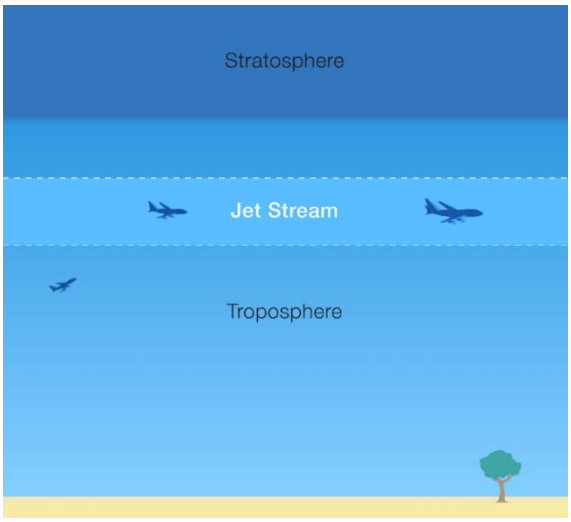
Τα φαινόμενα ENSO (El Niño-Southern Oscillation) αποτελούν μια φυσική διακύμανση της θερμοκρασίας της επιφάνειας της θάλασσας στον Ειρηνικό Ωκεανό, η οποία επηρεάζει το παγκόσμιο κλίμα. Διακρίνεται σε δύο φάσεις: Το *El Niño* είναι η θερμή φάση του φαινομένου που χαρακτηρίζεται από υψηλότερες από το κανονικό θερμοκρασίες της θάλασσας και επηρεάζει τα μοτίβα βροχοπτώσεων παγκοσμίως, συχνά προκαλώντας ξηρασίες και πλημμύρες σε διάφορες περιοχές. Αντίθετα, η *La Niña*, η ψυχρή φάση, προκαλεί ψυχρότερες θαλάσσιες θερμοκρασίες και μπορεί να οδηγήσει σε έντονες βροχοπτώσεις και υψηλότερες από το κανονικό θερμοκρασίες σε άλλες περιοχές, όπως οι νότιες Η.Π.Α. (πχ Τέξας, Φλόριντα). Τα φαινόμενα ENSO έχουν σημαντικό αντίκτυπο σε κλιματικά φαινόμενα όπως τυφώνες, ξηρασίες και πλημμύρες, επηρεάζοντας τη γεωργία, τη διαχείριση υδατικών πόρων και τη δημόσια υγεία σε παγκόσμιο επίπεδο (NOAA, 2021).

Η βαθύτερη κατανόηση των φαινομένων ENSO καθώς και του μηχανισμού γέννησης και λειτουργίας τους απαιτεί την προηγούμενη εμπέδωση της έννοιας των αεροχειμάρων. Ως αεροχείμαροι (jet streams) ορίζονται τα ρεύματα αερίων μαζών , που εντοπίζονται στην ανώτερη τροπόσφαιρα και την κατώτερη στρατόσφαιρα, και έχουν εκατοντάδες χιλιόμετρα πλάτος και μόνο λίγα χιλιόμετρα πάχος. Το κέντρο του ρεύματος αερίων μαζών, το οποίο ονομάζεται "πυρήνας ρεύματος" (jet core), συχνά υπερβαίνει τα 50 μέτρα ανά δευτερόλεπτο και μπορεί περιστασιακά να φτάσει ταχύτητες ανέμου της τάξεως των 100 μέτρων ανά δευτερόλεπτο (Koch et al., 2006; Schiemann et al., 2009; Ahrens and Henson, 2018).

Αναλυτικότερα, οι αεροχείμαρροι δημιουργούνται όταν οι θερμές αέριες μάζες συναντούν τις ψυχρές αέριες μάζες στην ατμόσφαιρα. Ο Ήλιος δεν θερμαίνει ομοιόμορφα όλη τη Γη. Γι' αυτό οι περιοχές κοντά στον ισημερινό είναι ζεστές και οι περιοχές κοντά στους πόλους είναι κρύες. Έτσι, όταν οι θερμότερες αέριες μάζες της Γης συναντούν τις πιο δροσερές αέριες μάζες, ο θερμότερος αέρας ανυψώνεται

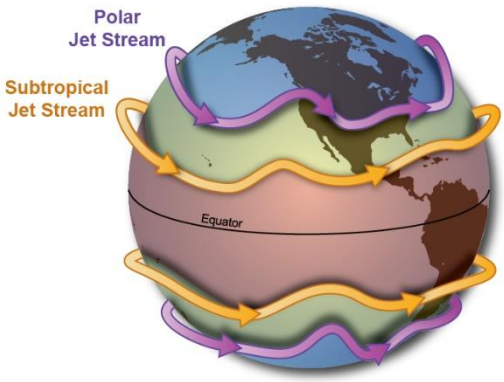
ψηλότερα στην ατμόσφαιρα ενώ ο δροσερός αέρας κατεβαίνει για να αντικαταστήσει τον θερμό αέρα.

Αυτή η κίνηση δημιουργεί ένα ρεύμα. Μια αερορροή/αεροχείμμαρος (jet stream) είναι ένας τύπος ρεύματος που δημιουργείται ψηλά στην ατμόσφαιρα. (NOAA/JPL-Caltech)



ΕΙΚΟΝΑ 2. 1- ΥΨΟΜΕΤΡΙΚΗ ΘΕΣΗ ΑΕΡΟΧΕΙΜΑΡΡΩΝ (JET STREAMS)

(NOAA/JLP-CALTECH)

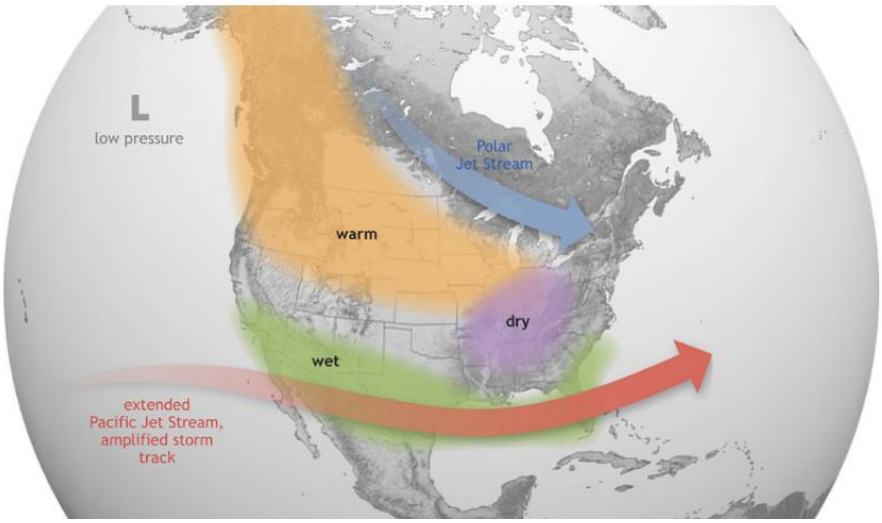


ΕΙΚΟΝΑ 2. 2- ΟΙ ΑΕΡΟΧΕΙΜΑΡΡΟΙ ΤΗΣ ΓΗΣ

(ΠΗΓΗ: HTTPS://WWW.NOAA.GOV/JETSTREAM/GLOBAL/JET-STREAM)

Στη Γη υπάρχουν 4 κύριοι αεροχείμαρροι: 2 πολικοί και 2 υποτροπικοί. Τα φαινόμενα ENSO πρωτίστως επηρεάζουν την μετακίνηση του Αεροχείμαρρου του Ειρηνικού ποικιλοτρόπως. Οι διαφορετικοί τρόποι , με τους οποίους επηρεάζεται η μετακίνηση

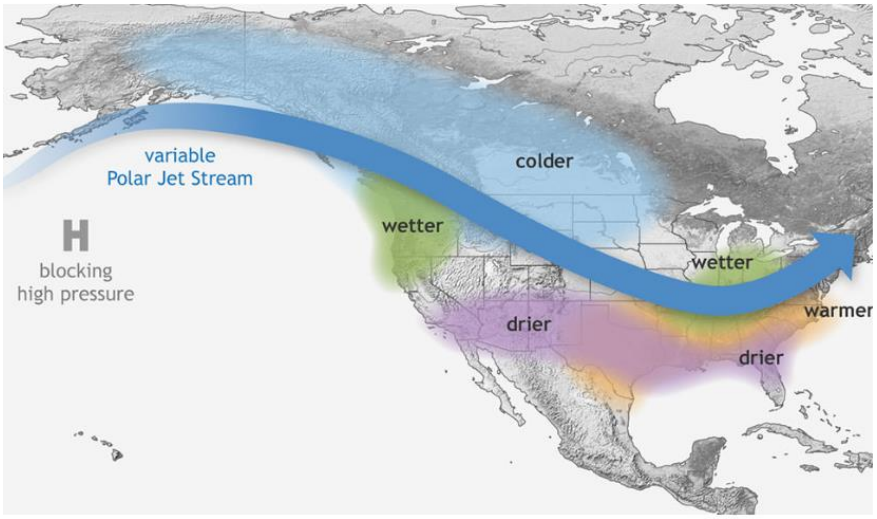
αυτή αναλύονται παρακάτω. Το φαινόμενο-El Niño προκαλεί τη μετακίνηση του αεροχειμάρρου του Ειρηνικού Ωκεανού νότια και πιο ανατολικά. Κατά τον χειμώνα, αυτό οδηγεί σε πιο υγρές συνθήκες από το κανονικό στο νότιο Η.Π.Α. και σε θερμότερες και πιο ξηρές συνθήκες στον Βορρά.



ΕΙΚΟΝΑ 2.3 EL NINO - ΜΕΤΑΚΙΝΗΣΗ ΡΕΥΜΑΤΟΣ ΕΙΡΗΝΙΚΟΥ ΩΚΕΑΝΟΥ

(ΠΗΓΗ: HTTPS://OCEANSERVICE.NOAA.GOV/FACTS/NINONINA.HTML)

Αντιθέτως, το φαινόμενο-La Niña έχει το αντίθετο αποτέλεσμα. Κατά τη διάρκεια των γεγονότων La Niña, οι αεροχείμαρροι είναι ακόμη ισχυρότεροι από το συνηθισμένο, σπρώχνοντας περισσότερο θερμό νερό προς την Ασία. Στα ανοιχτά της δυτικής ακτής της Αμερικής, η ανατροφοδότηση αυξάνεται, φέρνοντας κρύο, πλούσιο σε θρεπτικά συστατικά νερό στην επιφάνεια.



ΕΙΚΟΝΑ 2.4- LA NINA - ΜΕΤΑΚΙΝΗΣΗ ΡΕΥΜΑΤΟΣ ΕΙΡΗΝΙΚΟΥ ΩΚΕΑΝΟΥ (ΠΗΓΗ: HTTPS://OCEANSERVICE.NOAA.GOV/FACTS/NINONINA.HTML)

2.2 Ιστορικές Εκδηλώσεις και Επιπτώσεις

Μια σύντομη ανασκόπηση των πρόσφατων καταστροφικών επιπτώσεων των φαινομένων ENSO μπορεί να εκφράσει σε σαφέστερο τρόπο τη σημασία τους. Στον παρακάτω πίνακα 2.1 διακρίνονται σε κατηγορίες ανάλογα την ένταση των φαινομένων:

ΠΙΝΑΚΑΣ 2.1- ΠΙΝΑΚΑΣ ΚΑΤΑΝΟΜΗΣ ΕΚΔΗΛΩΣΕΩΝ ΦΑΙΝΟΜΕΝΩΝ ENSO ΣΕ ΚΑΤΗΓΟΡΙΕΣ (HTTPS://GGWEATHER.COM/ENSO/ONI.HTM)

El Niño - 27				La Niña - 25		
Weak - 11	Moderate - 7	Strong - 6	Very Strong - 3	Weak - 12	Moderate - 6	Strong - 7
1952-53	1951-52	1957-58	1982-83	1954-55	1955-56	1973-74
1953-54	1963-64	1965-66	1997-98	1964-65	1970-71	1975-76
1958-59	1968-69	1972-73	2015-16	1971-72	1995-96	1988-89
1969-70	1986-87	1987-88		1974-75	2011-12	1998-99
1976-77	1994-95	1991-92		1983-84	2020-21	1999-00
1977-78	2002-03	2023-24		1984-85	2021-22	2007-08
1979-80	2009-10			2000-01		2010-11
2004-05				2005-06		
2006-07				2008-09		
2014-15				2016-17		
2018-19				2017-18		
				2022-23		

Οι ισχυρότερες ιστορικά καταγεγραμμένες εκδηλώσεις των φαινομένων, όπως αυτές προέκυψαν συνθέτοντας και αναλύοντας πληροφορίες από πληθώρα επίσημων φορέων (National Oceanic and Atmospheric Administration (NOAA), World Meteorological Organization (WMO), Intergovernmental Panel on Climate Change (IPCC)) παρουσιάζονται παρακάτω:

- **El Niño 1982-1983**: Ένα από τα ισχυρότερα καταγεγραμμένα επεισόδια El Niño, που προκάλεσε σοβαρές πλημμύρες στη Νότια Αμερική και εξαιρετικές ξηρασίες στην Αυστραλία και την Ινδονησία. Η οικονομική ζημιά ήταν τεράστια, με εκτεταμένες καταστροφές σε υποδομές και αγροτικές καλλιέργειες.

- **El Niño 1997-1998**: Αυτό το επεισόδιο θεωρείται ένα από τα πιο έντονα του 20ού αιώνα και είχε παγκόσμιες επιπτώσεις, όπως πυρκαγιές στην Ινδονησία, ισχυρές πλημμύρες στην Καλιφόρνια και επιδεινωμένα καιρικά φαινόμενα από την Ανατολική Αφρική μέχρι τη Νότια Αμερική
- **La Niña 1998-2000**: Αμέσως μετά το μεγάλο El Niño του 1997-1998, μια έντονη La Niña ακολούθησε, προκαλώντας έντονες και καταστροφικές πλημμύρες στην Ανατολική Αφρική και ισχυρές ξηρασίες στον δυτικό Ειρηνικό και τις ΗΠΑ, με σημαντικές επιπτώσεις στη γεωργία και την παροχή νερού.
- **El Niño 2015-2016**: Ένα από τα ισχυρότερα επεισόδια στην πρόσφατη ιστορία, που προκάλεσε έντονες καταιγίδες, πλημμύρες και ξηρασίες σε πολλές περιοχές του κόσμου, όπως την Ανατολική Αφρική, την Αυστραλία και τη Νότια Αμερική.

Όλα τα παραπάνω φαινόμενα καθώς και φαινόμενα ENSO, τα οποία υπάγονται σε κατηγορίες μικρότερης έντασης επιφέρουν εξαιρετικά μεγάλο οικολογικό πλήγμα στον πλανήτη μας. Αναλυτικότερα, κατά τη διάρκεια των φαινομένων ENSO των ετών 1982–83, 1997–98 και 2015–16, μεγάλες εκτάσεις τροπικών δασών βίωσαν μια παρατεταμένη ξηρασία που οδήγησε σε εκτεταμένες πυρκαγιές, και δραστικές αλλαγές στη δομή των δασών και τη σύνθεση της χλωρίδας στα δάση του Αμαζονίου και του Βορνέο (França, F., Ferreira, J., Vaz-de-Mello, F.Z., et al. (2020)). Οι επιπτώσεις τους δεν περιορίζονται μόνο στη βλάστηση, καθώς παρατηρήθηκε μείωση των πληθυσμών εντόμων μετά την έντονη ξηρασία και τις φοβερές πυρκαγιές κατά τη διάρκεια του El Niño 2015–16. Παρατηρήθηκαν επίσης μειώσεις στα είδη πτηνών που εξειδικεύονται σε συγκεκριμένους βιοτόπους και είναι ευαίσθητα σε διαταραχές, καθώς και σε μεγάλα φρουγκιβόρα θηλαστικά στα καμένα δάση του Αμαζονίου, ενώ σε μια καμένη δασική περιοχή του Borneo σημειώθηκε προσωρινή εξαφάνιση περισσότερων από 100 ειδών πεταλούδων.

Στα εποχιακά ξηρά τροπικά δάση, τα οποία είναι πιο ανθεκτικά στην ξηρασία, οι ερευνητές (Phillips et al., 2009) διαπίστωσαν ότι η ξηρασία που προκλήθηκε από το El Niño αύξησε τη θνησιμότητα των νεοφυών (seedlings). Σε μια (Kobayashi, N., Nakagawa, M., & Kanzaki, M. (2022))έρευνα που δημοσιεύτηκε τον Οκτώβριο του

2022, οι ερευνητές μελέτησαν εποχιακά ξηρά τροπικά δάση σε εθνικό πάρκο στο Chiang Mai της Ταϊλάνδης για 7 χρόνια και παρατήρησαν ότι το El Niño αύξησε τη θνησιμότητα των νεοφυών ακόμα και στα εποχιακά ξηρά τροπικά δάση και μπορεί να επηρεάσει ολόκληρα τα δάση σε μακροπρόθεσμη βάση.

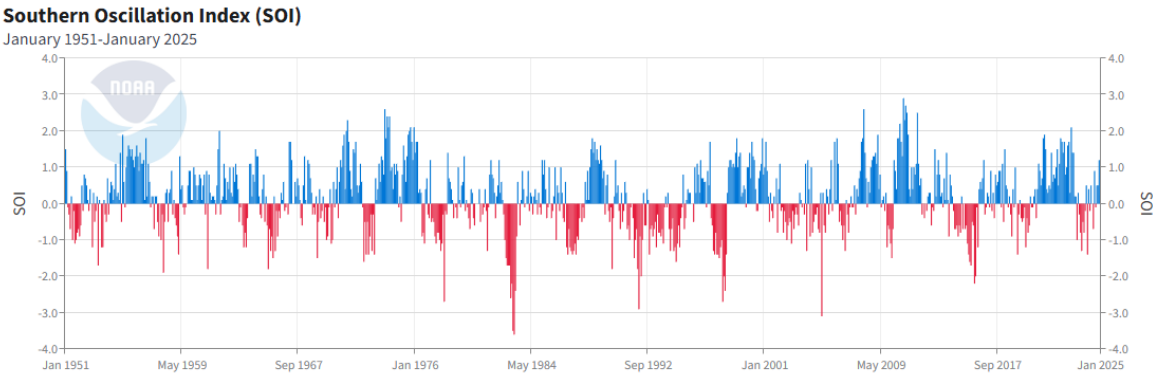
2.3 Δείκτες SOI (Southern Oscillation Index) & ONI (Oceanic Niño Index): Περιγραφή και Χρήση

Όπως προκύπτει από τα παραπάνω απαιτείται ένας τρόπος ποσοτικοποίησης της έντασης των φαινομένων. Αυτό επιτυγχάνεται με την χρήση των δεικτών (NOAA, 2024):

- SOI (Southern Oscillation Index)
- ONI (Oceanic Niño Index)

2.3.1. Δείκτης SOI (Southern Oscillation Index)

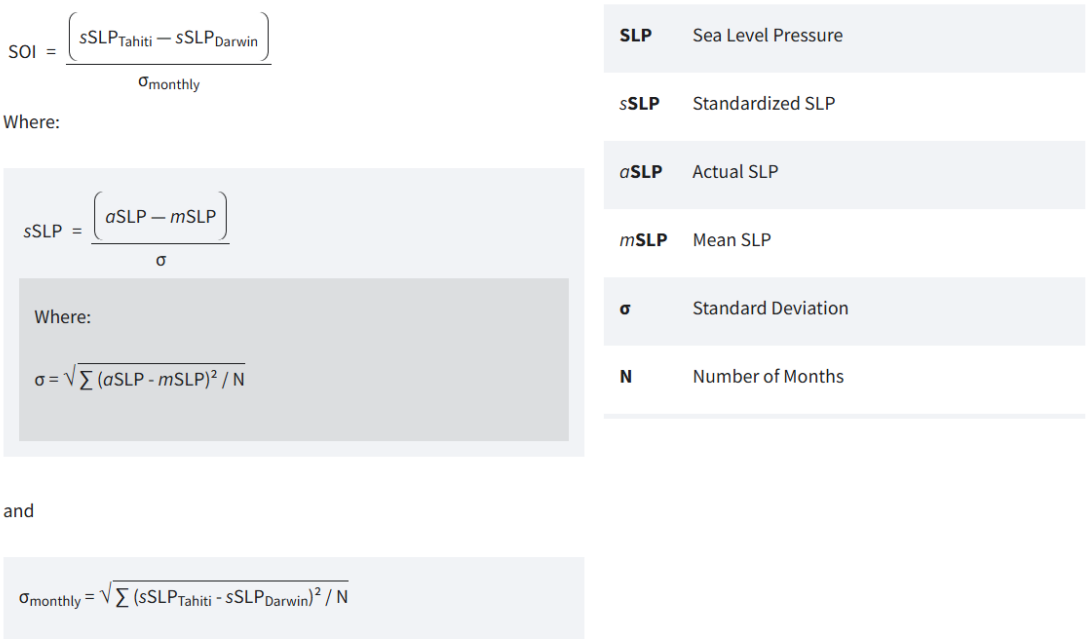
Ο Δείκτης SOI είναι ένας τυποποιημένος δείκτης που βασίζεται στις παρατηρούμενες διαφορές στην πίεση της θάλασσας στην επιφάνεια της θάλασσας (SLP, Sea Level Pressure) μεταξύ των περιοχών της Ταϊτή και του Δαρβίνου, στην Αυστραλία. Ο δείκτης SOI αποτελεί μία μέτρηση διακυμάνσεων μεγάλης κλίμακας στην ατμοσφαιρική πίεση μεταξύ του δυτικού και ανατολικού τροπικού Ειρηνικού κατά τη διάρκεια των επεισοδίων El Niño και La Niña. Γενικά, οι χρονοσειρές του δείκτη SOI παρουσιάζουν ανάλογη εικόνα με αυτές των θερμοκρασίες των ωκεανών στο ανατολικό τροπικό Ειρηνικό. Κατά την αρνητική φάση του δείκτη SOI παρατηρείται χαμηλότερη ατμοσφαιρική πίεση στην περιοχή της Ταϊτή και υψηλότερη από τα συνήθη επίπεδα στην περιοχή του Δαρβίνου. Οι παρατεταμένες περίοδοι αρνητικών (θετικών) τιμών SOI συμπίπτουν με θερμά (κρύα) ωκεάνια νερά στο ανατολικό τροπικό Ειρηνικό, που αποτελούν έκφραση των φαινομένων El Niño (La Niña).



ΕΙΚΟΝΑ 2.5 - ΓΡΑΦΙΚΗ ΑΠΕΙΚΟΝΙΣΗ ΔΕΙΚΤΗ SOI (NOAA, 2024)

Παρακάτω παρουσιάζεται συνοπτικά ο τρόπος υπολογισμού του παραπάνω δείκτη(NOAA.2024):

- **Συλλογή Δεδομένων**: Παίρνουμε τις μετρήσεις της ατμοσφαιρικής πίεσης από τα δύο προαναφερθέντα σημεία: (Ταίτη και Δαρβίνο (Darwin))
- **Υπολογισμός Ανωμαλιών Πίεσης**: Υπολογίζουμε τη διαφορά πίεσης μεταξύ των δύο τοποθεσιών, παίρνοντας την ατμοσφαιρική πίεση της Ταϊτή και αφαιρώντας αυτή του Δαρβίνου (Darwin).
- **Κανονικοποίηση**: Η διαφορά πίεσης κανονικοποιείται με βάση το γενικό μέσο όρο και την τυπική απόκλιση αυτής της διαφοράς για έναν ορισμένο μήνα, προσαρμοζόμενη σε μια κλίμακα συνήθως -10 μέχρι +10 (ή σε άλλες κλίμακες ανάλογα με τον οργανισμό που πραγματοποιεί τη μέτρηση).
- **Υπολογισμός SOI**: Ο τελικός δείκτης SOI προκύπτει από τον μέσο όρο των κανονικοποιημένων μηνιαίων διαφορών για ένα συγκεκριμένο χρονικό διάστημα, συνήθως τις τελευταίες 30 ημέρες.



ΕΙΚΟΝΑ 2.6 – ΔΙΑΔΙΚΑΣΙΑ ΥΠΟΛΟΓΙΣΜΟΥ ΔΕΙΚΤΗ SOI

(ΠΗΓΗ: HTTPS://WWW.NCEI.NOAA.GOV/ACCESS/MONITORING/ENSO/SOI)

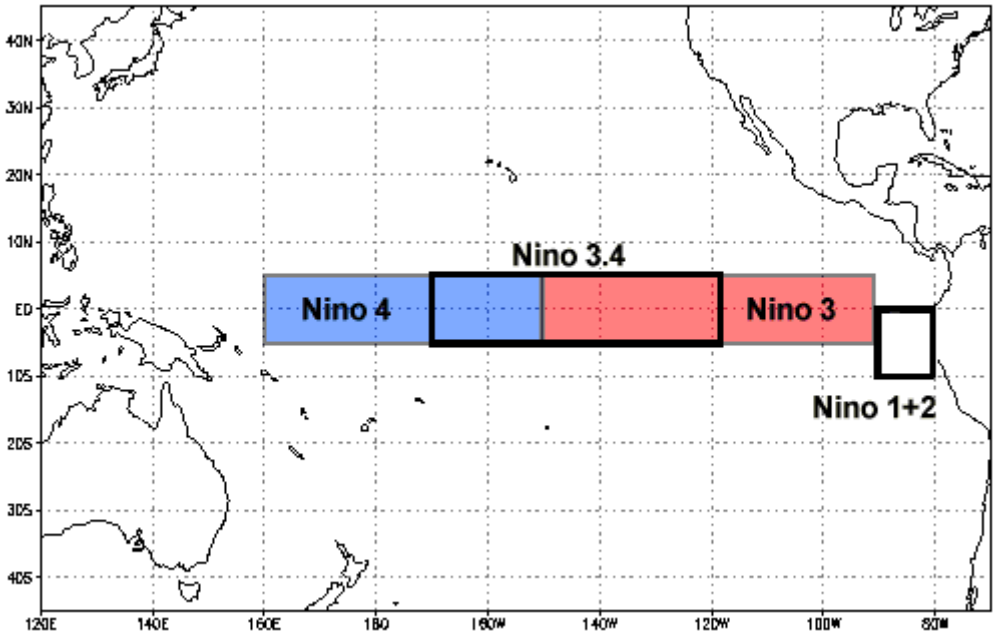
2.3.2. Δείκτης ONI (Oceanic Niño Index)

Ο Δείκτης ONI (Oceanic Niño Index) είναι ένας δείκτης που βασίζεται μονοσήμαντα στις Θερμοκρασίες Επιφάνειας Θάλασσας (SST, Sea Surface Temperature). Αναλυτικότερα, αυτός ο δείκτης ποσοτικοποίησης των φαινόμενων ENSO στον Ειρηνικό Ωκεανό προκύπτει από τη ανάλυση των πέντε συνεχόμενων τριμηνιαίων μέσων τιμών των ανωμαλιών της θερμοκρασίας στην επιφανεία της θάλασσας (SST), στην περιοχή Niño 3.4, με την προϋπόθεση ότι υπερβαίνει κατά απόλυτη τιμή το κατώφλι (threshold) των 0.5°C.

Ιστορικά, οι επιστήμονες έχουν κατηγοριοποιήσει την ένταση των φαινόμενων ENSO με βάση τις ανωμαλίες SST που υπερβαίνουν ένα προεπιλεγμένο κατώφλι σε μια συγκεκριμένη περιοχή του ισημερινού Ειρηνικού. Η πιο συνηθισμένα χρησιμοποιούμενη περιοχή είναι η Niño 3.4, και το πιο συνηθισμένο κατώφλι είναι μια θετική απόκλιση SST από το κανονικό μεγαλύτερη ή ίση με +0.5°C. Τα κριτήρια, τα οποία συχνά χρησιμοποιούνται για την κατηγοριοποίηση των επεισοδίων El Niño, είναι ότι πέντε συνεχόμενες τριμηνιαίες μέσες τιμές SST ανωμαλιών υπερβαίνουν το κατώφλι.

Παρακάτω παρουσιάζεται συνοπτικά ο τρόπος υπολογισμού του παραπάνω δείκτη(NOAA, 2024):

Επιλογή Περιοχής: Ο ONI εστιάζει στην περιοχή Niño 3.4, η οποία εκτείνεται στον κεντρικό Ειρηνικό μεταξύ των γεωγραφικών πλατών 5°Β-5°Ν και των μήκων 170°Δ-120°Δ.



ΕΙΚΟΝΑ 2.7- ΠΕΡΙΟΧΕΣ NINO (NOAA 2024)

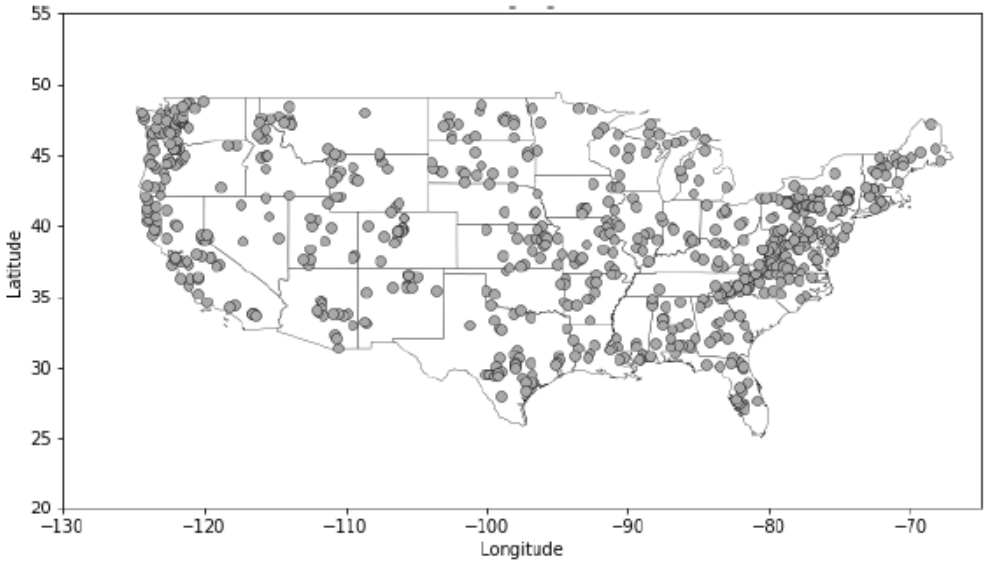
- **Συλλογή Δεδομένων**: Συλλέγονται δεδομένα θερμοκρασίας επιφανειακής θάλασσας (SST, Sea Surface Temperature) για την περιοχή Niño 3.4. Χρησιμοποιείται συνήθως το σύνολο δεδομένων Extended Reconstructed Sea Surface Temperature version 5 (ERSST.v5Huang, B., Thorne, P. W., Banzon, V. F., Boyer, T., Chepurin, G., Lawrimore, J. H., Menne, M. J., Smith, T. M., Vose, R. S., Zhang, H.-M., & Liu, W. (2017)
- **Υπολογισμός Διακυμάνσεων**: Υπολογίζονται οι ανωμαλίες SST, δηλαδή οι αποκλίσεις από τον μέσο όρο θερμοκρασίας, σε σύγκριση με ένα μακροχρόνιο κλιματολογικό διάστημα αναφοράς (μετρήσεις μεταξύ 1981-2010)

- **Μέσες Τριμηνιαίες Τιμές**: Για να εξομαλυνθούν οι βραχυπρόθεσμες διακυμάνσεις, υπολογίζεται η κινούμενη μέση των SST ανωμαλιών κάθε τρεις μήνες.
- **Εφαρμογή Κατωφλίου (Threshold)** : Για να καταταχθεί ένα φαινόμενο ως El Niño ή La Niña, οι τρίμηνες μέσες των ανωμαλιών πρέπει να υπερβαίνουν ένα κατώφλι ±0.5°C.
- **Διάρκεια**: Τέλος, για να επιβεβαιωθεί ένα φαινόμενο ENSO, οι ανωμαλίες πρέπει να πληρούν τα κριτήρια του κατωφλίου για τουλάχιστον πέντε συνεχείς τρίμηνες περιόδους.

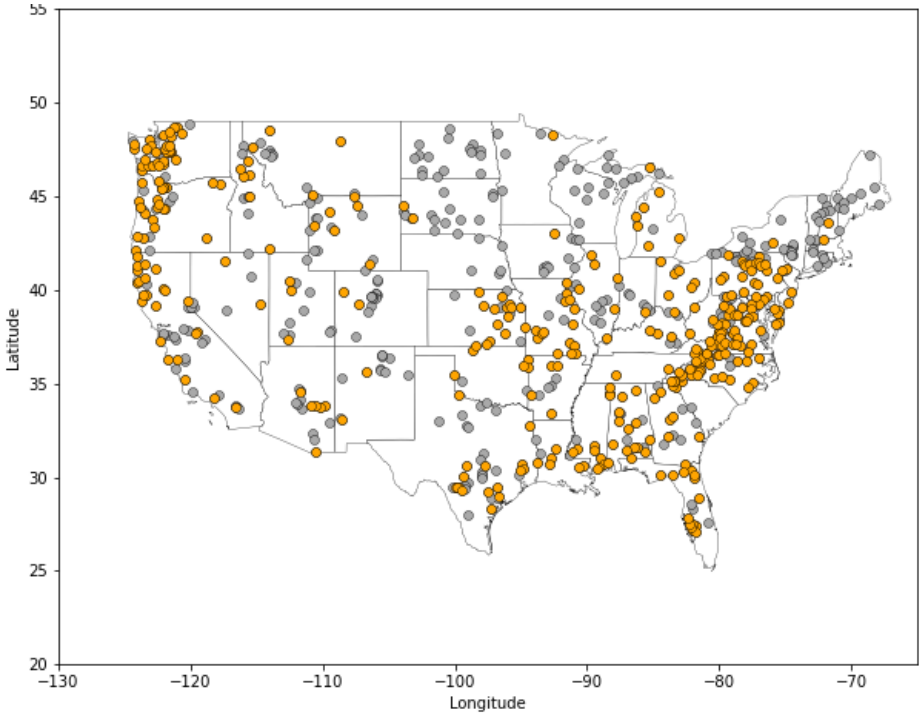
3. Βάσεις Δεδομένων

3.1 US-Camels Dataset

Αυτή η ανάλυση εφαρμόζεται στο σύνολο δεδομένων US-CAMELS, το οποίο περιλαμβάνει 671 ημερήσιες χρονοσειρές ροής ρευμάτων από απορροές στις πολιτείες των Ηνωμένων Πολιτειών (CONUS) που επηρεάζονται ελάχιστα από ανθρώπινες δραστηριότητες (Newman et al., 2014). Από αυτό το σύνολο δεδομένων, επιλέχθηκαν 360 χρονοσειρές ροής ρευμάτων με τη μέγιστη χρονική επικάλυψη (δηλαδή, 35 χρόνια από το 1980 έως το 2014) και λιγότερο από 10% απουσίες τιμών. Η Εικόνα 3.1 δείχνει την περιοχή μελέτης και τις τοποθεσίες των σταθμών μέτρησης ροής για το πλήρες σύνολο δεδομένων και η Εικόνα 3.2 δείχνει τις επιλεγμένες 360 τοποθεσίες των σταθμών μέτρησης ροής. (Papoulakos *et al*., 2025*)*



ΕΙΚΟΝΑ 3.1- ΟΙ 671 ΤΟΠΟΘΕΣΙΕΣ ΣΤΑΘΜΩΝ ΜΕΤΡΗΣΗΣ ΡΟΗΣ US-CAMELS (ΠΗΓΗ: PAPOULAKOS, 2025)



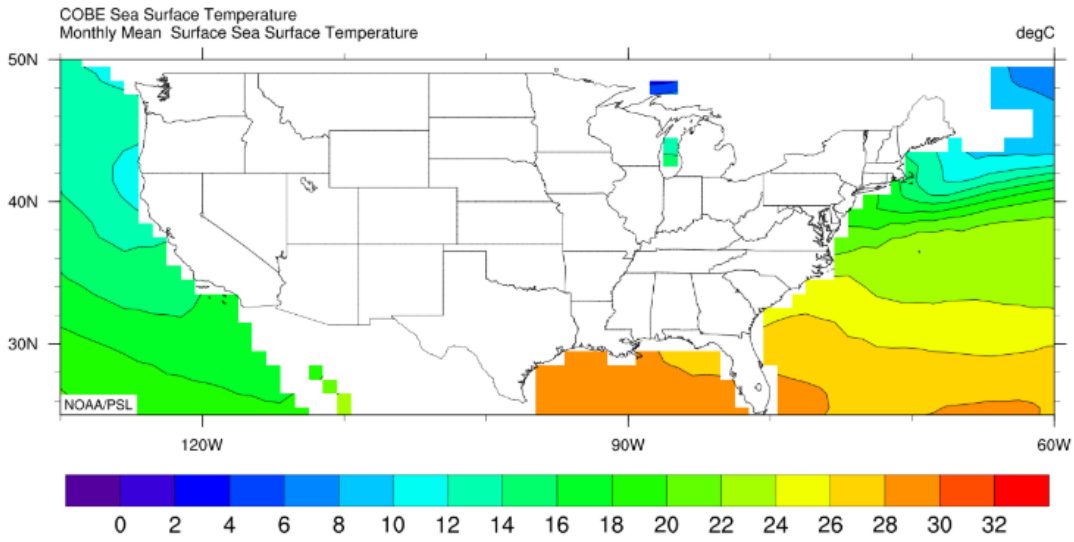
ΕΙΚΟΝΑ 3.2- ΟΙ 360 ΕΠΙΛΕΓΜΕΝΕΣ ΤΟΠΟΘΕΣΙΕΣ ΣΤΑΘΜΩΝ ΜΕΤΡΗΣΗΣ ΡΟΗΣ US-CAMELS (ΠΗΓΗ: PAPOULAKOS, 2025)

3.2. COBE Sea Surface Temperature

Το σύνολο δεδομένων θερμοκρασίας επιφανειακού θαλάσσιου νερού COBE (Centennial in-situ Observation-Based Estimates) (Hirahara, S., Ishii, M., & Fukuda, Y. (2014). παρέχει ιστορικά δεδομένα θερμοκρασίας στην επιφάνεια της θάλασσας (SST), τα οποία είναι κρίσιμα για την έρευνα και την παρακολούθηση των κλιματικών φαινομένων. Το σύνολο δεδομένων COBE συνήθως περιλαμβάνει δεδομένα από το τέλος του 19ου αιώνα και μετά, προσφέροντας ένα μακροχρόνιο ιστορικό αρχείο που βοηθά τους ερευνητές να κατανοήσουν την κλιματική αλλαγή και μεταβλητότητα για περισσότερο από έναν αιώνα. Το υποσύνολο, το οποίο χρησιμοποιείται στην παρούσα εργασία, διαθέτει δεδομένα για τη χρονική περίοδο 1978-2024. Επιλέχτηκε αυτή η χρονική περίοδος για να συμπίπτει χρονικά με δεδομένα άλλων βάσεων και συγκεκρίμενα η βάση δεδομένων Fema (βλ Κεφάλαιο 3.4). Επιπλέον, τα χωρικά όρια ορίστηκαν με τρόπο τέτοιο, ώστε να περιλαμβάνουν τις παραθαλάσσιες περιοχές των Ηνωμένων Πολιταίων Αμερικής. Ακριβέστερα η περιοχή μελέτης οριοθετείται από το

εύρος γεωγραφικού μήκους από 230.5 εως 299.5 και γεωγραφικό πλάτος από 49.5 έως 25.5.

Αξίζει να σημειωθεί ότι το σύνολο δεδομένων ενσωματώνει παρατηρήσεις από διάφορες πηγές, συμπεριλαμβανομένων πλοίων, και δορυφορικές μετρήσεις. Το σύνολο δεδομένων παρουσιάζει μηνιαίες μέσες τιμές των θερμοκρασιών του νερού στην επιφάνεια της θάλασσας.



ΕΙΚΟΝΑ 3.3 - ΠΕΡΙΟΧΗ ΔΕΔΟΜΕΝΩΝ 1.1 COBE SST DATASET ((HIRAHARA, S., ISHII, M., & FUKUDA, Y. (2014))

ΠΊΝΑΚΑΣ 3.1- ΤΜΗΜΑ ΔΕΔΟΜΕΝΩΝ COBE SST DATASET

lon	lat	time	sst
230,5	49,5	1/6/1950	10,62
230,5	48,5	1/6/1950	10,87
230,5	47,5	1/6/1950	11,125
230,5	46,5	1/6/1950	11,59
230,5	45,5	1/6/1950	12,015
231,5	49,5	1/6/1950	10,6775
231,5	48,5	1/6/1950	10,9475

3.3. Δείκτες ENSO

Οι δείκτες ENSO έχουν αναλυθεί λεπτομερώς στην Ενότητα 2.3. Χρησιμοποιήθηκαν τα επίσημα δεδομένα του ινστιτούτου NOAA (National Oceanic and Atmospheric Administration, 2017) Στον πίνακα 3.2 , που ακολουθεί παρατίθεται τμήμα των στοιχείων που χρησιμοποιήθηκαν με σκοπό την παρουσίαση της μορφής των δεδομένων. Συγκεκριμένα, όπως αναφέρθηκε στην Ενότητα 2.3. παρατηρούμε το τριμηνιαίο μέσο όρο των μετρήσεων της θερμοκρασίας στη στάθμη της θάλασσας, καθώς και τις ανωμαλίες SST, τις αποκλίσεις δηλαδή από τον μέσο όρο θερμοκρασίας, σε σύγκριση με ένα μακροχρόνιο κλιματολογικό διάστημα αναφοράς (μετρήσεις μεταξύ 1981-2010)

ΠΙΝΑΚΑΣ 3.2 – ΤΜΗΜΑ ΒΑΣΗΣ ΔΕΔΟΜΕΝΩΝ ΘΕΡΜΟΚΡΑΣΙΑΣ ΣΤΗ ΣΤΑΘΜΗ ΤΗΣ ΘΑΛΑΣΣΑΣ ΚΑΙ SOI – NINO 3.4 (ERSSTV5) (HUANG, B., THORNE, P. W., BANZON, V. F., BOYER, T., CHEPURIN, G., LAWRIMORE, J. H., MENNE, M. J., SMITH, T. M., VOSE, R. S., ZHANG, H.-M., & LIU, W. (2017)

SEAS	YR	TOTAL	ANOM
ASO	1986	27.31	0.71
SON	1986	27.49	0.94
OND	1986	27.64	1.14
NDJ	1986	27.70	1.22
DJF	1987	27.76	1.23
JFM	1987	27.95	1.19
FMA	1987	28.19	1.06
MAM	1987	28.41	0.95
AMJ	1987	28.53	0.97
MJJ	1987	28.59	1.22
JJA	1987	28.54	1.51
JAS	1987	28.45	1.70
ASO	1987	28.24	1.65
SON	1987	28.03	1.48
OND	1987	27.76	1.25
NDJ	1987	27.59	1.11
DJF	1988	27.34	0.81
JFM	1988	27.29	0.54
FMA	1988	27.27	0.14
MAM	1988	27.15	-0.31
AMJ	1988	26.69	-0.88
MJJ	1988	26.08	-1.30
JJA	1988	25.74	-1.30
JAS	1988	25.65	-1.11

3.4 Αρχεία Αιτημάτων Αποζημιώσεων Εθνικού Προγράμματος Ασφάλισης Πλημμυρών (NFIP)

Η Ομοσπονδιακή Υπηρεσία Διαχείρισης Έκτακτης Ανάγκης (FEMA) είναι μια υπηρεσία του Υπουργείου Εσωτερικής Ασφάλειας των Ηνωμένων Πολιτειών, και ο κύριος σκοπός της είναι να συντονίζει την αντίδραση σε καταστροφές που έχουν συμβεί εντός των ΗΠΑ και που υπερβαίνουν τις δυνατότητες των τοπικών και πολιτειακών αρχών.

Το πρόγραμμα επιτρέπει στους ιδιοκτήτες ακινήτων σε κοινότητες που συμμετέχουν να αγοράζουν κρατικά διαχειριζόμενη ασφάλιση έναντι ζημιών από πλημμύρες. Επίσης, απαιτεί την ασφάλιση για δάνεια που εξασφαλίζονται με υφιστάμενα ή υπό κατασκευή κτίρια, εφόσον αυτά βρίσκονται εντός ειδικών ζωνών πλημμυρικού κινδύνου (Special Flood Hazard Area) σε κοινότητες που συμμετέχουν στο πρόγραμμα. Το NFIP στοχεύει να παρέχει μια ασφαλιστική εναλλακτική αντί για οικονομική βοήθεια μετά την καταστροφή, ώστε να καλύπτει το κόστος επισκευής ζημιών σε κτίρια και περιεχόμενο τους (FEMA, 1986).

Στις 11 Ιουνίου 2019, η FEMA δημοσίευσε δεδομένα του NFIP που περιλαμβάνουν πάνω από δύο εκατομμύρια εγγραφές απαιτήσεων από το 1970 και πάνω από 47 εκατομμύρια εγγραφές ασφαλιστηρίων από τα τελευταία δέκα έτη, μέσω της πλατφόρμας OpenFEMA (FEMA, 2019). Αυτά τα δεδομένα εμπλουτίζουν τα ήδη διαθέσιμα δεδομένα του NFIP και προσφέρουν επιπλέον πληροφορίες που ενδιαφέρουν την επιστημονική και πολιτική κοινότητα.

Το δημοσιευμένο dataset επιτρέπει αναλύσεις για την εξέλιξη της κάλυψης πλημμυρικών ασφαλίσεων και καταγραφής απαιτήσεων σε χρονικό βάθος άνω των 40 ετών, περιλαμβάνοντας μεταβλητές όπως: πολιτεία, ζώνη απογραφής (census tract), ZIP code, έτος ζημίας και ποσό πληρωμής. Το σύνολο δεδομένων περιλαμβάνει περίπου 2.4 εκατομμύρια εγγραφές και 39 μεταβλητές. Οι μεταβλητές που χρησιμοποιήθηκαν στη μελέτη μας είναι:

- yearOfLoss: Το έτος κατά το οποίο συνέβη η πλημμύρα (σε μορφή YYYY)

- countyCode: Ο πενταψήφιος FIPS κωδικός της κομητείας
- state: Ο δίγραμμος κωδικός της πολιτείας
- latitude / longitude: Κατά προσέγγιση γεωγραφικές συντεταγμένες του ακινήτου (με ακρίβεια 1 δεκαδικού)

3.5 Γεωχωρικά Δεδομένα:

3.5.1 Χωρικά Όρια Πολιτειών/Κομητειών (TIGER/Line)

Στα πλαίσια της ανάπτυξης τμημάτων κώδικα για τις αναλύσεις προέκυψε η ανάγκη εντοπισμού και χρήσης διακριτών χωρικών ορίων (συνόρων) τόσο των Πολιτειών, όσο και των Κομητειών της εξεταζόμενης περιοχής. Η ανάγκη αυτή ικανοποιήθηκε με τη χρήση των γεωγραφικών όριων των πολιτειών και κομητειών των Ηνωμένων Πολιτειών, όπως αυτά ελήφθησαν από το σύστημα TIGER/Line Shapefiles του U.S. Census Bureau (U.S. Census Bureau. (2023)). Στα αρχεία αυτά περιλαμβάνονται γεωμετρικά και περιγραφικά δεδομένα που καθιστούν σαφή τα διοικητικά όρια των πολιτειών και κομητειών. Αναλυτικότερα, τα δεδομένα αυτά χρησιμοποιήθηκαν τόσο για την απομόνωση περιοχών από τις υπόλοιπες με σκοπό την ξεχωριστή μελέτη τους όσο και για τη οπτικοποίηση άλλων δεδομένων (υδρογραφική-οδική πυκνότητα, συσχετίσεις μεταβλητών σε χωρική βάση, Σύνολο αιτημάτων Αποζημιώσεων κ.α.)

3.5.2 Δίκτυο Υδρογραφίας (NHD)

Μία από τις μεταβλητές, που θα μελετηθεί εκτενώς στην παρούσα πτυχιακή εργασία είναι η Πυκνότητα Υδρογραφικού Δικτύου μια περιοχής (στην περίπτωση μας η Πολιτεία της Καλιφόρνια). Προκειμένου να αντλήσουμε δεδομένα για το χαρακτηριστικό αυτό επιστρατεύτηκε η βάση δεδομένων National Hydrography Dataset (NHD), η οποία παρέχεται από την U.S. Geological Survey (USGS, 2023). Τα δεδομένα αυτά περιγράφουν γραμμικά και επιφανειακά χαρακτηριστικά του υδρογραφικού δικτύου (πχ ποτάμια, ρέματα κλπ). Στην περίπτωση μας βασιστήκαμε στο επίπεδο (layer) NHDFlowline, στο οποίο συγκαταλέγονται όλες οι ροές

επιφανειακών υδάτων. Τα δεδομένα αυτά αξιοποιήθηκαν μέσω της ανάπτυξης τμημάτων κώδικα υπολογισμού του συνολικού μήκους των επιφανειακών αυτών ροών υδρογραφικού δικτύου καταλήγοντας στον υπολογισμό της υδρογραφική πυκνότητα της εξεταζόμενης περιοχής, εκφρασμένη ως χιλιόμετρα ροών ανά τετραγωνικό χιλιόμετρο (km/km²).

3.5.3. Οδικό Δίκτυο (OSM / NTD)

Μία ακόμα από τις μεταβλητές, που θα μελετηθεί εκτενώς στην παρούσα πτυχιακή εργασία είναι η Πυκνότητα Οδικού Δικτύου μια περιοχής (στην περίπτωση μας η Πολιτεία της Καλιφόρνια). Προκειμένου να αντλήσουμε δεδομένα για το χαρακτηριστικό αυτό αξιοποιήθηκαν δύο διαφορετικές πηγές δεδομένων: το OpenStreetMap (OSM - OpenStreetMap contributors. (2023) και το National Transportation Dataset (NTD – USGS, 2023). Τα δεδομένα από το OSM περιλαμβάνουν ένα ενημερωμένο σύνολο γραμμικών χαρακτηριστικών που αποτυπώνουν το οδικό δίκτυο, ενώ το NTD δεδομένα μεταφορών από ομοσπονδιακές υπηρεσίες των ΗΠΑ. Στην παρούσα πτυχιακή εργασία, από τη βάση δεδομένων NTD χρησιμοποιήθηκε το επίπεδο (layer) Trans_Road_Segment. Τα δεδομένα αυτά επίσης περιγράφουν γραμμικά χαρακτηριστικά του οδικού δικτύου (πχ οδούς, λεωφόρους κλπ). Όπως και με τα δεδομένα της προηγούμενης ενότητας έτσι και αυτά αξιοποιήθηκαν μέσω της ανάπτυξης τμημάτων κώδικα υπολογισμού του συνολικού μήκους των επιφανειακών αυτών ροών οδικού δικτύου καταλήγοντας στον υπολογισμό της οδική πυκνότητα της εξεταζόμενης περιοχής, εκφρασμένη ως χιλιόμετρα ροών ανά τετραγωνικό χιλιόμετρο (km/km²), δίνοντας έτσι μια εικόνα σχετικά με τα επίπεδα προσβασιμότητας και συγκοινωνιακής κάλυψης για κάθε Κομητεία.

3.5.4. Ψηφιακό Μοντέλο Υψομέτρου (DEM)

Μία ακόμα πολύτιμη πηγή δεδομένων, από αυτές, που χρησιμοποιήθηκαν στην παρούσα μελέτη αποδείχθηκε το Ψηφιακό Μοντέλο Υψομέτρου (DEM), που προέρχεται από το πρόγραμμα USGS 3D Elevation Program (3DEP) ((USGS, 2023)) και ακριβέστερα από τα αρχεία 1/3 arc-second historical elevation tiles. Το μοντέλο αυτό περιλαμβάνει όλα το σύνολο της έκτασης των Η.Π.Α. και περιέχει δεδομένα υψομετρικών πληροφοριών με ανάλυση περίπου 10 μέτρων. Το Ψηφιακό Μοντέλο Υψομέτρου (DEM) μέσω της ανάπτυξης των απαραιτήτων τμημάτων κώδικα και της εφαρμογής QGIS κατέστησε δυνατή την εξαγωγή μορφολογικών χαρακτηριστικών, όπως η μέση τιμή Υψομέτρου ανά Κομητεία.

3.6. U.S. Census Bureau – Πρόγραμμα Εκτιμήσεων Πληθυσμού (PEP)

Η τελευταία βάση δεδομένων που χρησιμοποιήθηκε κατά το στάδιο της ανάλυσης δεδομένων προέρχεται από το Πρόγραμμα Εκτιμήσεων Πληθυσμού (Population Estimates Program – PEP) του U.S. Census Bureau (U.S. Census Bureau. (2023))**.** Παρέχει ετήσιες εκτιμήσεις πληθυσμού στη βάση τόσο των διαφορετικών Πολιτείες όσο και Κομητειών, λαμβάνοντας υπόψη γεννήσεις, θανάτους και μεταναστευτικές ροές. Το χαρακτηριστικό του πληθυσμού των Κομητειών χρησιμοποιήθηκε κατά κύριο λόγο στο στάδιο της μοντελοποίησης του μοντέλου Μηχανικής Εκμάθησης, επιτρέποντας μας να αυξήσουμε την αξιοπιστία του μέσω της κανονικοποίησης των δηλωθέντων Αιτημάτων Αποζημίωσης κάποιας Κομητείας, επιτρέποντας με τον τρόπο αυτό τη σύγκριση μεταξύ Κομητειών διαφορετικού μεγέθους και πληθυσμιακής βάσης.

4. Μεθοδολογία

4.1. Κυλιόμενοι Μέσοι Όροι

Για την αξιολόγηση φαινομένων που εκτυλίσσονται σε συγκεκριμένα διαστήματα, όπως τα El Niño και La Niña, είναι απαραίτητη η εφαρμογή τεχνικών εξομάλυνσης χρονοσειρών όπως οι Κυλιόμενες Μέσες Τιμές. Η εξομάλυνση αυτή είναι πολύ σημαντική, καθώς δίχως αυτή το δείγμα θα χαρακτηρίζονται από έντονες σημειακές απότομα μεγάλες μεταβολές τιμών. Ο δείκτης ONI χρησιμοποιεί μία κυλιόμενη μέση τιμή τριών μηνών, ο οποίος υπολογίζεται ως:

$$R(t) = \frac{1}{n}\sum_{i=-k}^{k} A(t+i) \tag{4.1}$$

όπου:

- $R(t)$ = Κυλιόμενη μέση τιμή στο χρόνο t
- n=2 k+1 = Μέγεθος διαστήματος, με k να αντιπροσωπεύει τον αριθμό των περιόδων πριν και μετά το t
- A(t+i) = Ανωμαλία στο χρόνο t + i

Αυτή η μεθοδολογία τονίζει τη χρονική εξομάλυνση των δεδομένων, η οποία είναι καθοριστικής σημασίας για την ταυτοποίηση και την επιβεβαίωση των διαρκών κλιματικών μεταβολών που χαρακτηρίζουν τα φαινόμενα El Niño και La Niña κατά τη διάρκεια ενός τρίμηνου.

4.2 Συντελεστής συσχέτισης Pearson

Ο συντελεστής συσχέτισης Pearson (Pearson, 1895) όταν εφαρμόζεται σε ένα δείγμα αναπαρίσταται συνήθως ως r_{xy} και μπορεί να αναφέρεται ως ο συντελεστής συσχέτισης δείγματος ή ο συντελεστής συσχέτισης Pearson δείγματος. Μπορούμε να λάβουμε έναν τύπο για το r_{xy} αντικαθιστώντας εκτιμήσεις των συνδιακυμάνσεων και

των διακυμάνσεων με βάση ένα δείγμα στον παραπάνω τύπο. Δεδομένων $\{(x_1,y_1),\ldots,(x_n,y_n)\}$ που αποτελούνται από n ζεύγη, το r_{xy} ορίζεται ως:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

(4.2)

όπου:

- *n*: Μέγεθος Δείγματος
- x_i, y_i είναι τα επιμέρους δείγματα δεδομένων με δείκτη *i*
- $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Αυτή η εξίσωση δίνει μια στατιστικά αξιόπιστη μέτρηση του βαθμού στον οποίο τα δύο σετ δεδομένων συσχετίζονται γραμμικά, αποτελώντας μια από τις πιο διαδεδομένες λύσεις στη στατιστική ανάλυση για την αξιολόγηση σχέσεων μεταξύ μεταβλητών.

4.3. Ανάλυση Υπέρβασης Κατωφλιού & Μοντέλο Συλλογικού Ρίσκου

Η κατανόηση της συμπεριφοράς των ακραίων τιμών είναι θεμελιώδης στην κλιματική επιστήμη, ειδικά στην αξιολόγηση κινδύνων και επιπτώσεων που σχετίζονται με γεγονότα που δεν εμφανίζονται συχνά αλλά είναι μεγάλης σφοδρότητας. Στην παρούσα ερευνητική εργασία αναλύθηκε η μεταβλητή της παροχής νερού (streamflow) από πολλούς σταθμούς (βλ. κεφάλαιο 3.1). Η ανάλυση αυτή επετεύχθη με τη Μέθοδο Υπέρβασης Κατωφλιού. Αναλυτικότερα, η ανάλυση υπέρβασης κατωφλιών (peak over threshold analysis) είναι μια στατιστική προσέγγιση που χρησιμοποιείται για την αναγνώριση εκδηλώσεων φαινομένων ή τάσεων όταν οι

μελετώμενες μεταβλητές ξεπερνούν προκαθορισμένα κατώφλια. Στην περίπτωσή μας, χρησιμοποιούνται ως όρια υπέρβασης (κατώφλια) το 90%, 95%, 98% και 99%

Όσο υψηλότερο είναι το κατώφλι (π.χ. 99%), τόσο αυξάνεται η εμπιστοσύνη ότι σημειώνεται κάποιο πλημμυρικό φαινόμενο ενώ αντίστοιχα μειώνεται και ο αριθμός των σταθμών, στους οποίους οι μετρημένες παροχές πληρούν αυτό το αυστηρό κριτήριο. Αντίστροφα, η επιλογή χαμηλότερων κατωφλίων (όπως το 90%) επιτρέπουν μεγαλύτερο δείγμα ανάλυσης (περισσότερους σταθμούς), με μικρότερη όμως στατιστική σημασία με όρους ακραίων παροχών. Έτσι επιτυγχάνεται η εξέταση της της παροχής νερού (streamflow) σε διαφορετικά επίπεδα. Στην παρούσα ερευνητική εργασία, η παροχή νερού εκφράζεται μέσω των μεταβλητών του Πλήθους Υπερβάσεων (N) και του Συλλογικού Ρίσκου (Collective Risk) ανά κατώφλι Υπέρβασης (βλ. Κεφάλαιο 2.4).

Αναλυτικότερα, η μεταβλητή Πλήθους Υπερβάσεων (N) αναφέρεται στον ετήσιο αριθμό επεισοδίων όπου η παροχή υπερβαίνει κάποιο προκαθορισμένο ποσοστιαίο κατώφλι (90%, 95%, 98% και 99%) εκφράζοντας πρακτικά την ετήσια συχνότητα των έντονων πλημμυρικών γεγονότων. Το συλλογικό ρίσκο (S) ακολουθεί τη θεωρητική προσέγγιση του συλλογικού μοντέλου κινδύνου (collective risk model), όπως αυτό χρησιμοποιείται στην υδρολογική μοντελοποίηση (Kaas et al., 2008). Στο πλαίσιο αυτό, ορίζεται ως το άθροισμα της «σοβαρότητας» κάθε επιμέρους γεγονότος υπέρβασης εντός ενός έτους:

$$S = \sum_{j=1}^{N} Y_j \tag{4.3}$$

όπου:

Y_j = ένα μέτρο της έντασης ή σοβαρότητας του j-οστού επεισοδίου υπέρβασης (π.χ. απόκλιση ροής από το κατώφλι)

S = συλλογικό ρίσκο

Τα παραπάνω μεγέθη προκύπτουν από την ανάλυση των χρονοσειρών απορροής των 360 υδρολογικών σταθμών από τη Βάση Δεδομένων Camels (βλ. κεφάλαιο 3.1). Η ανάλυση αυτή επιτρέπει την παράλληλη αξιολόγηση της συχνότητας και έντασης των

υδρολογικών φαινομένων, σε ετήσια βάση (Serinaldi & Kilsby, 2016) και αποτελεί το θεμέλιο για τη συσχέτιση με κλιματικούς δείκτες όπως οι δείκτες ENSO στην παρούσα πτυχιακή εργασία.

4.4. Χαρτογραφική Ανάλυση

Ένα ακόμα εργαλείο, που χρησιμοποιήθηκε για την μελέτη και ανάλυση δεδομένων, σχέσεων μεταξύ τους και τη διεξαγωγή συμπερασμάτων ήταν η χαρτογραφική ανάλυση. Αναλυτικότερα, με την οπτικοποίηση των δεδομένων σε αληθινούς χάρτες κατέστη ευκολότερη η σύγκριση των τιμών των μελετώμενων μεταβλητών. Προκειμένου να επιτευχθούν τα παραπάνω έγινε εκτενής χρήση της εφαρμογής QGIS Desktop 3.40.5, μέσω του οποίου και δημιουργήθηκαν οι χρωματικοί χάρτες (choropleths). Σε αυτούς αποτυπώνεται η γεω-χωρική κατανομή διαφορετικών τιμών ανάλογα με την εξεταζόμενη μεταβλητή (πχ τιμές πυκνότητας δικτύων, συσχετίσεων μεταξύ δεικτών ENSO και των επιλεγμένων μεταβλητών, όπως αριθμός Αιτημάτων Αποζημίωσης ή παροχής (streamflow)) σε επίπεδο σταθμού ή πολιτείας. Ακόμα αναπτύχθηκαν χάρτες, κατά τους οποίους σημειώνονται διαφορετικά σημεία (για παράδειγμα θέσεις σταθμών, κεντροειδή κομητειών Πολιτείας) με χρωματική κωδικοποίηση, ανά περίπτωση. Με τον τρόπο αυτό καθίσταται πολύ εύκολη και γρήγορη η αναγνώριση μοτίβων αλλά και των περιοχών με ισχυρότερη επίδραση από το φαινόμενο ENSO.

4.5 Στατιστική Κατανομή Δεδομένων

Στα πλαίσια της παρούσας ερευνητικής εργασίας, για την καλύτερη επεξεργασία των δεδομένων έγινε εκτενή χρήση της στατιστικής ανάλυσης κατανομών δεδομένων μας. Ως η κατανομή ενός συνόλου δεδομένων ορίζεται ο τρόπος με τον οποίο κατανέμονται οι τιμές μιας μεταβλητής. Η εκπόνηση διαγραμμάτων απεικόνισης των κατανομών αυτών (πυκνοκατανομές ή ιστογράμματα) μας επέτρεψε τη διεξαγωγή συμπερασμάτων για το εξεταζόμενο δείγμα. Κατέστησε εύκολη την ανάγνωση του τύπου της κατανομής. Ακριβέστερα μία κατανομή μπορεί να είναι συμμετρική, (πχ κανονική κατανομή), ή ασύμμετρη, έχοντας τις περισσότερες τιμές συγκεντρωμένες

προς μία εκ των δύο πλευρών. Επίσης, κατά την ανάλυση κατανομών εξετάζονται κρίσιμα χαρακτηριστικά μεγέθη (της κατανομής) όπως:

- **Διάμεσος** (η τιμή που διακρίνει το σύνολο των δεδομένων σε δύο ίσα μέρη, όταν αυτά είναι ταξινομημένα κατά αύξουσα ή φθίνουσα σειρά. Οι μισές παρατηρήσεις είναι μικρότερες ή ίσες με τη διάμεσο, ενώ οι άλλες μισές είναι μεγαλύτερες)
- **Μέση τιμή** (άθροισμα όλων των τιμών διαιρεμένο με το πλήθος του δείγματος)
- **Διασπορά** (ο μέσος όρος των τετραγώνων των αποκλίσεων από τη μέση τιμή, εκφράζει τη μεταβλητότητα των διαφορετικών τιμών του δείγματος από τη μέση τιμή)
- **Ακραίες τιμές** (τιμές που αποκλίνουν από το υπόλοιπο σύνολο δεδομένων)

Για τα παραπάνω χαρακτηριστικά μεγέθη παρουσιάζονται διαγράμματα BoxPlots (βλ.Κεφάλαιο 4.6) και κρίνονται καθοριστικής σημασίας για την αναγνώριση μοτίβων και ανωμαλιών και στα δεδομένα.

4.6 Διαγράμματα BoxPlot και Dumbbell

Όπως λοιπόν αναφέρθηκε και στα προηγούμενα κεφάλαια, για την εποπτική αξιολόγηση των αποτελεσμάτων στην εργασία αυτή χρησιμοποιούνταθ τα διαγράμματα BoxPlot και Dumbbell,.

Τα διαγράμματα BoxPlot είναι ένα εργαλείο, που επιτυγχάνει την αποτύπωση της κατανομής ενός δείγματος δεδομένων με τη χρήση μέσω πέντε βασικών στατιστικών δεικτών:

- Την ελάχιστη τιμή
- Το 1ο τεταρτημόριο (Q1)
- Τη διάμεσο (Q2) (βλ. Κεφάλαιο 4.5)
- Το 3ο τεταρτημόριο (Q3)
- Τη μέγιστο τιμή

Ως τεταρτημόριο ορίζεται μία τιμή κάτω από την οποία βρίσκεται ένα ποσοστό των τιμών του δείγματος. Αναλυτικότερα ως πρώτο τεταρτημόριο ορίζεται η τιμή όριο, με τρόπο τέτοιο ώστε το 25% του εξεταζόμενου δείγματος να εμφανίζει μικρότερες τιμές από αυτή. Αντίστοιχα, το δεύτερο τεταρτημόριο (η διάμεσης) είναι η τιμή κάτω από την οποία βρίσκεται το 50% του δείγματος, ενώ το τρίτο τεταρτημόριο είναι η τιμή κάτω από την οποία βρίσκεται το 75% του δείγματος.

Κατά τη διαμόρφωση του διαγράμματος BoxPlot το «κουτί» (Box) έχει ως οριζόντιες πλευρές τις τιμές του πρώτου και τρίτου τεταρτημόριου (εύρος από το Q1 έως το Q3 - interquartile range – IQR), ενώ η οριζόντια γραμμή στο εσωτερικό του εκφράζει την τιμή της διαμέσου. Τα ευθύγραμμα τμήματα, που εκτείνονται μέχρι τις τιμές που δεν θεωρούνται ακραίες (συνήθως μέχρι 1.5 × IQR), ενώ πιθανές ακραίες τιμές (outliers) απεικονίζονται ως ξεχωριστά σημεία. Τα διαγράμματα BoxPlot, που χρησιμοποιήθηκαν στην παρούσα εργασίας αποδείχθηκαν ιδιαίτερα χρήσιμα για τη ευκολότερη σύγκριση κατανομών (κυρίως συσχετίσεων μεταξύ μεταβλητών στην εργασία αυτή).

Επιστρατεύτηκε ένα ακόμη εργαλείο οπτικοποίησης στοιχείων κατανομών, τα διαγράμματα Dumbbell, τα οποία αποτελούν έναν απλούστερο τύπο γραφήματος που χρησιμοποιείται για την οπτική σύγκριση δύο τιμών. Απεικονίζει δύο σημεία ενωμένα με μία οριζόντια γραμμή και είναι εξαιρετικά χρήσιμο σε περιπτώσεις όπου θέλουμε να αναδείξουμε τη διαφορά μεταξύ δύο μεταβλητών, καθώς δίνει έμφαση στη σύγκριση τους χωρίς να προσφέρει πολλές αποπροσανατολιστικές πληροφορίες. Η απλότητά αυτή τον άμεσο εντοπισμό των περιοχών όπου παρατηρείται η μεγαλύτερη απόκλιση μεταξύ των δύο τιμών.

Διαγράμματα BoxPlot και Dumbbell παρατίθενταιστο Κεφάλαιο 7.

5. Εξαγωγή Χαρακτηριστικών

5.1. Εισαγωγή

Στα πλαίσια της παρούσας εργασίας και συγκεκριμένα της διαμόρφωσης του μοντέλου μηχανικής εκμάθησης, με σκοπό τη διεξαγωγή συμπερασμάτων για την έρευνα μας, δόθηκε πολύ μεγάλη προσοχή στις διάφορές παραμέτρους που χρησιμοποιήθηκαν. Η διαδικασία άντλησης των διαφορετικών στοιχείων, που θα περιγράφουν παρακάτω αποτέλεσε νευραλγικό κομμάτι για την έρευνα μας και την αξιοπιστία των αποτελεσμάτων, στηρίχτηκε σε πολλές διαφορετικές βάσεις δεδομένων και επετεύχθη με τη χρήση διαφορετικών προγραμματιστικών εφαρμογών, επιλέγοντας ανά περίπτωση την πλέον αρμόζουσα. Στο κεφάλαιο αυτό θα περιγράφει η διαδικασία εξαγωγής των παρακάτω χαρακτηριστικών – παραμέτρων:

- Πυκνότητα Υδρογραφικού Δικτύου
- Πυκνότητα Οδικού Δικτύου
- Συντεταγμένες Κεντροειδών Κομητειών (Centroid Coordinates)
- Απόσταση Κεντροειδούς από Ακτογραμμή (Distance to Sea)
- Μέσο Υψόμετρο (Elevation)
- Πληθυσμιακή Βάση Κομητείας

5.2 Υπολογισμός Πυκνότητας Υδρογραφικού Δικτύου

Μια από τις μελετώμενες μεταβλητές είναι η Πυκνότητα Υδρογραφικού Δικτύου. Ως Πυκνότητα υδρογραφικού δικτύου ορίζεται ο λόγος του συνολικού μήκους των υδάτινων ροών (π.χ. ποτάμια, λίμνες, ρέματα) σε μία περιοχή, προς την επιφάνεια της. Εκφράζεται σε μονάδες km/km² και αποτελεί θεμελιώδη υδρογραφικό δείκτη για την εκτίμηση της απορροής, της διαπερατότητας εδάφους και του πλημμυρικού κινδύνου. Εκφράζεται ως:

$$D = \frac{L}{A} \tag{5.1}$$

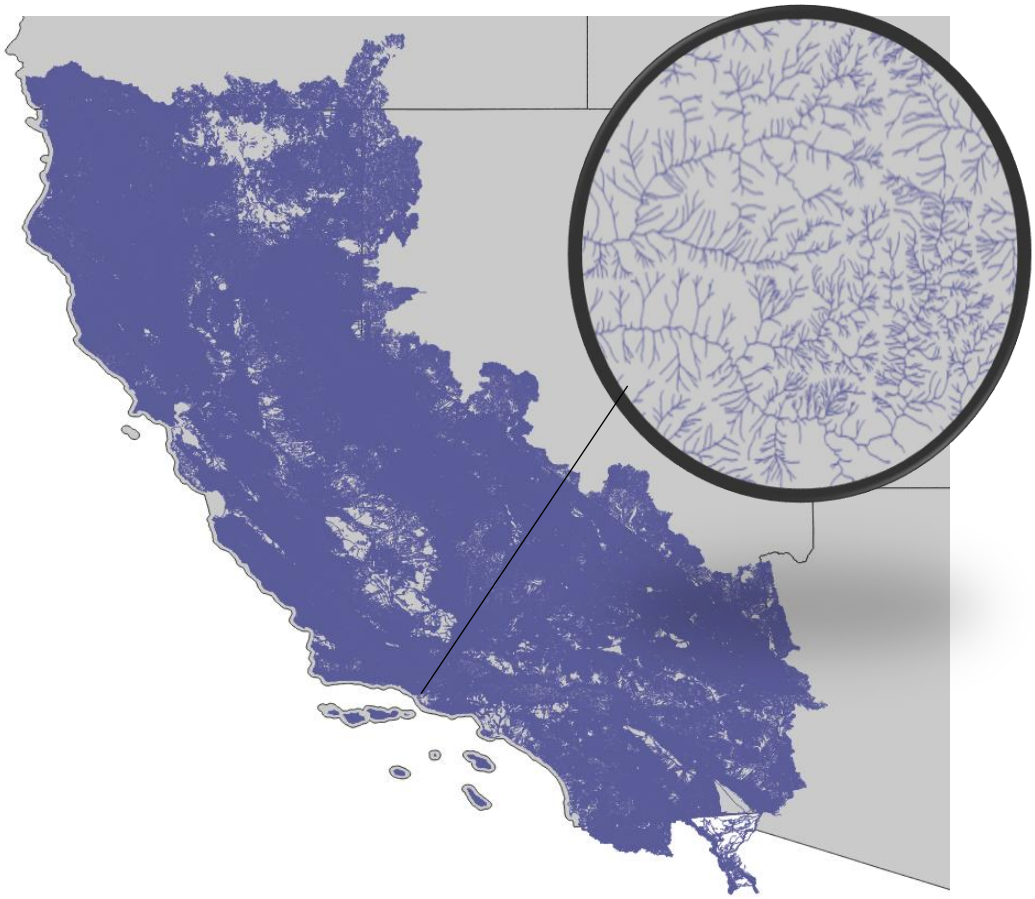
Όπου D: Πυκνότητα υδρογραφικού δικτύου (km/km^2),

L: το συνολικό μήκος των υδάτινων ροών (km)

A: η επιφάνεια της περιοχής (km^2)

Στην προκειμένη περίπτωση η ως περιοχή ορίζεται η επιφάνεια που περικλείεται από τα σύνορα της εκάστοτε κομητείας (county), τα οποία προέρχονται από το αρχείο TIGER/Line® Shapefiles της Υπηρεσίας Απογραφής των Ηνωμένων Πολιτειών (U.S. Census Bureau) (βλ. Κεφάλαιο 3.5.1). Για τον υπολογισμό του συνολικού μήκους υδάτινων ροών χρησιμοποιήθηκε ως πρώτη ύλη η βάση δεδομένων National Hydrography Dataset (NHD) σε υψηλή ανάλυση (High Resolution) (βλ. Κεφάλαιο 3.5.2).

Το συγκεκριμένο γεωχωρικό αρχείο (gpkg) περιέχει πληροφορίες για όλους τους υδρολογικούς σχηματισμούς (κανάλια, ρέματα κλπ), με ακριβή γεωμετρία σε μορφή γραμμών. Το υδρογραφικό δίκτυο αποτυπώνεται στο επίπεδο (Layer) NHDFlowline. Εν συνεχεία ακολούθησε η συγγραφή του σωστού τμήματος κώδικα (script) και η υλοποίηση του μέσω της εφαρμογής QGIS desktop. Στην παρακάτω Εικόνα 5.1 παρουσιάζεται η απεικόνιση του Υδρογραφικού δικτύου της Πολιτείας της Καλιφόρνια, καθώς και ένα μεγεθυμένο απόσπασμα για την καλύτερη εποπτεία του (του υδρογραφικού δικτύου).



ΕΙΚΟΝΑ 5.1- ΥΔΡΟΓΡΑΦΙΚΟ ΔΙΚΤΥΟ ΠΟΛΙΤΕΙΑΣ ΚΑΛΙΦΟΡΝΙΑ ΚΑΙ ΜΕΓΕΘΥΜΕΝΟ ΑΠΟΣΠΑΣΜΑ

Τα αποτελέσματα της παραπάνω διαδικασίας παρουσιάζονται στον Πίνακα 5.1.:

ΠΙΝΑΚΑΣ 5.1. ΠΥΚΝΟΤΗΤΑ ΥΔΡΟΓΡΑΦΙΚΟΥ ΔΙΚΤΥΟΥ ΚΑΛΙΦΟΡΝΙΑ ΣΕ ΕΠΙΠΕΔΟ ΚΟΜΗΤΕΙΑΣ

NAME	county Code	land (km^2)	Length (km)	hydro_density (km/km^2)
Sierra	6091	2468,695	8512999,653	3,448
Sacramento	6067	2500,063	5802009,967	2,321
Santa Barbara	6083	7080,875	24217426,213	3,420
Calaveras	6009	2641,830	10414420,372	3,942
Ventura	6111	4767,585	16896272,685	3,544
Los Angeles	6037	10515,988	27066547,216	2,574
Sonoma	6097	4080,104	9816148,818	2,406
Kings	6031	3602,602	4918642,734	1,365
San Diego	6073	10905,040	36768719,260	3,372
Placer	6061	3644,330	11844958,528	3,250
San Francisco	6075	120,952	126585,209	1,047
Marin	6041	1347,159	8269729,470	6,139
Mariposa	6043	3752,477	13813023,876	3,681
Lassen	6035	11762,091	18505581,275	1,573
Napa	6055	1947,606	5528215,483	2,838
Shasta	6089	9778,891	25714889,402	2,630
Monterey	6053	8499,610	27960598,317	3,290
Trinity	6105	8234,265	24701289,728	3,000
Mendocino	6045	9082,587	27463729,896	3,024
Inyo	6027	26410,783	73330434,640	2,777
Mono	6051	7896,625	22813745,261	2,889
Tuolumne	6109	5752,133	22624176,723	3,933
Solano	6095	2128,287	6043873,212	2,840
San Bernardino	6071	51976,311	143974209,775	2,770
Contra Costa	6013	1856,753	4658991,775	2,509
Alpine	6003	1912,293	8281698,585	4,331
El Dorado	6017	4423,288	15220601,029	3,441
Yolo	6113	2623,628	6320872,451	2,409
Yuba	6115	1636,888	4552795,144	2,781
San Benito	6069	3596,591	13403797,381	3,727
Humboldt	6023	9241,141	27112913,864	2,934
Riverside	6065	18671,880	52646433,746	2,820
Kern	6029	21068,705	51947914,005	2,466
Colusa	6011	2980,245	7949734,313	2,667
Del Norte	6015	2606,050	10343603,419	3,969
Modoc	6049	10225,877	16447783,494	1,608
Fresno	6019	15432,139	42932483,523	2,782
Madera	6039	5534,592	20259791,215	3,661
Santa Clara	6085	3343,893	10243470,823	3,063
Tehama	6103	7638,229	23355834,262	3,058

San Joaquin	6077	3605,879	9716271,194	2,695
Alameda	6001	1910,011	5989681,246	3,136
Nevada	6057	2480,587	9723859,411	3,920
Butte	6007	4238,489	11201519,274	2,643
Merced	6047	5019,407	12611691,160	2,513
Tulare	6107	12493,841	22600785,991	1,809
Stanislaus	6099	3875,011	8055904,564	2,079
Orange	6059	2053,499	5082160,311	2,475
Imperial	6025	10814,568	32064604,333	2,965
Sutter	6101	1560,976	3286776,647	2,106
Amador	6005	1539,967	6198990,679	4,025
Lake	6033	3250,648	13529886,854	4,162
Plumas	6063	6612,401	23713232,602	3,586
San Mateo	6081	1161,883	4087095,232	3,518
Siskiyou	6093	16261,931	29180415,268	1,794
Santa Cruz	6087	1152,818	3923841,230	3,404
Glenn	6021	3403,160	10341806,656	3,039
San Luis Obispo	6079	8549,141	29212045,712	3,417

Στη συνέχεια υλοποιήθηκε ο έλεγχος των τιμών που προέκυψαν με παλαιότερη βιβλιογραφία (U.S. Geological Survey. (2006, 2009), California Resources Agency (1999), Santa Barbara County Public Works Department (2015), North Coast Regional Water Quality Control Board (2012)). Στον παρακάτω πίνακα 5.2. παρατίθεται ο έλεγχος αυτός μέσω συγκρίσεων καθώς και οι πηγές τους. Οι διαφορές που παρατηρούνται είναι πολύ μικρές, συνεπώς το τμήμα του κώδικα (script) αποδεικνύεται αρκετά αξιόπιστο.

ΠΙΝΑΚΑΣ 5.2. ΈΛΕΓΧΟΣ ΑΠΟΤΕΛΕΣΜΑΤΩΝ ΔΙΑΔΙΚΑΣΙΑΣ ΜΕ ΤΙΜΕΣ ΒΙΒΛΙΟΓΡΑΦΙΑΣ ΚΑΙ ΠΑΡΑΘΕΣΗ ΠΗΓΩΝ

County	Calculated Hydro Density (km^2)	Reference Hydro Density (km^2)	Deviation_%	Source
Sierra	3,4484	3,4	1,42	Sierra Nevada headwaters mapping (USGS)
Sacramento	2,3207	2,2	5,49	CalWater planning documents
Santa Barbara	3,4201	3,4	0,59	Santa Barbara County hydrology studies
Los Angeles	2,5738	2,5	2,95	USGS urban stream density estimate
Mendocino	3,0238	2,8	7,99	North Coast watershed reports

5.3. Υπολογισμός Πυκνότητας Οδικού Δικτύου

Στις μελετώμενες μεταβλητές ανήκει και ένα ακόμα χαρακτηριστικό Πυκνότητας. Είναι η Πυκνότητα του Οδικού Δικτύου. Ως Πυκνότητα οδικού δικτύου (ή αλλιώς πυκνότητα κυκλοφορίας) ορίζεται ο λόγος του συνολικού μήκους των οδικών κατασκευών (π.χ. οδοί, λεωφόροι κ.α.) σε μία περιοχή, προς την επιφάνεια της. Εκφράζεται σε μονάδες km/km² και αποτελεί σημαντικό δείκτη ανάπτυξης της ευρύτερης περιοχής. Εκφράζεται ως:

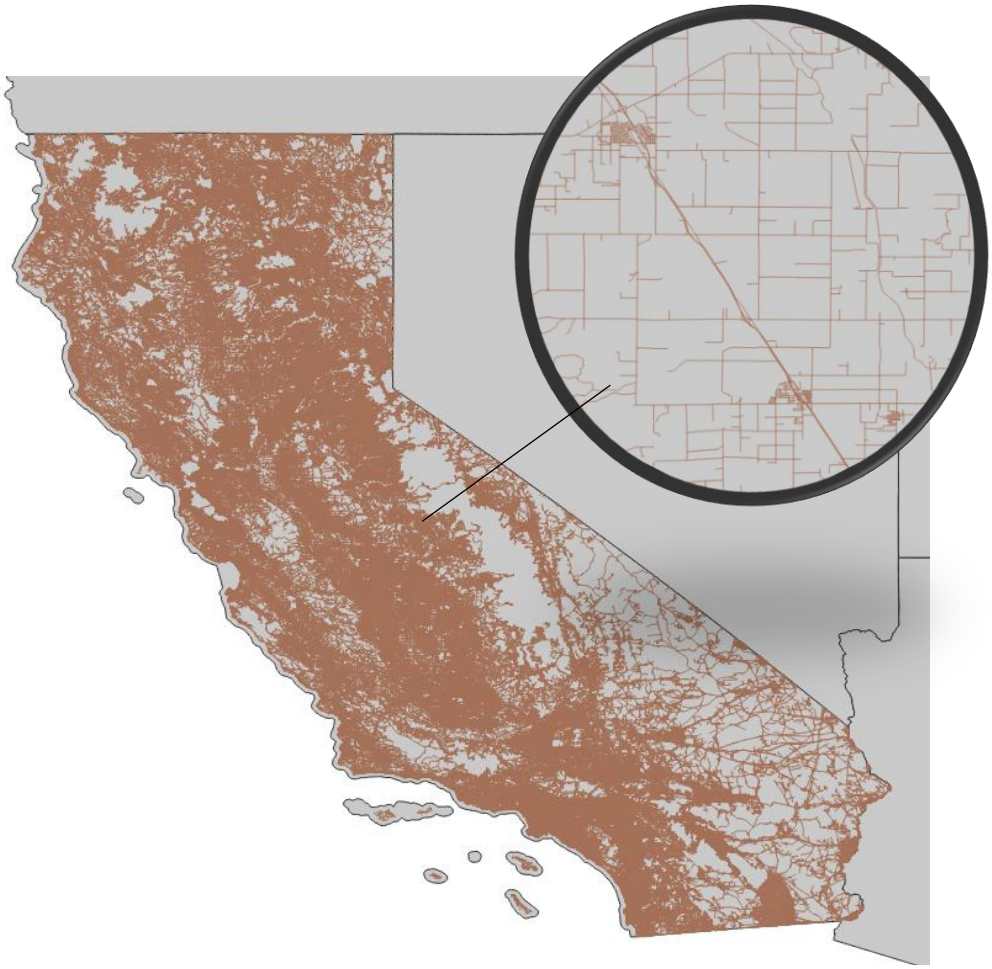
$$D = \frac{L}{A} \tag{5.2}$$

όπου D: Πυκνότητα οδικού δικτύου (km/km^2),

L: το συνολικό μήκος των οδών/λεωφόρων (km)

A: η επιφάνεια της περιοχής (km^2)

Όπως και με το Δίκτυο υδρογραφικού δικτύου η ως περιοχή ορίζεται η επιφάνεια που περικλείεται από τα σύνορα της εκάστοτε κομητείας (county), τα οποία προέρχονται από το αρχείο TIGER/Line® Shapefiles της Υπηρεσίας Απογραφής των Ηνωμένων Πολιτειών (U.S. Census Bureau) (βλ. Κεφάλαιο 3.5.1). Για τον υπολογισμό του συνολικού μήκους οδών κυκλοφορίας, χρησιμοποιήθηκε ως πρώτη ύλη η βάση δεδομένων TRAN_California_State_GPKG (βλ. Κεφάλαιο 3.5.3).
Το συγκεκριμένο γεωχωρικό αρχείο (gpkg) περιέχει πληροφορίες για όλους τους κυκλοφοριακούς σχηματισμούς (οδούς, λεωφόρους κλπ), με ακριβή γεωμετρία σε μορφή γραμμών. Το υδρογραφικό δίκτυο αποτυπώνεται στο επίπεδο (Layer) Trans_RoadSegment. Εν συνεχεία ακολούθησε η συγγραφή του σωστού τμήματος κώδικα (script) και η υλοποίηση του μέσω της εφαρμογής GGIS desktop. Στην παρακάτω Εικόνα 5.2 παρουσιάζεται η απεικόνιση του κυκλοφοριακού δικτύου της Πολιτείας της Καλιφόρνια, καθώς και ένα μεγεθυμένο απόσπασμα για την καλύτερη εποπτεία του (του υδρογραφικού δικτύου):



ΕΙΚΟΝΑ 5.2- ΟΔΙΚΟ ΔΙΚΤΥΟ ΠΟΛΙΤΕΙΑΣ ΚΑΛΙΦΟΡΝΙΑ ΚΑΙ ΜΕΓΕΘΥΜΕΝΟ ΑΠΟΣΠΑΣΜΑ

ΠΙΝΑΚΑΣ 5.3. ΠΥΚΝΟΤΗΤΑ ΟΔΙΚΟΥ ΔΙΚΤΥΟΥ ΚΑΛΙΦΟΡΝΙΑ ΣΕ ΕΠΙΠΕΔΟ ΚΟΜΗΤΕΙΑΣ

NAME	county Code	land (km^2)	length (km)	Transport_density (km/km^2)
Sierra	6091	2468,695	4220820,054	1,710
Sacramento	6067	2500,063	10205749,044	4,082
Santa Barbara	6083	7080,875	10108466,789	1,428
Calaveras	6009	2641,830	4618786,263	1,748
Ventura	6111	4767,585	9807901,299	2,057
Los Angeles	6037	10515,988	51002280,740	4,850
Sonoma	6097	4080,104	9529630,104	2,336
Kings	6031	3602,602	6462039,484	1,794
San Diego	6073	10905,040	27528662,718	2,524
Placer	6061	3644,330	8865745,903	2,433
San Francisco	6075	120,952	1956597,096	16,177
Marin	6041	1347,159	3263504,504	2,423
Mariposa	6043	3752,477	4137384,497	1,103
Lassen	6035	11762,091	12603610,746	1,072
Napa	6055	1947,606	2006203,290	1,030

Shasta	6089	9778,891	14789136,549	1,512
Monterey	6053	8499,610	12919560,441	1,520
Trinity	6105	8234,265	10006454,069	1,215
Mendocino	6045	9082,587	13448617,244	1,481
Inyo	6027	26410,783	10692330,661	0,405
Mono	6051	7896,625	6585367,027	0,834
Tuolumne	6109	5752,133	7160371,568	1,245
Solano	6095	2128,287	6986649,319	3,283
San Bernardino	6071	51976,311	49195329,399	0,946
Contra Costa	6013	1856,753	9457065,233	5,093
Alpine	6003	1912,293	898643,336	0,470
El Dorado	6017	4423,288	10770601,287	2,435
Yolo	6113	2623,628	3309573,293	1,261
Yuba	6115	1636,888	3566754,857	2,179
San Benito	6069	3596,591	4408324,437	1,226
Humboldt	6023	9241,141	12898191,223	1,396
Riverside	6065	18671,880	27931102,156	1,496
Kern	6029	21068,705	41163276,242	1,954
Colusa	6011	2980,245	3124696,129	1,048
Del Norte	6015	2606,050	3129487,424	1,201
Modoc	6049	10225,877	12105790,856	1,184
Fresno	6019	15432,139	23086219,791	1,496
Madera	6039	5534,592	8529716,260	1,541
Santa Clara	6085	3343,893	12058869,730	3,606
Tehama	6103	7638,229	10127335,141	1,326
San Joaquin	6077	3605,879	7895311,730	2,190
Alameda	6001	1910,011	9879138,017	5,172
Nevada	6057	2480,587	6143488,120	2,477
Butte	6007	4238,489	7546019,700	1,780
Merced	6047	5019,407	7890041,344	1,572
Tulare	6107	12493,841	13586909,554	1,087
Stanislaus	6099	3875,011	5575425,058	1,439
Orange	6059	2053,499	16522780,244	8,046
Imperial	6025	10814,568	9323786,983	0,862
Sutter	6101	1560,976	2929224,426	1,877
Amador	6005	1539,967	2600137,600	1,688
Lake	6033	3250,648	6354994,211	1,955
Plumas	6063	6612,401	10953482,491	1,657
San Mateo	6081	1161,883	4579027,829	3,941
Siskiyou	6093	16261,931	22718785,273	1,397
Santa Cruz	6087	1152,818	3753133,732	3,256
Glenn	6021	3403,160	4742264,142	1,393
San Luis Obispo	6079	8549,141	13296883,839	1,555

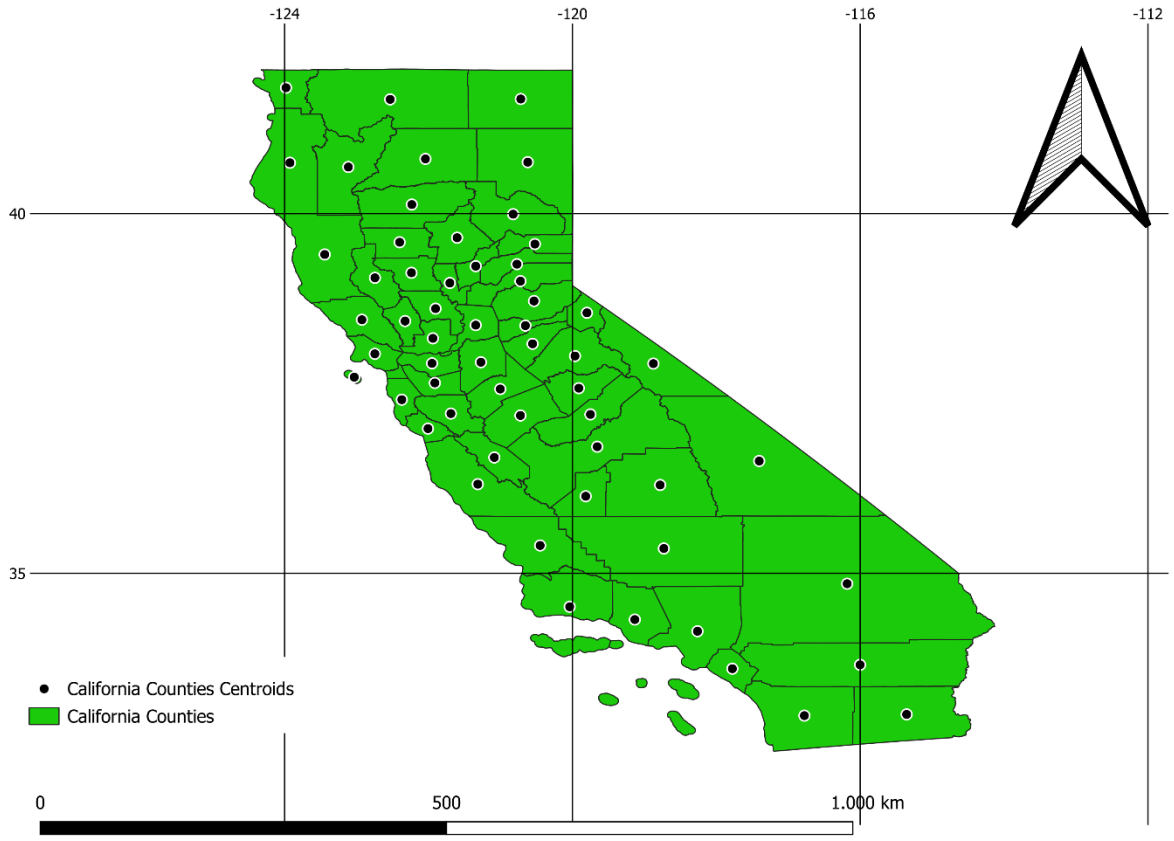
Στη συνέχεια υλοποιήθηκε ο έλεγχος των τιμών που προέκυψαν με παλαιότερη βιβλιογραφία(County of Santa Barbara. (2021), Los Angeles Metro. (2021), San Diego Association of Governments (SANDAG). (n.d.), San Francisco County Transportation Authority (SFCTA). (n.d.), San Bernardino County Transportation Authority (SBCTA). (n.d.), Fresno County Public Works and Planning Department. (n.d.)) . Στον παρακάτω πίνακα 5.4. παρατίθεται ο έλεγχος αυτός μέσω συγκρίσεων καθώς και οι πηγές τους. Οι διαφορές που παρατηρούνται είναι πολύ μικρές, συνεπώς το τμήμα του κώδικα (script) αποδεικνύεται αρκετά αξιόπιστο.

ΠΙΝΑΚΑΣ 5.4. ΈΛΕΓΧΟΣ ΑΠΟΤΕΛΕΣΜΑΤΩΝ ΔΙΑΔΙΚΑΣΙΑΣ ΜΕ ΤΙΜΕΣ ΒΙΒΛΙΟΓΡΑΦΙΑΣ ΚΑΙ ΠΑΡΑΘΕΣΗ ΠΗΓΩΝ

County	Calculated Transport Density (km^2)	Reference Transport Density (km^2)	Deviation_%	Source
Santa Barbara	1.43	1.5	-4.7	Caltrans PRD 2021 (urban fringe)
Ventura	2.06	1.9	8.4	Ventura General Plan 2040
Los Angeles	4.85	4.85	0.0	Caltrans PRD 2021 (urban roads)
San Diego	2.52	2.8	-10.0	Urban modeling estimate
San Francisco	16.18	16.18	0.0	City estimate based on land area and total roads
San Bernardino	0.95	0.9	5.6	Caltrans PRD + sparse county avg
Fresno	1.5	1.4	7.1	Caltrans + Ag-heavy county reference
Alameda	5.17	5.0	3.4	Urban studies + Caltrans PRD
Imperial	0.86	0.8	7.5	Caltrans rural county stats

5.4. Συντεταγμένες Κεντροειδών Κομητειών

Ένα ακόμη χαρακτηριστικό, που χρησιμοποιήθηκε ως μεταβλητή εισόδου είναι οι συντεταγμένες του κεντροειδούς (centroid) κάθε Κομητείας (INTPTLAT, τεταγμένη-γεωγρ. πλάτος και INTPTLON τετμημένη-γεωγρ. μήκος), συμπεριλαμβάνοντας έτσι τη χωρική μεταβλητότητα στη διαδικασία μοντελοποίησης. Οι συντεταγμένες αυτές προσφέρουν γεωγραφική πληροφορία για κάθε κομητεία και επιτρέπουν στο μοντέλο να εντοπίζει χωρικά πρότυπα και γεωγραφικές συσχετίσεις στα προβλεπόμενα σύνολα Αιτημάτων Αποζημίωσης. Αυτό επιτυγχάνεται καθώς το μοντέλο, με αυτή την παράμετρο μπορεί να εκπαιδευτεί και να προσαρμοστεί σε γεω-εξαρτώμενες σχέσεις. Στα πλαίσια της πτυχιακής μας εργασίας, αναλύσαμε την Πολιτεία της Καλιφόρνια, τα κέντρο ειδή των Κομητειών, της οποίας παρουσιάζονται στον παρακάτω Χάρτη (Εικόνα 5.3. και Πίνακα 5.5.):



ΕΙΚΟΝΑ 5.3. ΚΕΝΤΡΟΕΙΔΗ ΚΟΜΗΤΕΙΩΝ ΚΑΛΙΦΟΡΝΙΑ

ΠΙΝΑΚΑΣ 5.5. ΚΕΝΤΡΟΕΙΔΗ ΚΟΜΗΤΕΙΩΝ ΚΑΛΙΦΟΡΝΙΑ

NAME	INTPTLAT	INTPTLON	NAME	INTPTLAT	INTPTLON
Sierra	39,5769252	-120,5219926	Mono	37,9158363	-118,8751668
Sacramento	38,4501363	-121,3443291	Tuolumne	38,0214344	-119,9647335
Santa Barbara	34,5366774	-120,0383645	Solano	38,2672255	-121,939594
Calaveras	38,1910682	-120,5541065	San Bernardino	34,8566615	-116,1815707
Ventura	34,3587477	-119,1331453	Contra Costa	37,919479	-121,9515431
Los Angeles	34,1963983	-118,2618616	Alpine	38,6217831	-119,7983522
Sonoma	38,5251824	-122,9261095	El Dorado	38,7856116	-120,5342245
Kings	36,072478	-119,8155301	Yolo	38,6796091	-121,90275
San Diego	33,0236041	-116,7761174	Yuba	39,27013	-121,3442587
Placer	39,0620323	-120,7227181	San Benito	36,6116507	-121,0858108
San Francisco	37,7272391	-123,0322294	Humboldt	40,7066554	-123,9261757
Marin	38,0513667	-122,7463958	Riverside	33,7298275	-116,0022389
Mariposa	37,5743432	-119,9117215	Kern	35,3466288	-118,7295064
Lassen	40,7152866	-120,6212225	Colusa	39,1777385	-122,2375629
Napa	38,5070999	-122,3259045	Del Norte	41,7499033	-123,9809983
Shasta	40,7605142	-122,0435558	Modoc	41,5929185	-120,7183704
Monterey	36,2401044	-121,3155781	Fresno	36,7610058	-119,6550193
Trinity	40,6478582	-123,114666	Madera	37,2098213	-119,7498023
Mendocino	39,4323876	-123,4428811	Santa Clara	37,2216142	-121,6895401
Inyo	36,5621604	-117,4042092	Tehama	40,1261573	-122,2322737

NAME	INTPTLAT	INTPTLON
San Joaquin	37,9349815	-121,272244
Alameda	37,6471385	-121,912488
Nevada	39,2975082	-120,7713429
Butte	39,6653362	-121,6032086
Merced	37,1948063	-120,7228019
Tulare	36,2288339	-118,7810551
Stanislaus	37,5623162	-121,0028311
Orange	33,6756872	-117,7772068
Imperial	33,0408143	-115,3554001
Sutter	39,0361898	-121,7039397
Amador	38,4435493	-120,653858
Lake	39,1080094	-122,7452171
Plumas	39,9922953	-120,8243709
San Mateo	37,4146725	-122,3715457
Siskiyou	41,5879861	-122,5332868
Santa Cruz	37,0123466	-122,0077889
Glenn	39,6025462	-122,4016998
San Luis Obispo	35,3881891	-120,4488974

5.5. Απόσταση Κεντροειδούς από Ακτογραμμή

Ένα ακόμη χαρακτηριστικό, που χρησιμοποιήθηκε ως μεταβλητή εισόδου είναι οι συντεταγμένες του η απόσταση Κεντροειδούς Κομητείας από τη θάλασσα (distance to sea). Η παράμετρος αυτή κρίνεται πολύ σημαντική καθώς είναι συνυφασμένη με την έκθεση σε παράκτιους κινδύνους, (πχ πλημμύρες από καταιγίδες ή άνοδο της στάθμης της θάλασσας). Τα φαινόμενα ENSO, όπως αναφέρθηκε και σε προηγούμενες ενότητες επηρεάζουν περιοχές που βρίσκονται κοντά στη θάλασσα. Η απόσταση λοιπόν του κεντροειδούς από την ακτογραμμή μας δίνει πολύτιμες πληροφορίες για το βαθμό επιρροής φαινομένων ENSO και τις προβλέψεις των Αιτημάτων Αποζημίωσης.

Για τον υπολογισμό των αποστάσεων αυτών χρησιμοποιήθηκε η εφαρμογή QGIS, μέσω της κωδικοποίησης κατάλληλου script και τα αποτελέσματα της διαδικασίας προβάλλονται στον πίνακα 5.6., που ακολουθεί:

ΠΙΝΑΚΑΣ 5.6. ΑΠΟΣΤΑΣΕΙΣ ΚΕΝΤΡΟΕΙΔΩΝ ΚΟΜΗΤΕΙΩΝ ΤΗΣ ΚΑΛΙΦΟΡΝΙΑ ΑΠΟ ΤΗ ΘΑΛΑΣΣΑ

NAME	distance_to Sea_km	NAME	distance_to Sea_km	NAME	distance_to Sea_km
Sierra	190,176	San Benito	62,890	Inyo	293,015
Sacramento	43,851	Humboldt	19,695	Mono	224,910
Santa Barbara	8,348	Riverside	142,211	Tuolumne	130,923
Calaveras	81,282	Kern	129,086	Solano	15,326
Ventura	17,386	Colusa	114,555	San Bernardino	207,333
Los Angeles	30,104	Del Norte	16,363	Contra Costa	13,486
Sonoma	20,072	Modoc	279,089	Alpine	156,637
Kings	120,576	Fresno	184,531	El Dorado	117,219
San Diego	50,622	Madera	171,858	Yolo	61,593
Placer	131,700	Santa Clara	34,754	Yuba	127,824
San Francisco	14,705	Tehama	142,124	Imperial	122,640
Marin	6,997	San Joaquin	16,736	Sutter	98,466
Mariposa	143,229	Alameda	20,990	Amador	85,144
Lassen	295,727	Nevada	150,811	Lake	77,887
Napa	39,686	Butte	168,861	Plumas	218,761
Shasta	185,581	Merced	110,004	San Mateo	7,382
Monterey	31,714	Tulare	211,968	Siskiyou	127,587
Trinity	84,662	Stanislaus	62,526	Santa Cruz	6,564
Mendocino	31,599	Orange	13,246	Glenn	141,363
San Luis Obispo	32,632				

5.6. Μέσο Υψόμετρο

Μια ακόμα μεταβλητή εισαγωγής στο μοντέλο μας απετέλεσε το Μέσο Υψόμετρο Κομητείας. Επιλέχθηκε το παραπάνω στοιχείο , καθώς έχουν παρατηρηθεί λιγότερα πλημμυρικά φαινόμενα (και κατ' επέκταση Αιτήματα για Αποζημίωση) σε περιοχές με μεγαλύτερο υψόμετρο, οπότε μέσω της ανάλυσης του μας δόθηκε η ευκαιρία βελτιστοποίησης του μοντέλου μας. Για τον υπολογισμό του Μέσου Υψομέτρου ανά κομητεία στην Καλιφόρνια, δημιουργήθηκαν script για την εφαρμογή QGIS. Αρχικά πραγματοποιήθηκε συγχώνευση 72 αρχείων υψομετρικών δεδομένων GeoTIFF ("USGS 3D Elevation Program (3DEP) – 1/3 arc-second historical elevation tiles"), τα οποία περιλαμβάνουν την έκταση της πολιτείας στο σύνολο της. Αφού δημιουργήθηκε το ενιαίο ψηφιδωτό (merged DEM), εφαρμόστηκε ζωνική στατιστική (zonal statistics) με βάση τα σύνορα των κομητειών, εξάγοντας έτσι τη μέση τιμή υψομέτρου (mean elevation) για κάθε μία. Το αποτελέσματα παρουσιάζονται στον παρακάτω πίνακα 5.7.

ΠΙΝΑΚΑΣ 5.7 – ΜΕΣΟ ΥΨΟΜΕΤΡΟ ΚΟΜΗΤΕΙΩΝ ΤΗΣ ΚΑΛΙΦΟΡΝΙΑ

Name	Mean Elevation (m)	Name	Mean Elevation (m)	Name	Mean Elevation (m)
Imperial	96,222	Inyo	240,003	Santa Clara	187,640
Sacramento	29,642	Mono	255,000	Tehama	214,400
Santa Barbara	154,534	Tuolumne	253,530	San Joaquin	36,398
Calaveras	238,441	Solano	41,008	Alameda	142,195
Ventura	185,941	San Bernardino	251,223	Nevada	253,126
Los Angeles	187,530	Contra Costa	111,386	Butte	161,994
Sonoma	141,522	Alpine	255,000	Merced	89,758
Kings	97,580	El Dorado	251,795	Tulare	206,006
San Diego	202,256	Mariposa	247,749	Stanislaus	100,997
Placer	213,443	Yuba	139,080	Orange	110,489
San Francisco	11,113	San Benito	235,651	Sierra	247,000
Marin	75,215	Humboldt	195,854	Sutter	23,388
Yolo	75,791	Riverside	222,165	Amador	220,585
Lassen	255,000	Kern	217,025	Lake	254,924
Napa	191,674	Colusa	121,917	Plumas	255,000
Shasta	247,354	Del Norte	184,385	San Mateo	87,175
Monterey	187,256	Modoc	255,000	Siskiyou	254,947
Trinity	254,734	Fresno	184,327	Santa Cruz	132,134
Mendocino	211,089	Madera	181,877	Glenn	156,087
San Luis Obispo	212,283				

5.7. Πληθυσμιακή Βάση Κομητείας

Η τελευταία μεταβλητή, που χρησιμοποιήθηκε για την εκπαίδευση του μοντέλου είναι ο πληθυσμός της κάθε Κομητείας της Καλιφόρνια. Μάλιστα απετέλεσε μία καθοριστική προσθήκη στη διαδικασία εκπαίδευσης του μοντέλου μας, καθώς μας επέτρεψε να διακρίνουμε αν το πλήθος αιτημάτων αποζημίωσης σε μία κομητεία είναι υψηλό ή χαμηλό εξαιτίας του βαθμού επιρροής, που έχουν τα φαινόμενα ENSO σε αυτή ή απλώς επειδή έχει περισσότερους / λιγότερους κατοίκους. Αναλυτικότερα, υπολογίστηκε ο αριθμός των αιτημάτων ανά 100.000 κατοίκους για κάθε κομητεία. Αυτή η κανονικοποίηση είναι απαραίτητη μας επέτρεψε να έχουμε μια καθαρότερη εικόνα και έτσι το μοντέλο κατάφερε να εντοπίσει αληθινά υφιστάμενα μοτίβα τρωτότητας, ανεξαρτήτως πληθυσμού. Στον πίνακα 5.8 παρουσιάζεται ο πληθυσμός της κάθε Κομητείας της Καλιφόρνια, όπως αυτός προέκυψε με βάση το U.S. Census Bureau – Population Estimates Program (PEP):

ΠΙΝΑΚΑΣ 5.8 – ΠΛΗΘΥΣΜΟΣ ΚΟΜΗΤΕΙΩΝ ΤΗΣ ΚΑΛΙΦΟΡΝΙΑ

Name	Population (2023)	Name	Population (2023)	Name	Population (2023)
Alameda	1622188	Madera	162858	San Luis Obispo	281639
Alpine	1141	Marin	254407	San Mateo	726353
Amador	41811	Mariposa	16919	Santa Barbara	441257
Butte	207172	Mendocino	89108	Santa Clara	1877592
Calaveras	46565	Merced	291920	Santa Cruz	261547
California	38965193	Modoc	8500	Shasta	180366
Colusa	22037	Mono	13066	Sierra	3200
Contra Costa	1155025	Monterey	430723	Siskiyou	42905
Del Norte	26589	Napa	133216	Solano	449218
El Dorado	192215	Nevada	102037	Sonoma	481812
Fresno	1017162	Orange	3135755	Stanislaus	551430
Glenn	28129	Placer	423561	Sutter	97948
Humboldt	133985	Plumas	19131	Tehama	64896
Imperial	179057	Riverside	2492442	Trinity	15670
Inyo	18527	Sacramento	1584288	Tulare	479468
Kern	913820	San Benito	68175	Tuolumne	54204
Kings	152682	San Bernardino	2195611	Ventura	829590
Lake	67878	San Diego	3269973	Yolo	220544
Lassen	28861	San Francisco	808988	Yuba	85722
Los Angeles	9663345	San Joaquin	800965		

5.8. Ενοποίηση Δεδομένων και Ορισμός ως Δυναμικά & Στατικά Χαρακτηριστικά

Τελικό στάδιο της διαδικασίας , αμέσως μετά την εξαγωγή χαρακτηριστικών για τις Κομητείες της Καλιφόρνια είναι η ενοποίηση των χαρακτηριστικών αυτών σε μία ενιαία βάση δεδομένων. Το στάδιο κρίνεται μείζονας σημασίας, καθώς πάνω σε αυτή τη βάση δεδομένων, που δημιουργήσαμε βασίστηκε ή διαμόρφωση και η εκπαίδευση του μοντέλου μηχανικής εκμάθησης μας. Αναλυτικότερα, έχοντας μια πλήρη λίστα με αξιόπιστα δεδομένα μπορούμε ευκολότερα να διαχειριστούμε το μοντέλο, να προβούμε στην σωστή παραμετροποίηση του και κατ' επέκταση να διεξάγουμε σωστά συμπεράσματα.

Έχοντας λοιπόν δημιουργήσει μια συνολική λίστα με όλες τις μεταβλητές, που περιεγράφηκαν σε αυτό το κεφάλαιο, την ενοποιήσαμε με τα Αιτήματα αποζημιώσεων (βλ. Κεφάλαιο 3.4) για κάθε χρόνο για όλες τις Κομητείες της Καλιφόρνια αλλά και το δείκτη ENSO (βλ. Κεφάλαιο 3.3) . Έτσι καταλήξαμε να έχουμε

μια βάση δεδομένων 1572 καταχωρήσεων, με 8 διαφορικές μεταβλητές (Πυκνότητα Υδρογραφικού Δικτύου, Πυκνότητα Οδικού Δικτύου, Συντεταγμένες Κεντροειδών Counties, Απόσταση Κεντροειδούς από Ακτογραμμή (Distance to Sea), Μέσο Υψόμετρο (Elevation), Πληθυσμιακή Βάση Κομητείας, δείκτης ENSO). Πρόκειται για επαρκή πληροφορία, οργανωμένη με κατάλληλο τρόπο και δομή, που μας επιτρέπει να τροποποιήσουμε το διαμορφωθέν μοντέλο μηχανικής εκμάθησης με διαφορετικά σενάρια (διαφορετικές χρονικές περίοδοι-στόχοι). Αξίζει να σημειωθεί ότι για ορθότερη διαχείριση και επεξεργασίας των πληροφοριών κάναμε την παραδοχή της διάκρισης των μεταβλητών- χαρακτηριστικών σε Στατικά & Δυναμικά Χαρακτηριστικά, όπως παρουσιάζονται στον Πίνακα 5.9. Ως στατικό χαρακτηριστικό ορίζονται οι μεταβλητές, που δε μεταβάλλονται με την πάροδο του χρόνου, ενώ ως δυναμικές, αυτές που μεταβάλλονται.

ΠΙΝΑΚΑΣ 5.9 ΔΙΑΚΡΙΣΗ ΣΕ ΣΤΑΤΙΚΑ & ΔΥΝΑΜΙΚΑ ΧΑΡΑΚΤΗΡΙΣΤΙΚΑ

Στατικά Χαρακτηριστικά	Δυναμικά Χαρακτηριστικά
Πυκνότητα Υδρογραφικού Δικτύου	Δείκτης ENSO
Πυκνότητα Κυκλοφοριακού Δικτύου	Αριθμός Αιτημάτων Αποζημίωσης
Συντεταγμένες Κεντροειδούς Κομητειών	
Απόσταση Κεντρειδών από Ακτογραμμή	
Μέσο Υψόμετρο	
Πληθυσμιακή Βάση (2023)	

6. Μοντέλο Μηχανικής Εκμάθησης (Machine Learning)

6.1 Περιγραφή Προβλήματος

Σε αυτή την ενότητα εξηγείται αναλυτικά η ανάπτυξη και η διαμόρφωση του μοντέλου μηχανικής εκμάθησης. Στις παραπάνω ενότητες έχουν αναφερθεί πολλά για τις φυσικές καταστροφές, και ιδιαίτερα τις πλημμύρες, που αποτελούν ένα από τα πλέον καταστροφικά φαινόμενα παγκοσμίως. Η ερευνητική μας μελέτη εστιάζει στις Ηνωμένες Πολιτείες Αμερικής. Αναλυτικότερα, στις Η.Π.Α οι φυσικές καταστροφές λόγω πλημμυρών αποτελούν αντικείμενο διαχείρισης του οργανισμού FEMA και συγκεκριμένα μέσω του Εθνικού Προγράμματος Ασφάλισης Πλημμυρών (NFIP) (βλ. Κεφάλαιο 3.4.). Κομμάτι του NFIP αποτελεί η καταγραφή των ετήσιων αιτημάτων αποζημιώσεων για ζημιές που σχετίζονται με πλημμύρες. Παρά τη συστηματική συλλογή των δεδομένων αυτών, η πρόβλεψη τόσο των περιοχών, που θα εμφανίσουν υψηλή ασφαλιστική ιδιότητα στο μέλλον, όσο και ο βαθμός επικινδυνότητας παραμένει δύσκολη και με χαμηλή αξιοπιστία. Αυτό το κενό καλείται να γεμίσει το μοντέλο μηχανικής εκμάθησης που διαμορφώθηκε, με στόχο την πρόβλεψη του αριθμού Αιτημάτων Αποζημίωσης σε μια Πολιτεία / Κομητεία, στηριζόμενο σε μεταβλητές, που θα αναφερθούν στις επόμενες ενότητες, από βάσεις δεδομένων, που αναφέρθηκαν σε προηγούμενες.

Στα πλαίσια της παρούσας πτυχιακής μελέτης, το διαμορφωθέν μοντέλο έχει ως πεδίο εφαρμογής τα Όρια της Πολιτείας της Καλιφόρνια, η οποία επιλέχθηκε με βάση τα πολυάριθμα Αιτήματα Αποζημιώσεων, που έλαβαν χώρα αλλά και το γεγονός ότι αποτελεί μια από τις περισσότερο επηρεαζόμενες περιοχές από τα φαινόμενα ENSO (βλ. Κεφάλαιο 7.3). Η φιλοσοφία, φυσικά, η οποία διέπει την ανάπτυξη του μοντέλου καθώς και οι βασικές αρχές, μπορούν με μικρές μετατροπές να προσαρμοστούν σε κάθε Πολιτεία και φυσικά σε όλες τις Η.Π.Α. στο σύνολο τους.

6.2 Επιλογή Μεταβλητών

Έχοντας λοιπόν περιγράψει το πρόβλημα που καλείται να λύσει το μοντέλο Μηχανικής Εκμάθησης της παρούσας εργασίας, συνέχεια έχει η περιγραφή του τρόπου με τον οποίο αυτό επιτυγχάνεται. Εναρκτήριο στάδιο απετέλεσε η επιλογή των χαρακτηριστικών, τα οποία θα αποτελούσαν δεδομένα εισαγωγής (input) για το μοντέλο μας. Δεδομένου λοιπόν ότι στόχος του (μοντέλου) είναι η διερεύνηση της σχέσης μεταξύ περιβαλλοντικών, κλιματικών και γεωχωρικών παραμέτρων με τον αριθμό Αιτημάτων Αποζημίωσης για πλημμυρικές καταστροφές στην Πολιτεία της Καλιφόρνια και ως αποτέλεσμα τις εκτενούς βιβλιογραφικής έρευνας καταλήξαμε στην επιλογή των παρακάτω χαρακτηριστικών για κάθε Κομητεία της Καλιφόρνια, τα οποία και έχουν αναλυθεί στο Κεφάλαιο 5:

- Πυκνότητα υδρογραφικού δικτύου, ως ένδειξη φυσικών υδάτινων ροών,
- Πυκνότητα οδικού δικτύου, ως δείκτη αστικοποίησης,
- Απόσταση από τη θάλασσα, ως δείκτη έκθεσης κομητείας σε παράκτια φαινόμενα,
- Γεωγραφική θέση (συντεταγμένες centroid), ώστε να ληφθεί υπόψη η ύπαρξη χωρικών μοτίβων
- Υψόμετρο, ως πιθανός δείκτης πλημμυρικού κινδύνου,
- Δείκτες ENSO, ως πιθανός δείκτης κινδύνου,
- Πληθυσμός, ώστε να ποσοτικοποιηθεί ο βαθμός έκθεσης στον κίνδυνο,
- Καταγεγραμμένα Αιτήματα Αποζημιώσεων στην Καλιφόρνια προηγούμενων ετών, προκειμένου να ληφθεί υπόψη η ιστορικότητα των φαινομένων

Εύκολα παρατηρεί κανείς ότι οι παραπάνω μεταβλητές χαρακτηρίζονται από υψηλή ετερογένεια μεταξύ τους και πολυπλοκότητα. Το διαμορφωθέν μοντέλο μηχανικής εκμάθησης λοιπόν καλείται να εντοπίσει μοτίβα και κρυφές συσχετίσεις μεταξύ των μεταβλητών, που να μπορούν να εξηγήσουν τη γεωγραφική διασπορά των αιτημάτων αποζημίωσης και πετύχει το μεγαλύτερο δυνατό βαθμό αξιοπιστίας προβλέψεων.

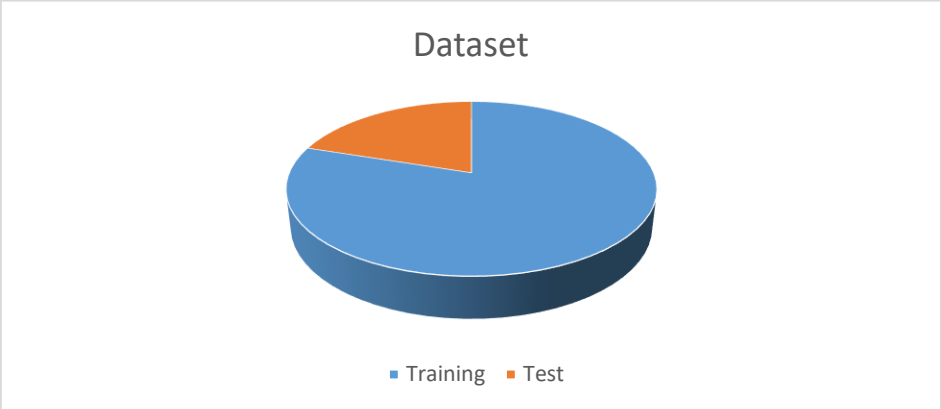
6.3 Εκπαίδευση Μοντέλου & Διαχωρισμός Δεδομένων

Έχοντας καθορίσει τις μεταβλητές εισαγωγής, το επόμενο βήμα ήταν η εκπαίδευση του μοντέλου. Στα πλαίσια αυτής, υιοθετήθηκε η προσέγγιση μηχανικής μάθησης παλινδρόμησης (regression), με στόχο την εκτίμηση του αριθμού ετήσιων απαιτήσεων αποζημιώσεων λόγω πλημμύρας στις Κομητείες της Καλιφόρνια.

Η τεχνική της παλινδρόμησης αποτελεί μία από τις βασικότερες τεχνικές στον τομέα της μηχανικής μάθησης (supervised learning), η οποία χρησιμοποιείται όταν ο στόχος είναι η πρόβλεψη μιας συνεχούς αριθμητικής μεταβλητής, όπως και στην περίπτωση μας. Το μοντέλο εκπαιδεύεται ώστε να αναγνωρίζει τη σχέση μεταξύ των μεταβλητών εισαγωγής και της μεταβλητής στόχου από το σύνολο των δεδομένων δεδομένα, τα οποία χωρίζονται σε σύνολο δεδομένων εκπαίδευσης και δοκιμής. Στα πλαίσια της παρούσας πτυχιακής εργασίας τα δεδομένα διακρίθηκαν σε δύο σύνολα με την εξής αναλογία, που αποτελεί μια τυπική επιλογή:

- Σύνολο εκπαίδευσης (training set): 80% των δεδομένων (1978-2006,2012-2024)
- Σύνολο δοκιμής (test set): 20% των δεδομένων (2007-2011)

Το εκπαιδευτικό σύνολο χρησιμοποιήθηκε για την εκπαίδευση του μοντέλο στις υποκείμενες σχέσεις μεταξύ των ανεξάρτητων μεταβλητών (features) και της προβλεπόμενης μεταβλητής. Αντίθετα αξιοποιήθηκε αποκλειστικά στο στάδιο της αξιολόγησης της απόδοσης του μοντέλου σε άγνωστα δεδομένα — στοιχείο κρίσιμο για τον έλεγχο της ικανότητας γενίκευσης του μοντέλου σε νέα, μελλοντικά παραδείγματα και θα αναλυθεί στην επόμενη ενότητα. Η κατανομή των στα δύο σύνολά έγινε με γνώμονα την παρουσίαση των αποτελεσμάτων με την μεγαλύτερη ακρίβεια για εποπτικούς λόγους , στα πλαίσια της εργασίας.



ΕΙΚΟΝΑ 6.1 – ΔΙΑΚΡΙΣΗ ΔΕΔΟΜΕΝΩΝ ΣΕ ΕΚΠΑΙΔΕΥΤΙΚΟ ΚΑΙ ΔΟΚΙΜΑΣΤΙΚΟ ΣΥΝΟΛΟ

6.4 Αξιολόγηση Απόδοσης (R^2, RMSE, MAE)

Η απόδοση ενός μοντέλου αποτελεί τον πλέον σημαντικό δείκτη αξιοπιστίας του καθώς εκφράζει την πιθανότητα πρόβλεψης σωστών (ή σε ποιο βαθμό προσεγγίζουν τις σωστές) τιμών. Η αξιολόγηση της λοιπόν είναι μια σημαντική διαδικασία, η οποία θα εξηγηθεί στο παρόν κεφάλαιο και πραγματοποιείται με τη χρήση τριών βασικών στατιστικών δεικτών:

- R^2, δηλαδή συντελεστή προσδιορισμού
- Ριζική Μέση Τυπική Απόκλιση (RMSE - Root Mean Squared Error)
- Μέσο Απόλυτο Σφάλμα (MAE- Mean Absolute Error)

Ο δείκτης R^2 (συντελεστής προσδιορισμού) υπολογίζεται το ποσοστό διακύμανσης της προβλεπόμενης μεταβλητής του μοντέλου και εκφράζει το βαθμό προσαρμογής του μοντέλου στις μεταβλητές του δείγματος. Οι τιμές που λαμβάνει κυμαίνονται μεταξύ έως 1, όπου το 1 υποδηλώνει υψηλή ακρίβεια πρόβλεψης του μοντέλου, ενώ τιμές κοντά στο 0 ή αρνητικές δείχνουν ότι το μοντέλο δε λειτουργεί. Όσο η τιμή του λοιπόν αυξάνεται και προσεγγίζει το 1, τόσο μεγαλύτερη αποδοτικότητα χαρακτηρίζει το μοντέλο, ενώ αντίθετα, όσο μειώνεται η τιμή του, τόσο μειώνεται και η απόδοση του μοντέλου. (James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). Στην παρούσα πτυχιακή μελέτη, χρησιμοποιήθηκε ως το βασικό κριτήριο αποδοτικότητας μοντέλου κατά τη διάρκεια της ανάπτυξης του (μοντέλου), αφού σε περιπτώσεις μοντέλων παλινδρόμησης (regression models) κρίνεται το πλέον καθοριστικό κριτήριο. Το R^2 υπολογίστηκε ως εξής:

$$R^2 = 1 - \frac{\Sigma(y_i - y_{προβ})^2}{\Sigma(y_i - \check{y})^2} \quad (6.1)$$

Όπου:

$Y_ι$ = πραγματική τιμή

$Y_{προβ}$= προβλεπόμενη τιμή

$\check{y}$ = μέσος όρος όλων των πραγματικών τιμών

Ένας άλλος δείκτης, ο οποίος ελέγχθηκε , είναι η Ριζική Μέση Τυπική Απόκλιση (Root Mean Squared Error – RMSE) εκφράζει το τετραγωνικό μέσο σφάλμα και είναι ευαίσθητο σε μεγάλες αποκλίσεις, καθώς υπολογίζεται από το τετράγωνο της διαφοράς μεταξύ πραγματικής και προβλεπόμενης τιμής. Εξαιτίας της χρήσης του τετραγώνου στον τύπο είναι εμφανές ότι μία τιμή με μεγάλη απόκλιση θα εκτινάξει το δείκτη, οπότε χαρακτηρίζεται από μία ευαισθησία σε αυτές. Σημειώνεται ότι αντίθετα με το συντελεστή προσδιορισμού, υψηλός δείκτης RMSE εκφράζει κακή απόδοση. Παρατίθεται ο τύπος υπολογισμού:

$$RMSE = \sqrt{\frac{1}{n} \Sigma (y_i - y_{\pi\rho o\beta})^2} \quad (6.2)$$

Όπου:

$Y_ι$ = πραγματική τιμή

$Y_{προβ}$ = προβλεπόμενη τιμή

n = Πλήθος εξεταζόμενου δείγματος

Αντίστοιχα, ο δείκτης του Μέσου Απόλυτου Σφάλματος Μέσο Απόλυτο Σφάλμα (MAE - Mean Absolute Error) χρησιμοποιείται σε αυτές τις περιπτώσεις. Δεν είναι τόσο ευαίσθητος όσο ο δείκτης RMSE στις μεγάλες αποκλίσεις (μιας και δεν χρησιμοποιείται το τετράγωνο της διαφοράς , αλλά η ίδια η διαφορά ανάμεσα στην πραγματική και την προβλεπόμενη τιμή), αλλά ελέγχει το σύνολο των σφαλμάτων. Σημειώνεται ότι όπως και με το δείκτη RMSE, υψηλός δείκτης MAE εκφράζει κακή απόδοση. Παρατίθεται ο τύπος υπολογισμού:

$$MAE = \frac{1}{n} \Sigma \left| y_i - y_{\pi\rho o\beta} \right| \quad (6.3)$$

Όπου:

$Y_ι$ = πραγματική τιμή

$Y_{προβ}$ = προβλεπόμενη τιμή

n = Πλήθος εξεταζόμενου δείγματος

Επισημαίνεται ότι στην παρούσα πτυχιακή εργασία δόθηκε μεγαλύτερη προσοχή στον δείκτη συντελεστή προσδιορισμού (R^2) καθώς δοκιμάστηκαν διαφορετικά μοντέλα προγραμματισμού και σκοπός ήταν να επιλέξουμε αυτό που ταιριάζει καλύτερα στη δική μας περίπτωση. Στον παρακάτω πίνακα 6.1 παρουσιάζονται οι τιμές, που εμφάνισε το διαμορφωθέν μοντέλο στους παραπάνω δείκτες:

Το μοντέλο πέτυχε $R^2 = 0.638$, γεγονός που σημαίνει ότι είναι ικανό να εξηγήσει πάνω από το 60% της διασποράς της εξαρτημένης μεταβλητής. Η τιμή αυτή κρίνεται ικανοποιητική για το συγκεκριμένο πρόβλημα, που καλούμαστε να αντιμετωπίσουμε, δεδομένου του υψηλού βαθμού πολυπλοκότητας και ετερογένειας των μεταβλητών εισαγωγής.

Ταυτόχρονα, οι υπόλοιποι δείκτες κατέγραψαν αποτελέσματα της τάξης των 0.81 (RMSE) και 0.59 (MAE). Στα περισσότερα μοντέλα, οι τιμές αυτές είναι μεγαλύτερες του 1 και περιγράφουν τιμές αντίστοιχες των μεγεθών της προβλεπόμενης μεταβλητής. Στην περίπτωση μας όμως, η προβλεπόμενη μεταβλητή-στόχος (Αιτήματα Αποζημιώσεων) έχει υποστεί κανονικοποίηση ως "αριθμός αιτημάτων αποζημιώσεων ανά 100.000 κατοίκους" (βλ. κεφάλαιο 6.5) . Ως εκ τούτου, οι προβλέψεις του μοντέλου χαρακτηρίζονται από μέσο σφάλμα μικρότερο της μίας μονάδας, δηλαδή λιγότερο από 1 Αίτημα ανά 100.000 κατοίκους, στατιστικό που υποδηλώνει καλή απόδοση.

Κατά γενική ομολογία και συνυπολογίζοντας το σύνολο και των τριών δεικτών αξιολόγησης επιβεβαιώνεται ότι το μοντέλο χαρακτηρίζεται από ισορροπημένη και αξιόπιστη συμπεριφορά.

6.5 Διερεύνηση μοντέλων μηχανικής μάθησης & Σταδιακή βελτιστοποίηση απόδοσης

Σε αυτή την ενότητα, έχοντας εξηγήσει τα βασικά για το μοντέλο μηχανικής εκμάθησης αλλά και τον τρόπο αξιολόγησης του (δείκτες απόδοσης) περιγράφεται η εξέλιξη του και η διαδικασία διαμόρφωσης του μέσα από διαφορετικά στάδια και συνεχείς αλλαγές. Στα πλαίσια της διαδικασίας βελτιστοποίησης του δοκιμάστηκαν αλλαγές τόσο ως προς τις μεταβλητές εισόδου όσο και ως τη σύνταξη των χρησιμοποιούμενων αλγορίθμων με σκοπό πάντα την αύξηση της απόδοσης (όπως αυτή εκφράζεται μέσα από το συντελεστή προσδιορισμού R^2.

Η αρχική έκδοση του μοντέλου βασίστηκε στον τύπο μοντελοποίησης Random Forest (RF). Η επιλογή αυτή έγινε αφού είναι ο συνηθέστερος και επικρατέστερος τύπος μοντέλου, και δεν απαιτεί πολλές ρυθμίσεις παραμέτρων. Στόχος ήταν απλή πρόβλεψη του απόλυτου αριθμού ασφαλιστικών αιτημάτων αποζημιώσεων ανά Κομητεία της Καλιφόρνια με τη χρήση περιορισμένων μεταβλητών εισαγωγής (Αιτήματα Αποζημιώσεων προηγούμενων ετών, πυκνότητα υδρογραφικού δικτύου, πυκνότητα οδικού δικτύου) χωρίς κανονικοποίηση (δε συνηθίζεται στο RF). Η απόδοση του μοντέλου ήταν εξαιρετικά χαμηλή ($R^2 = 0.090$), γεγονός που κατέστησε σαφή την ανάγκη για αύξησης της.

Η πρώτη λογική κίνηση λοιπόν ήταν η αύξηση της πληροφορίας και συγκεκριμένα των δεδομένων εισόδου. Ύστερα από μελέτη της βιβλιογραφίας επιλέχθηκαν κάποιες συμπληρωματικές μεταβλητές εισόδου, οι οποίες ήταν γνωστό ότι επηρεάζουν τα Αιτήματα Αποζημιώσεων. Προστέθηκαν λοιπόν σημαντικά δυναμικά και στατικά χαρακτηριστικά, όπως δείκτες ENSO, ο πληθυσμός, το μέσο υψόμετρο αλλά και γεωγραφικά στοιχεία όπως οι συντεταγμένες κεντροειδούς Κομητείας και η απόσταση αυτού από την ακτογραμμή (βλ. Κεφάλαιο 5). Η εισαγωγή των χαρακτηριστικών αυτών οδήγησε σε σημαντική βελτίωση της απόδοσης ($R^2 \approx 0.26$).

Η παραπάνω βελτίωση του δείκτη προσδιορισμού έδειξε ότι βρισκόμαστε στο σωστό δρόμο αλλά δεν επαρκούσε. Ακολούθησε η δοκιμή άλλων τύπων μοντελοποίησης με γνώμονα και βασικό κριτήριο την καλύτερη εφαρμογή στις συνθήκες, όπως είχαν

διαμορφωθεί με τις νέες μεταβλητές εισόδου. Δοκιμάστηκε το μοντέλο XGBoost Chen, T., & Guestrin, C. (2016)., το οποίο ενδείκνυται για τη διαχείριση στατικών και δυναμικών χωρικών δεδομένων και υποστηρίζει τη δυνατότητα κανονικοποίησης (ανάγκη άμεσα εμφανής δεδομένης της πολυπλοκότητας του προβλήματος) και αμέσως έδειξε να ανταποκρίνεται καλύτερα από το προηγούμενο μοντέλο που εφαρμόστηκε (Random Forest). Αξίζει να αναφερθεί ότι ο λόγος που το μοντέλου XGBoost είχε καλύτερη προσαρμογή στο πρόβλημα μας είναι η ικανότητα του να «διορθώνει» τα προηγούμενα δέντρα επιλογής (decision trees) (αντίθετα με το RF) και έτσι στα πολυδιάστατα προβλήματα δίνει ακριβέστερες προβλέψεις. Εν συνέχεια ορίστηκαν οι μεταβλητές εισόδου σε στατικά και δυναμικά δεδομένα (βλ. Κεφάλαιο 5.8) και ρυθμίστηκαν οι υπερ-παράμετροι για τη μέγιστη δυνατή απόδοση στο πρόβλημα μας, με αποτέλεσμα την αύξηση της απόδοσης ($R^2 \approx 0.39$). Η ρύθμιση υπερ-παραμέτρων είναι επιβεβλημένη σε πολύπλοκα προβλήματα, για να επιτευχθεί ο βέλτιστος βαθμός απόδοσης (Tepetidis, N., Koutsoyiannis, D., Iliopoulou, T. and Dimitriadis, P., 2024.)Τέτοιες παράμετροι είναι στην περίπτωση του XGBoost είναι:

- Αριθμός δέντρων (n_estimators)
- Μέγιστο βάθος δέντρων (max_depth)
- Ρυθμός εκμάθησης (learning_rate)
- Ποσοστά δεδομένων και χαρακτηριστικών που χρησιμοποιούνται (subsample / colsample_bytree)

Ακολούθησε η προσθήκη αλληλεπιδράσεων μεταξύ χαρακτηριστικών (feature interactions), με σκοπό την ενίσχυση της απόδοσης. Αναλυτικότερα έχει παρατηρηθεί ότι οι αλληλεπιδράσεις αφορούν στον συνδυασμό δύο ή περισσότερων μεταβλητών εισόδου, οι οποίες μαζί επηρεάζουν τη προβλεπόμενη μεταβλητή με τρόπο μη γραμμικό. Για παράδειγμα, η επίδραση του υψομέτρου μπορεί να διαφέρει ανάλογα με την απόσταση από τη θάλασσα, και έτσι να δημιουργείται πρακτικά ένα νέο χαρακτηριστικό, ως συνδυασμός των δύο αυτών παραμέτρων. Πρόκειται για μία τεχνική η οποία χρησιμοποιείται σε μεγάλο βαθμό στα προβλήματα με πολλές μεταβλητές εισόδου (όπως το δικό μας) και αυξάνει την αξιοπιστία των προβλέψεων. Στο ίδιο στάδιο εισήχθη το στοιχείο της κανονικοποίησης της μεταβλητής στόχου.

Αναλυτικότερα τα Αιτήματα Αποζημιώσεων έναντι πλημμυρικών φαινομένων εκφράστηκαν ανά 100.000 κατοίκους. Αυτό συνέβη, έτσι ώστε συνεπικουρικά με τη μεταβλητή της πληθυσμιακής πυκνότητας να εξαλειφθεί η εξάρτηση από το πληθυσμιακό μέγεθος των Κομητειών. Έτσι κατέστη δυνατή η σύγκριση μεταξύ Κομητειών με διαφορετικά δημογραφικά χαρακτηριστικά. Οι δύο αυτές προσθήκες βελτίωσαν τη σταθερότητα του μοντέλου, οδηγώντας σε υψηλότερη τιμή R^2 (0.50) και χαμηλότερα σφάλματα (MAE/RMSE). Ο συνδυασμός των παραπάνω στρατηγικών συνέβαλε σημαντικά στη βελτίωση της συνολικής απόδοσης του συστήματος πρόβλεψης.

Επόμενο στάδιο βελτίωσης απετέλεσε ο λογαριθμικός μετασχηματισμός στη μεταβλητή στόχο, κατόπιν κανονικοποίησης ανά πληθυσμό του προηγούμενου σταδίου (log(claims / population)). Με τη χρήση του λογαρίθμου περιορίστηκε η επίδραση των ακραίων τιμών (outliers) και επετεύχθη καλύτερη κατανομή των δεδομένων, διευκολύνοντας με τον τρόπο αυτό την εκπαίδευση του μοντέλου. Εν αντίθεση με τις αρχικές εκτιμήσεις μας, δεν επέφερε την αναμενόμενη αύξηση απόδοσης (R^2=0.56).

Έχοντας πλέον καταλήξει σε μία έκδοση του μοντέλου μας, ικανή να δώσει αξιόπιστες προβλέψεις, δεν υπήρχε περιθώριο για μικρές επεμβάσεις, όπως στα προηγούμενα στάδια. Στα πλαίσια λοιπόν της περαιτέρω αύξησης του βαθμού απόδοσης και λαμβάνοντας υπόψιν τις αλλαγές, οι οποίες έχουν σημειωθεί από την αρχική φάση του προβλήματος δοκιμάστηκαν και άλλοι τύποι μοντέλων. (Gradient Boosting (HistGB) Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., et al. (2011). και το CatBoost (Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018)). Το μοντέλο CatBoost αποδείχθηκε αυτό με την καλύτερη εφαρμογή στο πρόβλημα μας. Αυτό οφείλεται στο γεγονός ότι το μοντέλο CatBoost, εμφανίζει λίγο καλύτερες επιδόσεις σε προβλήματα με περιπτώσεις feature interaction (το κάνει και αυτόματα) και διαθέτει εσωτερικό boosting scheme ,το οποίο είχε καλή προσαρμογή στο μοντέλο μας. Οι δείκτες απόδοσης του τελικού μοντέλου μας περιγράφονται στον Πίνακα 6.1 (R^2=0.638, RMSE=0.81 και MAE=0.59)

Η εξελικτική διαδικασία, όπως αυτή προέκυψε μέσα από τα διαφορετικά στάδια, που αναλύθηκαν ανωτέρω παρουσιάζεται γραφικά στο Διάγραμμα 6.1:

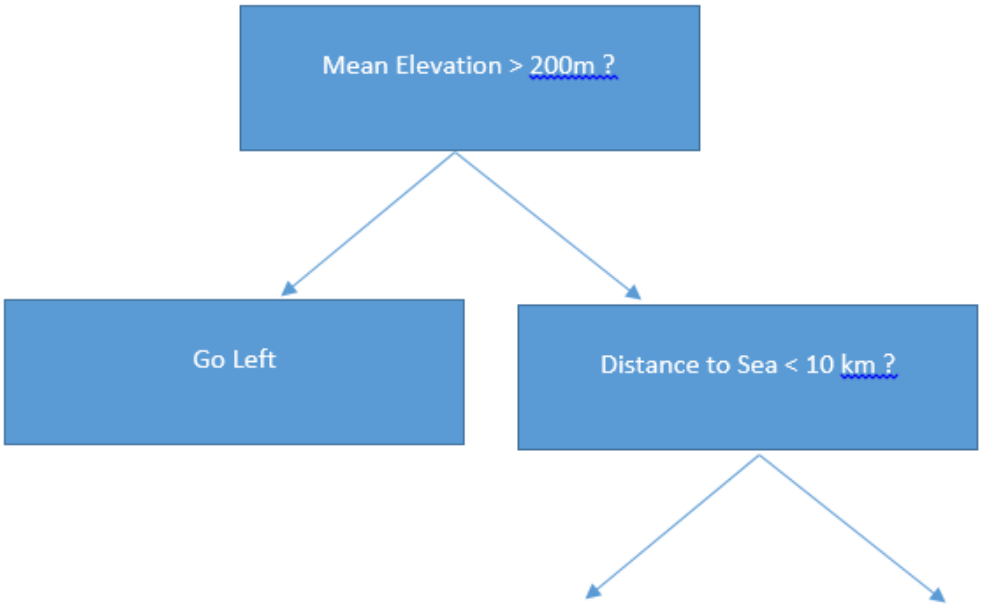
6.6 Μοντέλο CatBoost & Βασικές Αρχές Μηχανικής Εκμάθησης

Από τις παραπάνω ενότητες προέκυψε ότι το καταλληλότερο μοντέλο για πρόβλημα, που ετέθη είναι το μοντέλο CatBoost. Προκειμένου να γίνουν πλήρως κατανοητοί οι λόγοι της επιλογής αυτής και οι ιδιαιτερότητες του μοντέλου αυτού πρώτα θα γίνει μία σύντομη αναφορά στις βασικές αρχές της μηχανικής εκμάθησης και τους θεμελιώδεις μηχανισμούς της. (James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013))

Τα μοντέλα μηχανικής εκμάθησης γίνονται ευρέως αποδεκτά από την επιστημονική κοινότητα και χρησιμοποιούνται όλο και περισσότερο τα τελευταία χρόνια. Οι σύγχρονες καινοτομίες στον τομέα του προγραμματισμού και των υπολογιστών έφεραν σαρωτικές αλλαγές και επέτρεψαν τη χρήση αυτών των εργαλείων. Παλαιότερα ήταν σύνηθες να λειτουργούν τα προγραμματιστικά εργαλεία με μία μονοσήμαντη σχέση (πχ. Αν «σενάριο Α» , τότε «ενέργεια Β»). Τα μοντέλα μηχανικής εκμάθησης όμως δουλεύουν χωρίς κάποια οδηγία όπως ανωτέρω. Αντιθέτως, δέχονται ιστορικά δεδομένα ως μεταβλητές εισαγωγής και καλούνται να εντοπίσουν μόνα τους τις σχέσεις μεταξύ των μεταβλητών και εν κατακλείδι των προβλεπόμενων τιμών.

Ο τρόπος με τον οποίο εντοπίζει τα μοτίβα, που περιγράφονται παραπάνω το μηχανικό μοντέλο εκμάθησης και στον οποίο στηρίζεται όλη η φιλοσοφία της είναι η αρχή του δέντρου αποφάσεων (decision tree). Πιο αναλυτικά , το μοντέλο , με βάση τα δεδομένα , που έχει θέτει ερωτήματα/παρατηρήσεις και ανάλογα την απάντηση δημιουργείται ένα δέντρο αποφάσεων. Πρόκειται για ένα μονοπάτι του οποίο η κατεύθυνση καθορίζεται από τις ερωτήσεις. Στο τέλος λοιπόν του μονοπατιού το μοντέλο κάνει μια πρόβλεψη σχετικά με την προβλεπόμενη τιμή με βάση το μονοπάτι, που ακολούθησε (τις σχέσεις και ερωτήσεις δηλαδή που έθεσε στα δεδομένα εκπαίδευσης). Στην εικόνα 6.2. παρουσιάζεται ένα τυχαίο παράδειγμα από

το μοντέλο μηχανικής εκμάθησης, που αναπτύχθηκε κατά την έρευνα μας με δύο τυχαίες από τις μεταβλητές εισόδου μας. Στη συνέχεια του δέντρου προς τα κάτω θα προκύπτουν και άλλα κριτήρια, με βάση τα οποία θα δημιουργούνται και άλλα διαφορετικά μονοπάτια.



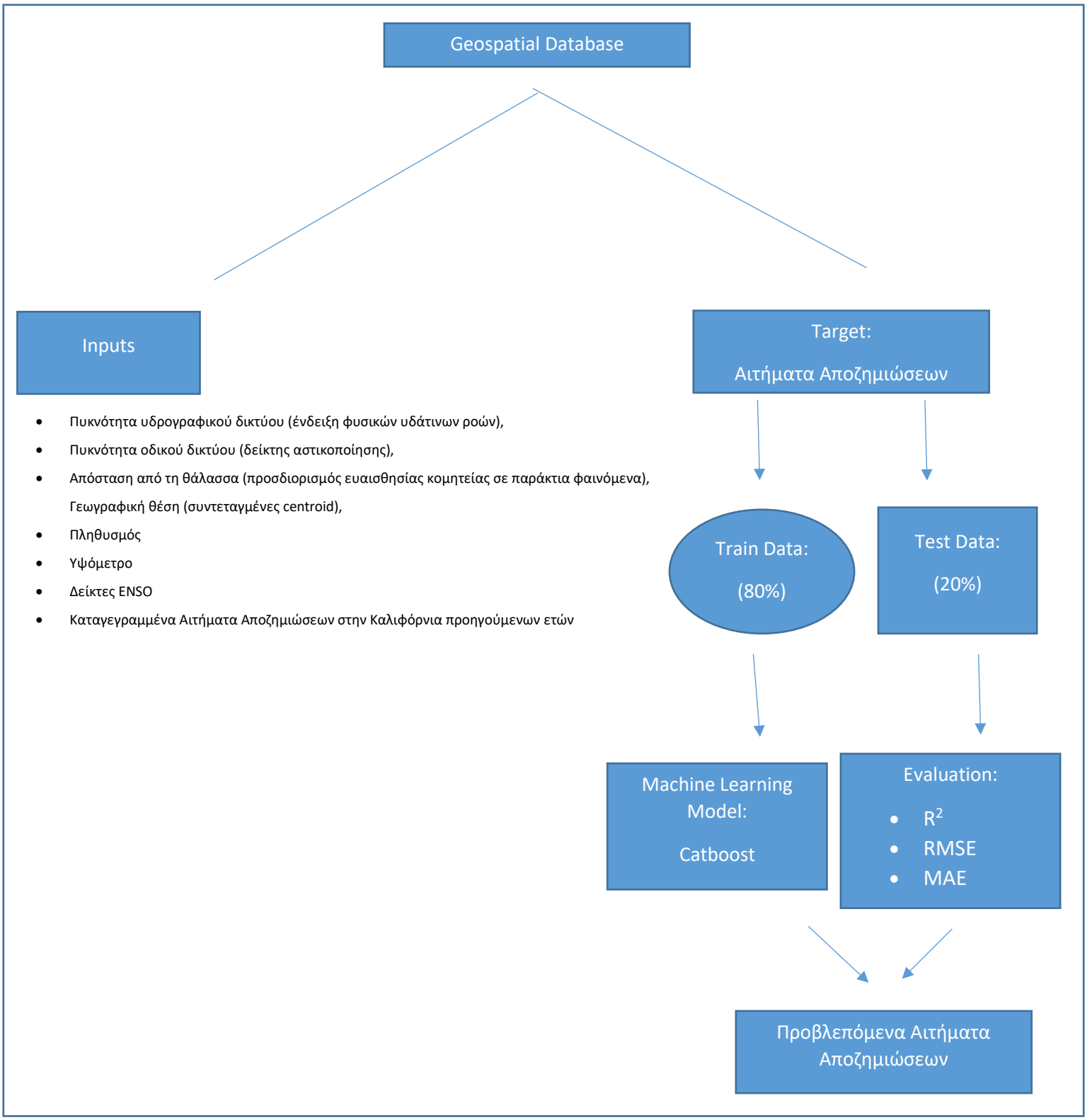
ΕΙΚΟΝΑ 6.2 – ΑΠΟΣΠΑΣΜΑ ΔΕΝΤΡΟΥ ΑΠΟΦΑΣΕΩΝ ΤΟΥ ΜΟΝΤΕΛΟΥ ΜΗΧ. ΕΚΜΑΘΗΣΗΣ ΠΟΥ ΑΝΑΤΠΥΧΘΗΚΕ ΣΤΗΝ ΠΑΡΟΥΣΑ ΕΡΓΑΣΙΑ

Σύντομα όμως παρατηρήθηκε το πρόβλημα της υπερπροσαρμογής (overlifting) με τα απλά δέντρα. Αναλυτικότερα, ακριβώς επειδή το δέντρο αυτό βασίζεται σε όλα τα δεδομένα, μαθαίνει αυτούσιες τις σχέσεις μεταξύ τους και αποτυγχάνει να «γενικεύσει» επιφέροντας έτσι μεγάλη αστάθεια. Δόθηκαν δύο λύσεις σε αυτό το πρόβλημα. Η πρώτη και απλούστερη είναι η τεχνική Random Forest, η οποία και απετέλεσε το πρώτο πρότυπο μοντέλου που δοκιμάσαμε, όπως περιγράφεται στην παραπάνω ενότητα. Με βάση την τεχνική αυτή, αντί για ένα δέντρο, δημιουργούνται περισσότερα, τα οποία όμως εκπαιδεύονται σε διαφορετικά τυχαία (random) υποσύνολα δεδομένων. Η απόφαση λοιπόν πια (ή πρόβλεψη της τιμής στόχου) δεν

είναι θέμα ενός μόνο δέντρου (συγκεκριμένων δεδομένων) αλλά συνιστώσα ενός ολόκληρου δάσους δέντρων και μάλιστα διαφορετικών τιμών μεταβλητών εισόδου. Η δεύτερη λύση που επινοήθηκε αφορά την τεχνική Gradient Boosting. Αποτελεί μία άλλη προσέγγιση συνεχούς βελτιστοποίησης του δέντρου απόφασης. Δημιουργείται αρχικά ένα δέντρο αποφάσεων. Το επόμενο δέντρο, εξαρτάται από το πρώτο, υπολογίζοντας τη διαφορά μεταξύ της προβλεπόμενης τιμής και της πραγματικής. Το τρίτο δέντρο ελέγχει το δεύτερο κ.ο.κ. Έτσι, με κάθε νέο δέντρο βελτιώνεται η ακρίβεια της πρόβλεψης. Σημαντικό πλεονέκτημα της τεχνικής αυτής είναι η ικανότητα να ενσωματώνει σχέσεις αλληλεπίδρασης μεταξύ των μεταβλητών εισόδου (αξιοποιήσαμε αυτή την δυνατότητα στο μοντέλο μας), στοιχείο που σε συνδυασμό με την υψηλή ακρίβεια που επιφέρει η σωστή ρύθμιση του μοντέλου, καθιστά την τεχνική αυτή πολύ καλή λύση.

Για το μοντέλο μας, όπως αναφέρθηκε παραπάνω, επιλέχθηκε ο τύπος CatBoost. Το CatBoost ανήκει στην δεύτερη κατηγορία, στην οικογένεια των Gradient Boosting Decision Trees (GBDT). Εμφανίζει εξαιρετικά πλεονεκτήματα, που το κατέστησαν την πιο ταιριαστή λύση στο πρόβλημα που μελετάται. Αρχικά ανιχνεύει αυτόματα αλληλεπιδράσεις μεταξύ χαρακτηριστικών ενώ χαρακτηρίζεται από υψηλή ακρίβεια. Επικράτησε σε σχέση με τα άλλα μοντέλα της ίδιας κατηγορίας καθώς εμφανίζει σπάνια prediction shift (συχνό πρόβλημα στην κατηγορία αυτή, αφού ενέχει το στοιχείο τα ης σύγκρισης με τις πραγματικές τιμές της προβλεπόμενης τιμής και έτσι χρησιμοποιείται και αυτή ως δεδομένο εκπαίδευσης, τείνουν δηλαδή τα αποτελέσματα προς τις σωστές τιμές, χωρίς να εκπαιδεύεται σωστά το μοντέλο, σε αλλά δεδομένα, θα αστοχούσε). Το μοντέλο CatBoost λοιπόν χρησιμοποιεί την τεχνική Ordered Boosting , και αποφεύγεται το πρόβλημα του prediction shift, ενώ με την ισορροπημένη διόρθωση που κάνει σε όλα τα δέντρα αποφεύγεται και η υπερπροσαρμογή στα δεδομένα (overlifting), που αναφέρθηκε παραπάνω.

Με βάση όλα τα παραπάνω διαμορφώθηκε και το τελικό μοντέλο μας CatBoost, του οποίου τα βασικά στοιχεία φαίνονται παρακάτω στην εικονα 6.



- Πυκνότητα υδρογραφικού δικτύου (ένδειξη φυσικών υδάτινων ροών),
- Πυκνότητα οδικού δικτύου (δείκτης αστικοποίησης),
- Απόσταση από τη θάλασσα (προσδιορισμός ευαισθησίας κομητείας σε παράκτια φαινόμενα), Γεωγραφική θέση (συντεταγμένες centroid),
- Πληθυσμός
- Υψόμετρο
- Δείκτες ENSO
- Καταγεγραμμένα Αιτήματα Αποζημιώσεων στην Καλιφόρνια προηγούμενων ετών

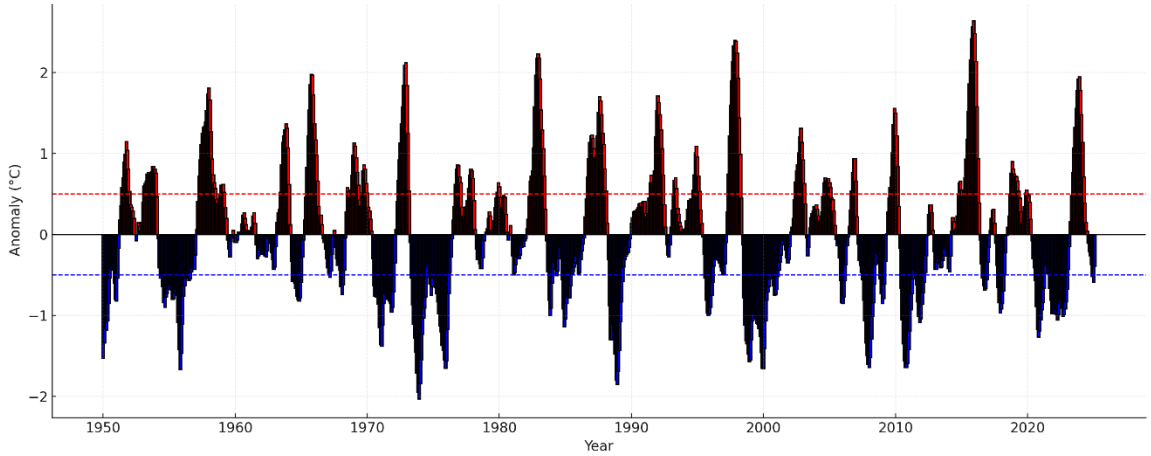
ΕΙΚΟΝΑ 6.3- ΔΙΑΓΡΑΜΜΑ ΡΟΗΣ ΤΟΥ ΤΕΛΙΚΟΥ ΜΟΝΤΕΛΟΥ CAT BOOST

7.Αποτελέσματα

7.1. Ερμηνεία Τάσεων ENSO (1950–2024)

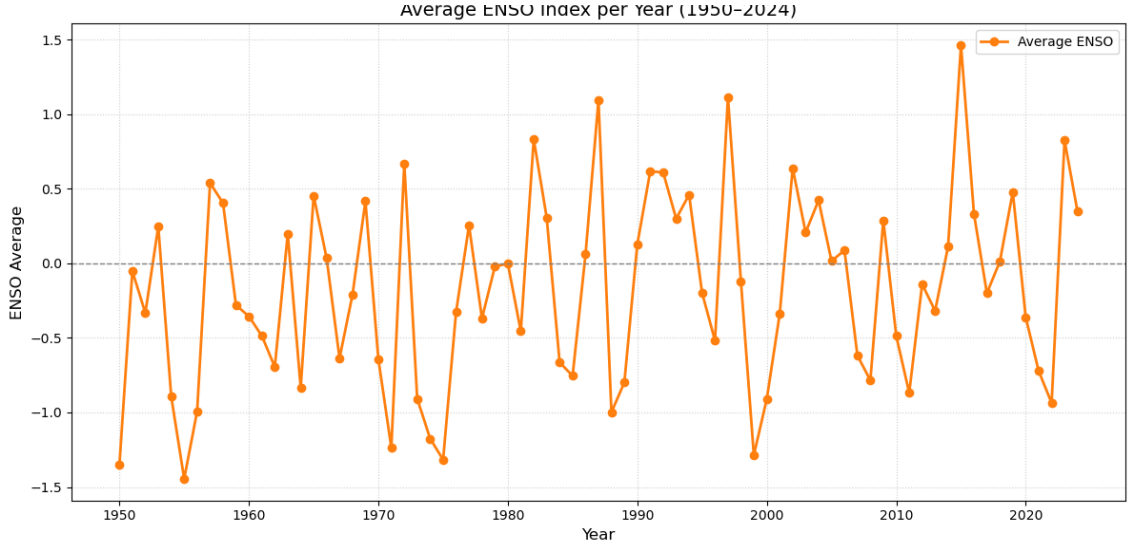
Τα φαινόμενα ENSO (El Niño–Southern Oscillation) αποτελούν έναν από τους σημαντικότερους σύγχρονους τομείς μελέτης και έρευνας σε παγκόσμιο επίπεδο, με ευρύτατες επιπτώσεις στην κλιματολογική σταθερότητα πολλών περιοχών του πλανήτη. Η βαθύτερη κατανόηση της μακροχρόνιας συμπεριφοράς του ENSO κρίνεται καθοριστική για την αξιολόγηση του πλημμυρικού κινδύνου και των συνεπακόλουθων κοινωνικοοικονομικών επιπτώσεων. Στο παρόν κεφάλαιο παρουσιάζονται οι ετήσιες τάσεις των φαινομένων ENSO για την περίοδο 1950–2024, μέσω ανάλυσης των μέσων, μέγιστων και ελάχιστων τιμών του δείκτη ENSO. Στόχος είναι η αναγνώριση κλιματικών μοτίβων, η αξιολόγηση της συχνότητας και έντασης των ακραίων επεισοδίων El Niño και La Niña, καθώς και η σύνδεσή τους με την παρατηρούμενη πλημμυρική δραστηριότητα στις Ηνωμένες Πολιτείες.

Εναρκτήριο στάδιο της ανάλυσης μας είναι η επεξεργασία της πρωτογενούς χρονοσειράς , δημοσιευμένης από το NOAA Physical Sciences Laboratory (PSL) (NOAA Physical Sciences Laboratory (PSL). (2023). Για την αποτύπωση της διαχρονικής μεταβολής του φαινομένου ENSO, χρησιμοποιήθηκαν οι τιμές του δείκτη ONI (Oceanic Niño Index), ο οποίος προκύπτει από τις ανωμαλίες στις τιμές της θερμοκρασίας στη στάθμη της Θάλασσας (Sea Surface Temperature) στην περιοχή Niño 3.4 του τροπικού Ειρηνικού (βλ. Κεφάλαιο 2 (2.3.2). Η χρονολογική βάση των δεδομένων εκτείνεται μεταξύ των ετών 1950 - 2024. Κατά την επεξεργασία, όλα τα τριμηνιαία παράθυρα μετατράπηκαν σε συνεχές χρονικό με σκοπό να απεικονιστούν με τη μεγαλύτερη δυνατή ακρίβεια οι μεταβολές της τιμής. Οι τιμές που υπερβαίνουν το όριο των +0.5 °C υποδηλώνουν επεισόδια El Niño, ενώ τιμές μικρότερες των −0.5 °C επεισόδια La Niña.

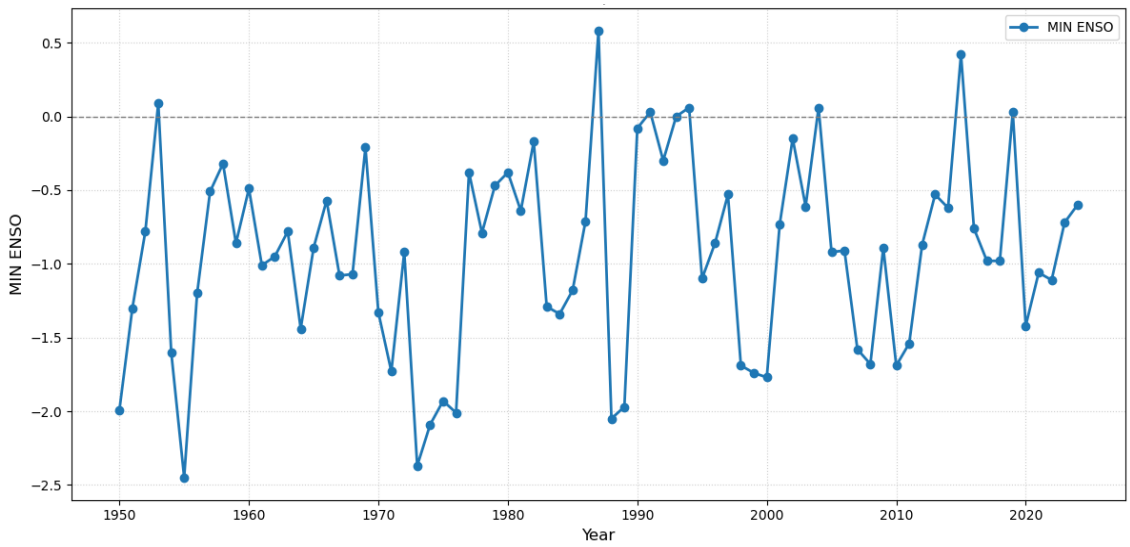


ΔΙΑΓΡΑΜΜΑ 7.1 ΤΡΙΜΗΝΙΑΙΕΣ ΜΕΣΕΣ ΤΙΜΕΣ ΤΩΝ ΘΕΡΜΟΚΡΑΣΙΑΚΩΝ ΑΝΩΜΑΛΙΩΝ (ΣΕ °C) ΣΤΗΝ ΠΕΡΙΟΧΗ ΝΙΝΟ 3.4 ΤΟΥ ΕΙΡΗΝΙΚΟΥ ΩΚΕΑΝΟΥ, ΓΙΑ ΤΗΝ ΠΕΡΙΟΔΟ 1950–2025.

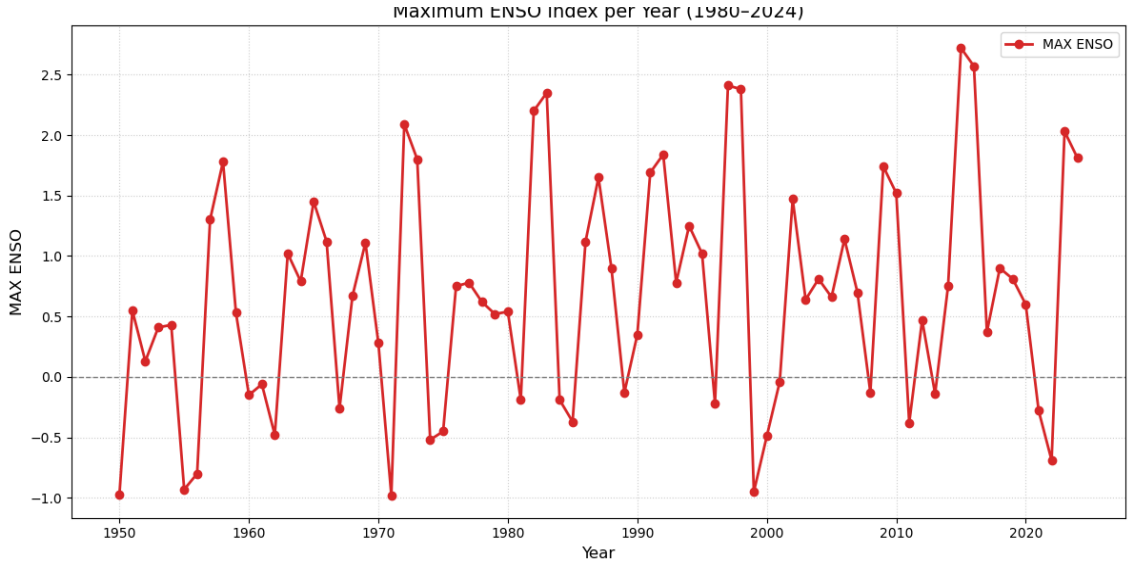
Η παρουσίαση των ετήσιων δεικτών ENSO για την περίοδο 1950–2024, που ακολουθεί στα παρακάτω διαγράμματα καθιστά εμφανείς κλιματικές τάσεις, άρρηκτα συνδεδεμένες με τη συχνότητα και την ένταση των φαινομένων El Niño και La Niña. Τα τρία διαγράμματα που παρουσιάζονται (Διάγραμμα 7.2. (μέσος όρος δείκτη ENSO), Διάγραμμα 7.3. (ελάχιστη τιμή δείκτη ENSO) και Διάγραμμα 7.4. (μέγιστη τιμή δείκτη ENSO) επιτυγχάνουν τη μεταφορά μιας ενδεικτικής εικόνας στον αναγνώστη για τα παραπάνω φαινόμενα.



ΔΙΑΓΡΑΜΜΑ 7.2. ΕΤΗΣΙΑ ΕΞΕΛΙΞΗ ΤΟΥ ΜΕΣΟΥ ΔΕΙΚΤΗ ENSO ΓΙΑ ΤΗΝ ΠΕΡΙΟΔΟ 1950–2024



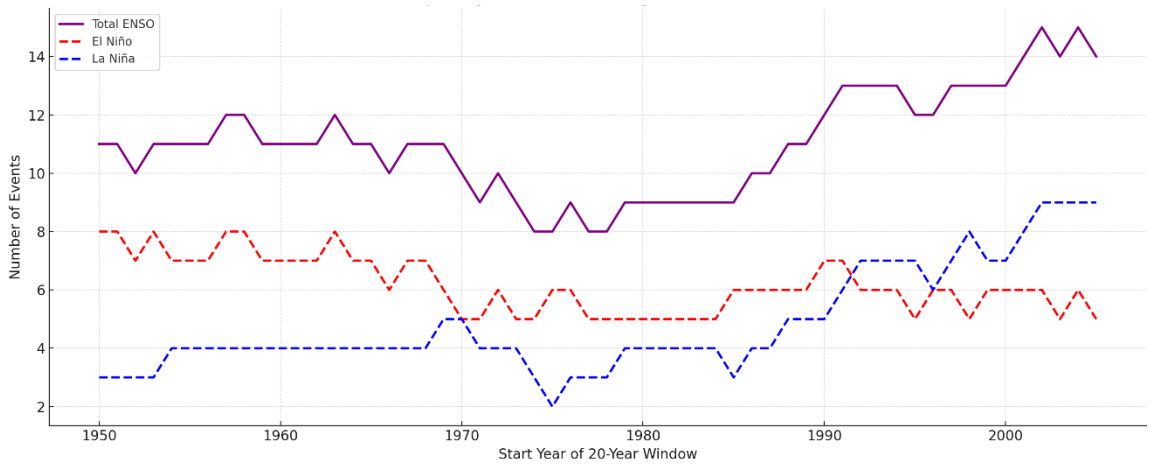
ΔΙΑΓΡΑΜΜΑ 7.3 ΕΤΗΣΙΑ ΕΞΕΛΙΞΗ ΤΟΥ ΕΛΑΧΙΣΤΟΥ ΔΕΙΚΤΗ ENSO ΓΙΑ ΤΗΝ ΠΕΡΙΟΔΟ 1950–2024



ΔΙΑΓΡΑΜΜΑ 7.4 ΕΤΗΣΙΑ ΕΞΕΛΙΞΗ ΤΟΥ ΜΕΓΙΣΤΟΥ ΔΕΙΚΤΗ ENSO ΓΙΑ ΤΗΝ ΠΕΡΙΟΔΟ 1950–2024

Το διάγραμμα του μέσου δείκτη ENSO χαρακτηρίζεται από επαναλαμβανόμενες φάσεις θετικών και αρνητικών τιμών. Ωστόσο, ιδιαίτερη σημασία έχουν οι μέγιστες και ελάχιστες τιμές, οι οποίες εκφράζουν την παρουσία επεισοδίων El Nino / La Nina. Από τη δεκαετία του 1980 και μετά, παρατηρείται αύξηση της συχνότητας

(Διάγραμμα 7.5.) των φαινομένων αυτών, όπως επισημαίνεται και σε πρόσφατες διεθνείς μελέτες (Cai et al. 2020).



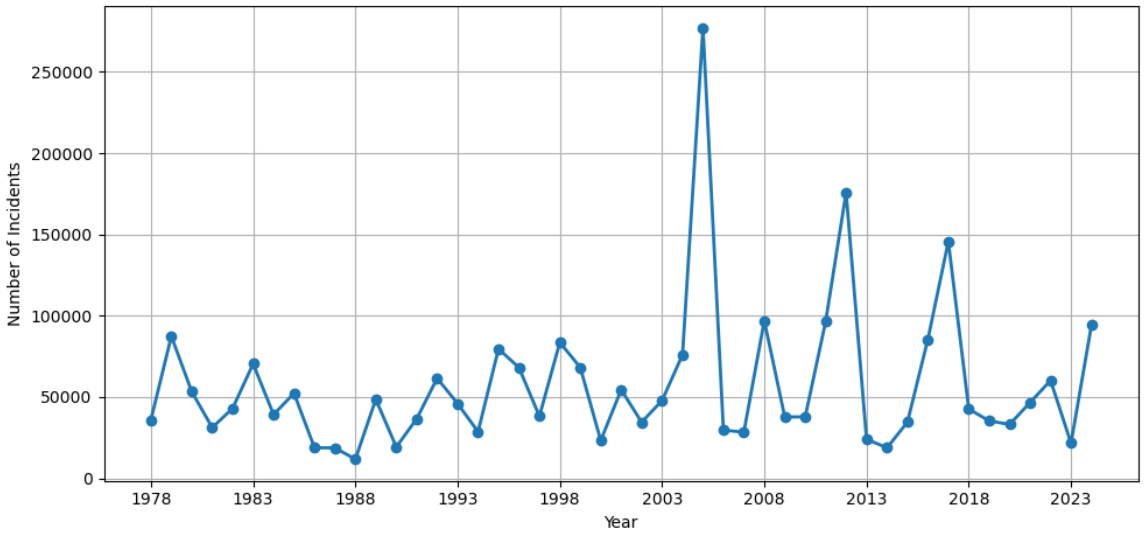
ΔΙΑΓΡΑΜΜΑ 4.5.: ΣΥΧΝΟΤΗΤΑ ΕΠΕΙΣΟΔΙΩΝ ENSO (EL NIÑO ΚΑΙ LA NIÑA) ΣΕ ΚΙΝΟΥΜΕΝΑ ΠΑΡΑΘΥΡΑ 20 ΕΤΩΝ, ΓΙΑ ΤΗΝ ΠΕΡΙΟΔΟ 1950–2025. ΚΑΘΕ ΣΗΜΕΙΟ ΤΗΣ ΚΑΜΠΥΛΗΣ ΑΝΤΙΣΤΟΙΧΕΙ ΣΤΟΝ ΑΡΙΘΜΟ ΤΩΝ ENSO ΕΠΕΙΣΟΔΙΩΝ ΠΟΥ ΚΑΤΑΓΡΑΦΗΚΑΝ ΕΝΤΟΣ ΤΟΥ ΑΝΤΙΣΤΟΙΧΟΥ 20ΕΤΟΥΣ ΔΙΑΣΤΗΜΑΤΟΣ

Αναλύοντας το μέσο δείκτη ENSO γίνεται εύκολα αντιληπτό ότι χαρακτηρίζεται από ημι-περιοδική εμφάνιση, με φάσεις El Niño και La Niña να εκδηλώνονται κάθε 3–7 έτη, χωρίς όμως αυστηρά σταθερή περιοδικότητα (Trenberth, 1997).

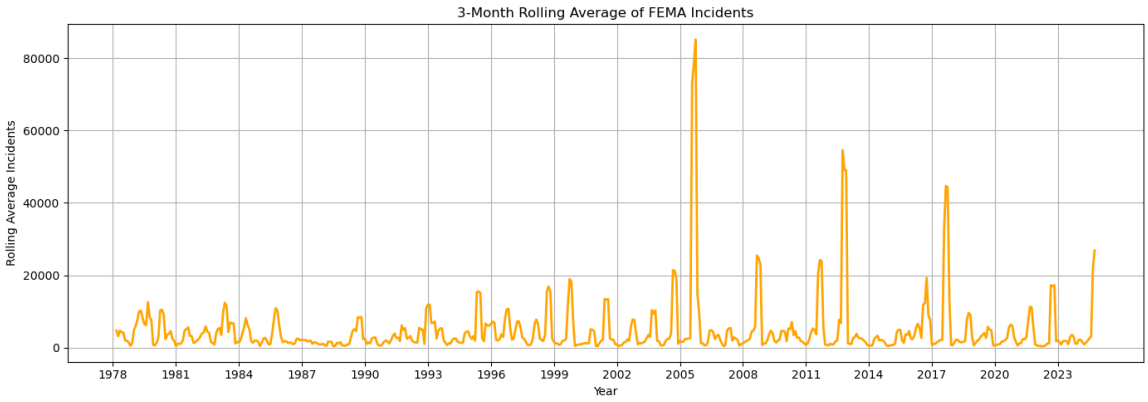
Εύκολα διαπιστώνει κανείς ότι τα έτη 1982–83, 1997–98, 2015–16 κάνουν την εμφάνιση τους έντονα El Niño, φαινόμενα ενώ τα 1973–74, 2007–08, 2010–11 χαρακτηρίζονται από έντονες συνθήκες φαινομένων La Niña.
Η παρουσία τέτοιων διακυμάνσεων δικαιολογεί και την ένταξή τους σε μοντέλα πρόβλεψης κινδύνου, καθώς συχνά συνδέονται με αιχμές στις αποζημιώσεις πλημμύρας.

7.2 Χρονική και Χωρική Ανάλυση Αιτημάτων Αποζημίωσης

Στα πλαίσια της παρούσας μελέτης πραγματοποιήθηκε ανάλυση των δεδομένων αιτήσεων αποζημίωσης FEMA (βλ. Ενότητα 3.4). Στο Διάγραμμα 7.6. απεικονίζεται ο συνολικός αριθμός αιτημάτων ανά έτος. Οι κυριότερες αιχμές παρατηρούνται το 2005, έτος κατά το οποίο έλαβε χώρα ο καταστροφικός τυφώνας Katrina, καθώς και το 2012 (τυφώνας Sandy). Επίσης αναδεικνύονται εύκολα οι αυξήσεις των αιτημάτων τα έτη 2017 και 2021 σχετίζονται αντίστοιχα με τους τυφώνες Harvey και Ida. Βέβαια οι χρονικές περίοδοι κατά τις οποίες σημειώθηκαν οι αιχμές αυτές αποτελούν κρίσιμες περιόδους για την ανάλυση της σχέσης φαινομένων ENSO–πλημμυρικών φαινομένων, καθώς ανήκουν σε έτη με ενεργά φαινόμενα El Niño ή La Niña. Στο Διάγραμμα 7.7. παρουσιάζεται ο κυλιόμενος μέσος όρος 3 μηνών, επιτρέποντας την σαφέστερη οπτικοποίηση εποχιακών τάσεων που οφείλονται σε ακραία καιρικά φαινόμενα. Πρόκειται για τα ίδια δεδομένα αλλά με έναν αναλυτικότερο τρόπο παράθεσης τους.

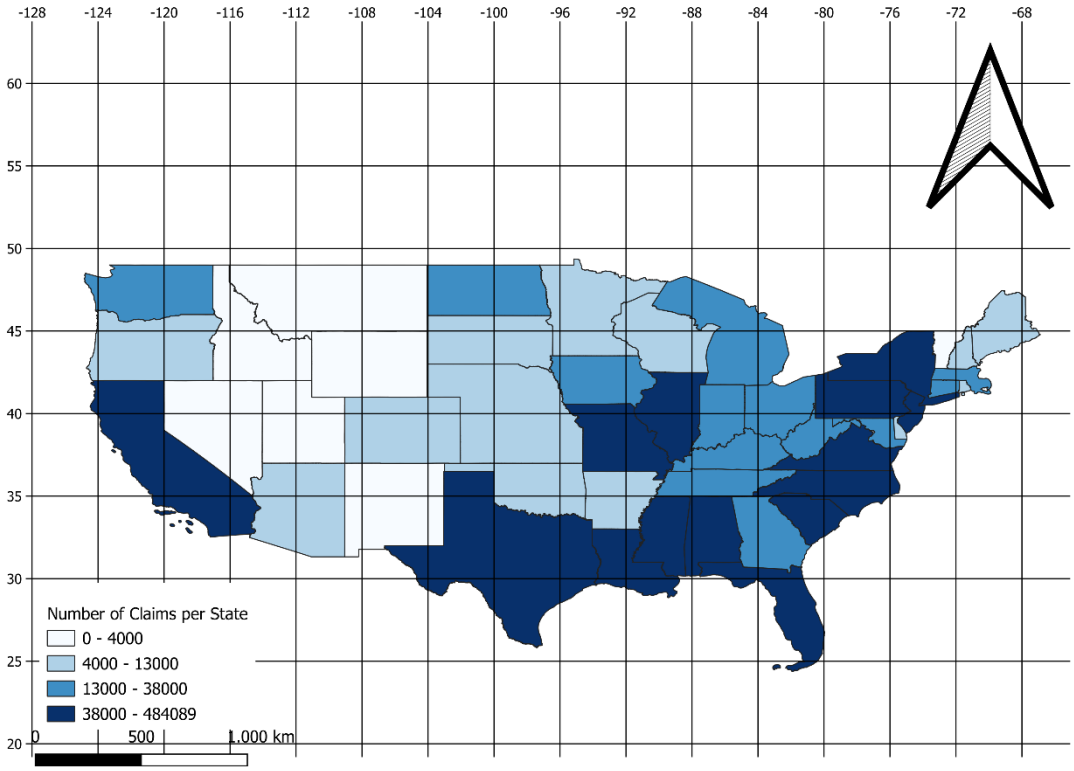


ΔΙΑΓΡΑΜΜΑ 4.6-ΕΤΗΣΙΟΣ ΣΥΝΟΛΙΚΟΣ ΑΡΙΘΜΟΣ ΑΙΤΗΣΕΩΝ ΑΠΟΖΗΜΙΩΣΗΣ FEMA ΑΝΑ ΕΤΟΣ



ΔΙΑΓΡΑΜΜΑ 4.7 ΚΥΛΙΟΜΕΝΟΣ ΜΕΣΟΣ ΟΡΟΣ 3 ΜΗΝΩΝ ΓΙΑ ΤΟΝ ΑΡΙΘΜΟ ΤΩΝ ΑΙΤΗΣΕΩΝ ΑΠΟΖΗΜΙΩΣΗΣ FEMA (1977–2024)

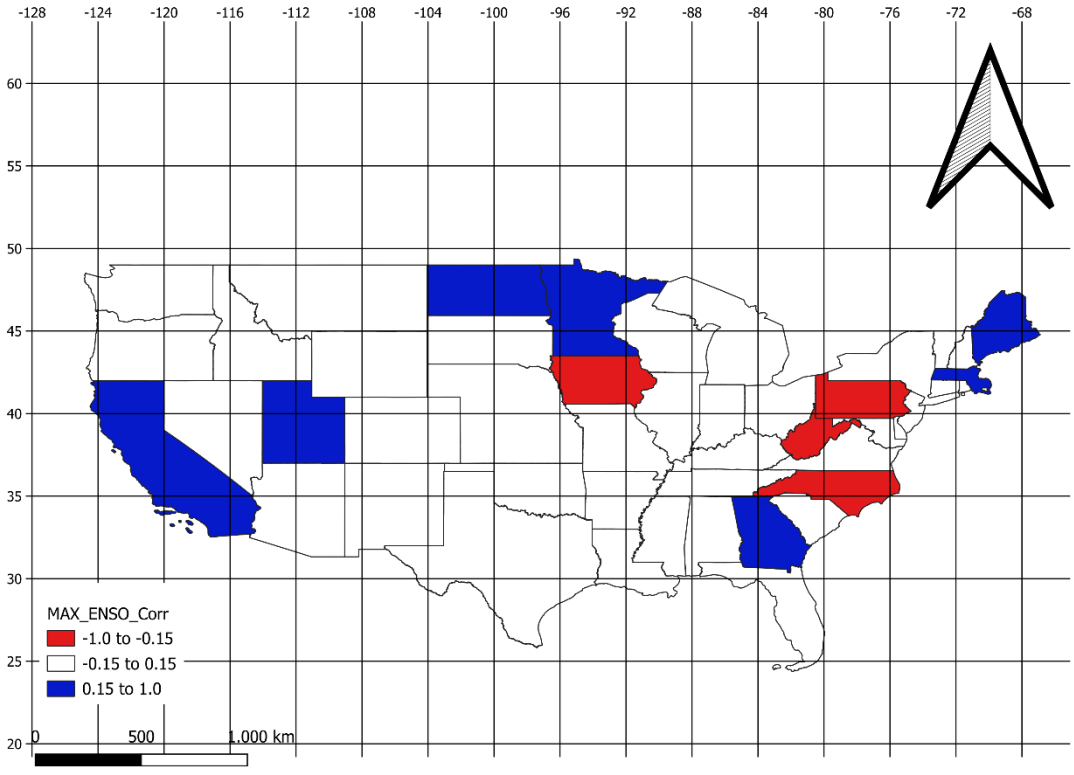
Η παραπάνω ανάλυση αφορά τη χρονική διασπορά των δεδομένων. Εξίσου ενδιαφέρουσα αποδεικνύεται η μελέτη της χωρικής διασποράς των Αιτημάτων αποζημίωσης της βάσης δεδομένων Fema.



ΕΙΚΟΝΑ 7.1.- ΧΩΡΙΚΗ ΚΑΤΑΝΟΜΗ ΤΟΥ ΣΥΝΟΛΙΚΟΥ ΑΡΙΘΜΟΥ ΑΙΤΗΜΑΤΩΝ ΑΠΟΖΗΜΙΩΣΗΣ ΑΝΑ ΠΟΛΙΤΕΙΑ ΣΤΙΣ ΗΝΩΜΕΝΕΣ ΠΟΛΙΤΕΙΕΣ

Ο χάρτης της εικόνας 7.1 απεικονίζει τον συνολικό αριθμό αιτημάτων αποζημίωσης λόγω πλημμυρών σε επίπεδο Πολιτείας στις Ηνωμένες Πολιτείες. Εύκολα παρατηρεί κανείς ότι τα περισσότερα αιτήματα εντοπίζονται στις νοτιοανατολικές και νοτιοκεντρικές Πολιτείες όπως η Καλιφόρνια, η Φλόριντα, το Τέξας, η Λουιζιάνα και η Βόρεια Καρολίνα (Smith et al. (2013)). Αυτό συνδέεται άμεσα με τη συχνότητα εμφάνισης τροπικών τυφώνων και καταιγίδων στις προαναφερθείσες περιοχές. Αντίθετα, Πολιτείες στο εσωτερικό της χώρας ή στα βορειοδυτικά εμφανίζουν σαφώς χαμηλότερα αιτήματα αποζημίωσης έναντι πλημμυρικής δραστηριότητας.

7.3 Συσχετίσεις Μεταξύ ENSO και Αιτημάτων Αποζημίωσης



ΕΙΚΟΝΑ 7.2: ΧΩΡΙΚΗ ΣΥΣΧΕΤΙΣΗ ΜΕΓΙΣΤΟΥ ΕΤΗΣΙΟΥ ΔΕΙΚΤΗ ENSO ΚΑΙ ΑΙΤΗΣΕΩΝ ΑΠΟΖΗΜΙΩΣΕΩΝ ΠΛΗΜΜΥΡΩΝ ΣΤΙΣ Η.Π.Α. (1980–2024)

Στην παραπάνω εικόνα 7.2. απεικονίζεται η συσχέτιση μεταξύ του μέγιστου ετήσιου δείκτη ENSO (El Niño–Southern Oscillation) και του αριθμού αιτήσεων αποζημίωσης

λόγω πλημμυρών για κάθε Πολιτεία των Ηνωμένων Πολιτειών. Η συσχέτιση έχει προσδιοριστεί με χρήση του συντελεστή συσχέτισης, κατανεμημένο σε τρεις κατηγορίες:Οι κατηγορίες αυτές είναι οι εξής:

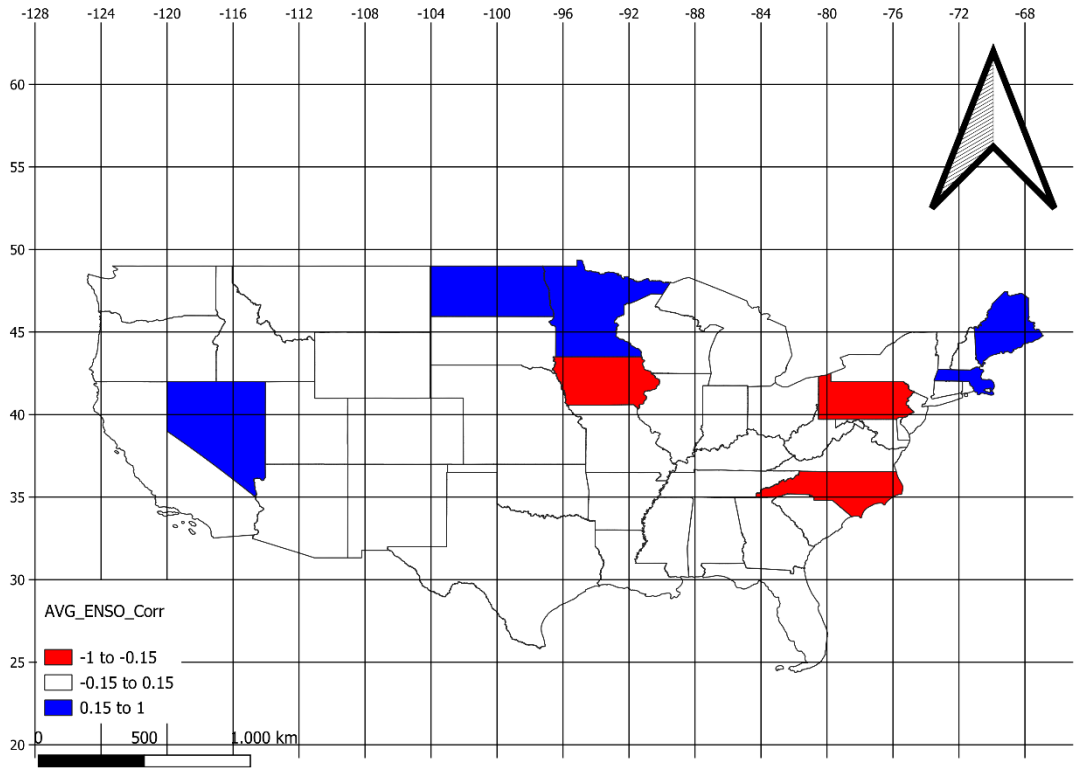
Για τιμές βαθμού συσχέτισης:

- Κατηγορία 1: από -1.00 έως -0.15 (κόκκινη απόχρωση)
- Κατηγορία 2: από -0.15 έως 0.15 (λευκή απόχρωση)
- Κατηγορία 3: από 0.15 έως 1.00 (μπλέ απόχρωση)

Παρατηρείται σημαντική χωρική μεταβλητότητα, γεγονός που αποτυπώνει τη μη ομοιόμορφη επίδραση του ENSO στον υδρολογικό κίνδυνο σε διαφορετικές περιοχές (Kunkel et al., 2003· Hamlet & Lettenmaier, 2007).

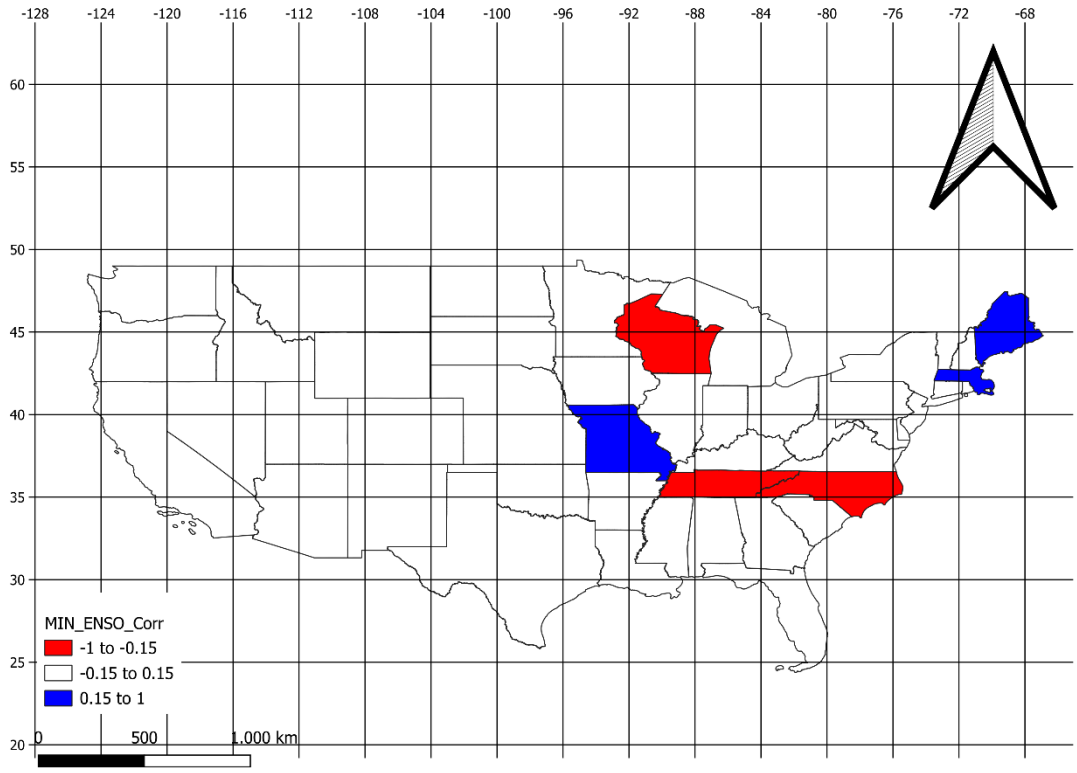
Αναλυτικότερα, η Πολιτεία της Καλιφόρνια σημειώνει τη μεγαλύτερη θετική συσχέτιση (0.345), όπως φαίνεται και από τα αποτελέσματα προηγούμενων μελετών, σύμφωνα με τα οποία οι έντονες φάσεις του φαινομένου El Niño προκαλούν αυξημένες βροχοπτώσεις και πλημμυρικά φαινόμενα στην πολιτεία (Cayan et al., 1999). Αντίστοιχα, οι πολιτείες του ανατολικού και νοτιοανατολικού τμήματος, όπως η Βόρεια Καρολίνα, η Πενσυλβάνια και η Αϊόβα, παρουσιάζουν αρνητική συσχέτιση, η οποία πιθανώς συνδέεται με την τάση του El Niño να μειώνει τα ύψη βροχόπτωσης σε αυτές τις περιοχές κατά τη διάρκεια των χειμερινών μηνών (Ropelewski & Halpert, 1986).

Η γεωγραφική κατανομή που προκύπτει ενισχύει την αντίληψη ότι το ENSO δεν επηρεάζει με τον ίδιο τρόπο όλες τις περιοχές των Η.Π.Α. Αντίθετα, τοπικά χαρακτηριστικά, όπως η μορφολογία του εδάφους και η αστικοποίηση καθιστά τη συσχέτιση φαινομένων ENSO και πλημμυρικών καταστροφών περίπλοκη (Hamlet & Lettenmaier, 2007).



ΕΙΚΟΝΑ 7.3 ΧΩΡΙΚΗ ΣΥΣΧΕΤΙΣΗ ΜΕΣΟΥ (AVERAGE) ΕΤΗΣΙΟΥ ΔΕΙΚΤΗ ENSO ΚΑΙ ΑΙΤΗΣΕΩΝ ΑΠΟΖΗΜΙΩΣΕΩΝ ΠΛΗΜΜΥΡΩΝ ΣΤΙΣ Η.Π.Α. (1980–2024)

Η παραπάνω εικόνα – χάρτης 7.3 παρουσιάζει τη συσχέτιση μεταξύ του μέσου ετήσιου δείκτη ENSO (Average ENSO) και του συνολικού αριθμού αιτημάτων Αποζημίωσης ανά Πολιτεία στις Ηνωμένες Πολιτείες. Κατά το πρότυπο, που τηρείται στα πλαίσια αυτής της πτυχιακής μελέτης με μπλε απόχρωση παρουσιάζονται οι περιοχές με τις θετικότερες τιμές, άνω του ορίου 0.15 (όπως η Νεβάδα, Νέα Υόρκη, Βόρεια Ντακότα και Μέιν). Αντίθετα, οι Πολιτείες με αρνητική συσχέτιση (όπως η Πενσυλβάνια, Καρολίνα του Βορρά, Βιρτζίνια και Αϊόβα) παρουσιάζονται με κόκκινη απόχρωση, σημειώνοντας τιμές κάτω του ορίου -0.15

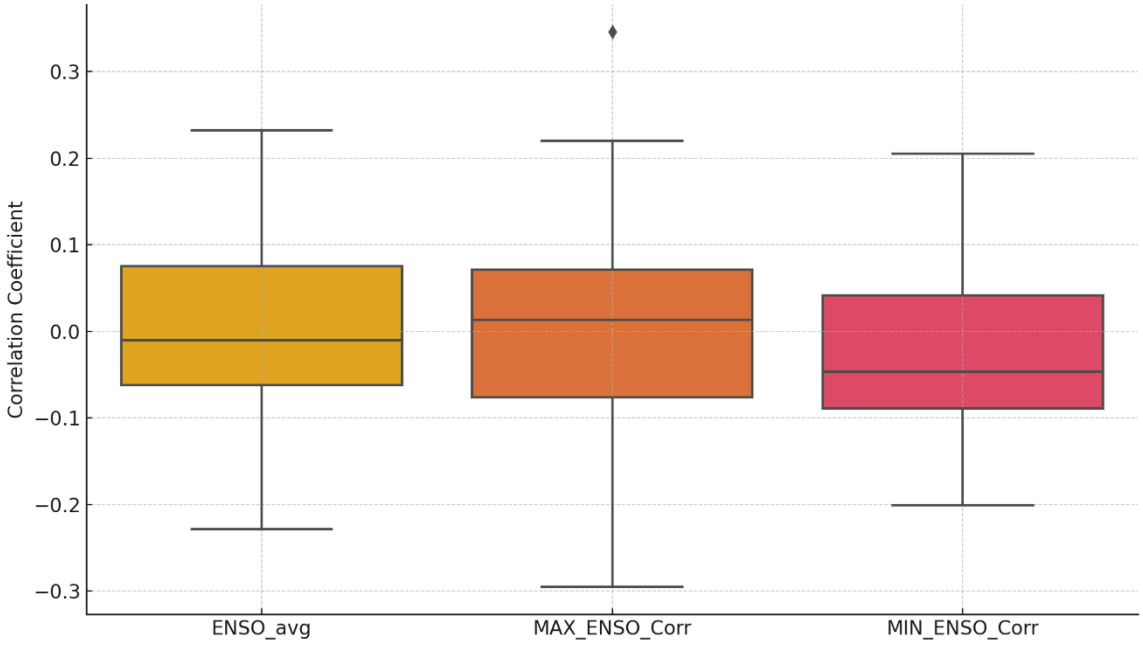


ΕΙΚΟΝΑ 7.4: ΧΩΡΙΚΗ ΣΥΣΧΕΤΙΣΗ ΕΛΑΧΙΣΤΟΥ ΕΤΗΣΙΟΥ ΔΕΙΚΤΗ ENSO ΚΑΙ ΑΙΤΗΣΕΩΝ ΑΠΟΖΗΜΙΩΣΕΩΝ ΠΛΗΜΜΥΡΩΝ ΣΤΙΣ Η.Π.Α. (1980–2024)

Η παραπάνω εικόνα – χάρτης 7.4 απεικονίζει τη συσχέτιση μεταξύ των ελάχιστων ετήσιων τιμών του δείκτη ENSO (MIN ENSO, συνθήκες La Niña) και των αιτημάτων Αποζημίωσης ανά Πολιτεία. Κατά το πρότυπο, που τηρείται στα πλαίσια αυτής της πτυχιακής μελέτης με μπλε απόχρωση παρουσιάζονται οι περιοχές με τις θετικότερες τιμές, άνω του ορίου 0.15 (όπως η Νέα Υόρκη και το Μιζούρι). Αντίθετα οι Πολιτείες με αρνητική συσχέτιση, (όπως η Ουισκόνσιν, Καρολίνα του Βορρά, Τενεσί και Κεντάκι) παρουσιάζονται με κόκκινη απόχρωση με όριο το -0.15.

Συμπερασματικά προκύπτει ότι τα Αιτήματα Αποζημιώσεων, ενόσω συνδέονται με τα φαινόμενά ENSO δε σχετίζονται μονοσήμαντα. Αντίθετα επηρεάζονται από πλήθος άλλων παραγόντων όπως τοπογραφικές και κλιματολογικές συνθήκες, γεωγραφική θέση κάθε περιοχής κ.α.

Ακολουθεί στατιστική ανάλυση των αποτελεσμάτων στα παρακάτω διαγράμματα και πίνακες

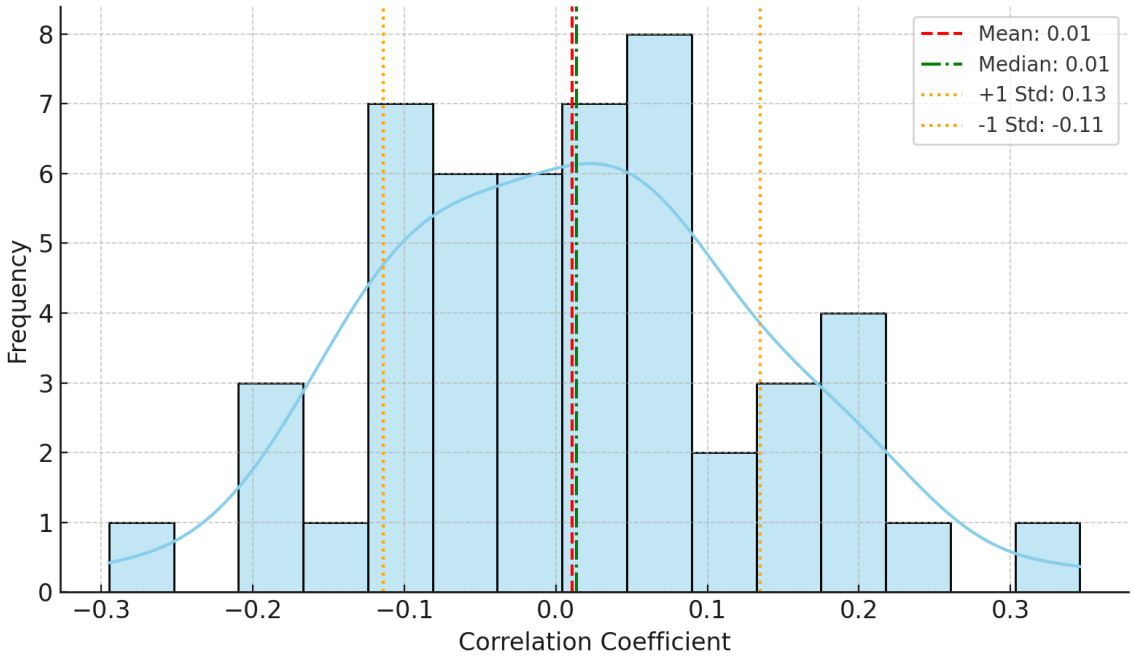


ΔΙΑΓΡΑΜΜΑ BOXPLOT 4.8 ΔΙΑΣΠΟΡΑ ΤΩΝ ΣΥΣΧΕΤΙΣΕΩΝ ΜΕΤΑΞΥ ΔΕΙΚΤΩΝ ENSO (ΜΕΣΟΣ ΟΡΟΣ, ΜΕΓΙΣΤΟ ΚΑΙ ΕΛΑΧΙΣΤΟ) ΚΑΙ ΤΩΝ ΑΙΤΗΜΑΤΩΝ ΑΠΟΖΗΜΙΩΣΗΣ ΑΝΑ ΠΟΛΙΤΕΙΑ ΤΩΝ Η.Π.Α.

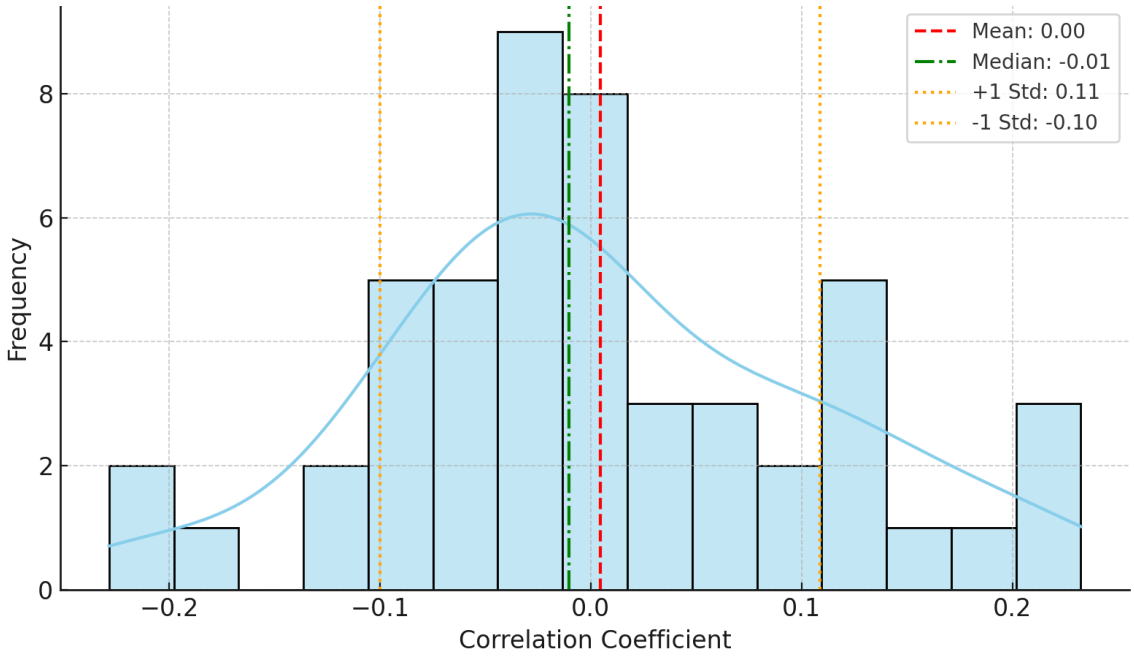
Το διάγραμμα boxplot 7.8 εκφράζει χαρακτηριστικά της κατανομής των συσχετίσεων μεταξύ δεικτών ENSO (διάμεσος, 1ο και 3ο τεταρτημόριο, μέγιστο και ελάχιστο) και των Αιτημάτων Αποζημίωσης ανά Πολιτεία των Η.Π.Α.Η γραμμή εντός του κουτιού δείχνει το διάμεσο (median), ενώ τα όρια του κουτιού δείχνουν το 1ο και 3ο τεταρτημόριο (Q1 και Q3). Τα ευθύγραμμα τμήματα, εκτός του κουτιού οριοθετούνται από τα ελάχιστα και μέγιστα μη ακραία σημεία, ενώ τυχόν ακραίες τιμές (outliers) εμφανίζονται ως ξεχωριστά σημεία εκτός των τμημάτων αυτών. Οι τιμές που διαμορφώνουν το παρακάτω διάγραμμα προβάλλονται στον παρακάτω πίνακα 7.1. αλλά και στις γραφικές απεικονίσεις (Διαγράμματα 7.9 – 7.11) των κατανομών, που ακολουθούν:

ΠΙΝΑΚΑΣ 7.1 ΠΙΝΑΚΑΣ ΣΤΑΤΙΣΤΙΚΗΣ ΣΥΣΧΕΤΙΣΕΩΝ ΜΕΤΑΞΥ ΔΕΙΚΤΩΝ ENSO (ΜΕΣΟΣ ΟΡΟΣ, ΜΕΓΙΣΤΟ ΚΑΙ ΕΛΑΧΙΣΤΟ) ΚΑΙ ΤΩΝ ΑΙΤΗΜΑΤΩΝ ΑΠΟΖΗΜΙΩΣΗΣ ΑΝΑ ΠΟΛΙΤΕΙΑ ΤΩΝ Η.Π.Α. (ΜΕΣΟΣ ΟΡΟΣ, ΤΥΠΙΚΗ ΑΠΟΚΛΙΣΗ, ΔΙΑΜΕΣΟΣ, ΚΑΘΩΣ ΚΑΙ ΤΩΝ ΤΕΤΑΡΤΗΜΟΡΙΩΝ (Q1 ΚΑΙ Q3) ΓΙΑ ΚΑΘΕ ΔΕΙΚΤΗ)

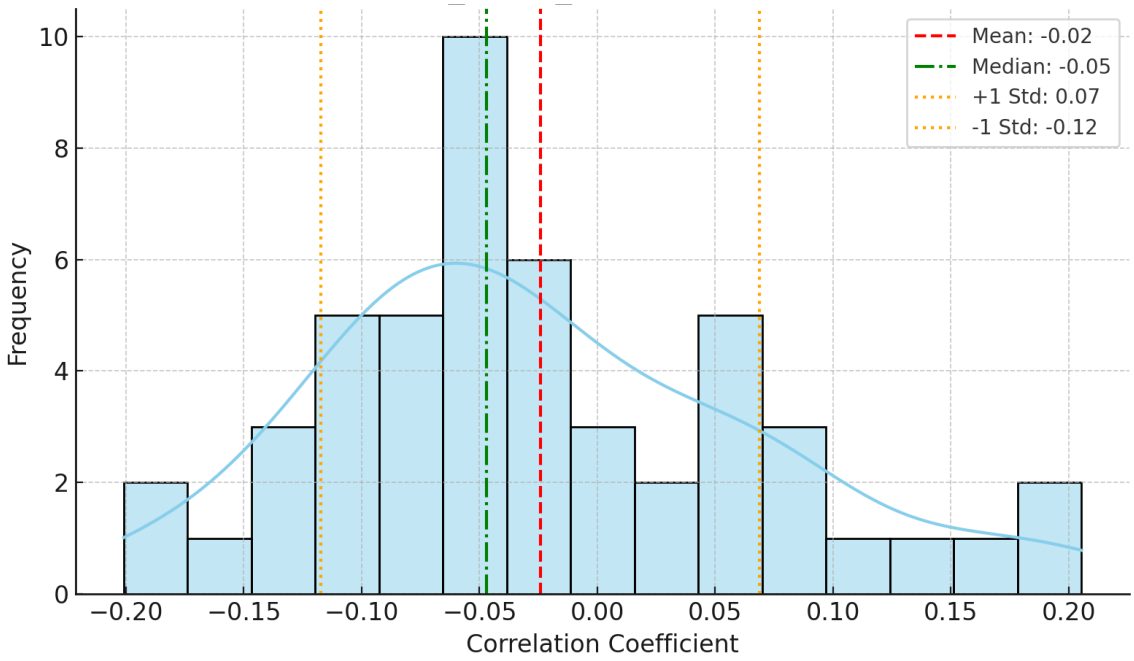
	Mean	Standard Deviation	Median (Q2)	1st Quartile (Q1)	3rd Quartile (Q3)
MAX_ENSO_Corr	0,0105	0,1244	0,0133	-0,0761	0,0711
AVG_ENSO_Corr	0,0045	0,1043	-0,0103	-0,0621	0,0753
MIN_ENSO_Corr	-0,0242	0,0931	-0,0469	-0,0894	0,0417



ΔΙΑΓΡΑΜΜΑ 7.9: ΔΙΑΓΡΑΜΜΑ ΚΑΤΑΝΟΜΗΣ ΣΥΣΧΕΤΙΣΗΣ ΔΕΙΚΤΗ ΜΕΓΙΣΤΟΥ ENSO (MAX_ENSO) ΣΕ ΟΛΕΣ ΤΙΣ ΠΟΛΙΤΕΙΕΣ ΤΩΝ Η.Π.Α.



ΔΙΑΓΡΑΜΜΑ 7.10 ΔΙΑΓΡΑΜΜΑ ΚΑΤΑΝΟΜΗΣ ΣΥΣΧΕΤΙΣΗΣ ΔΕΙΚΤΗ Μ.Ο. ENSO (ENSO_AVG) ΣΕ ΟΛΕΣ ΤΙΣ ΠΟΛΙΤΕΙΕΣ ΤΩΝ Η.Π.Α

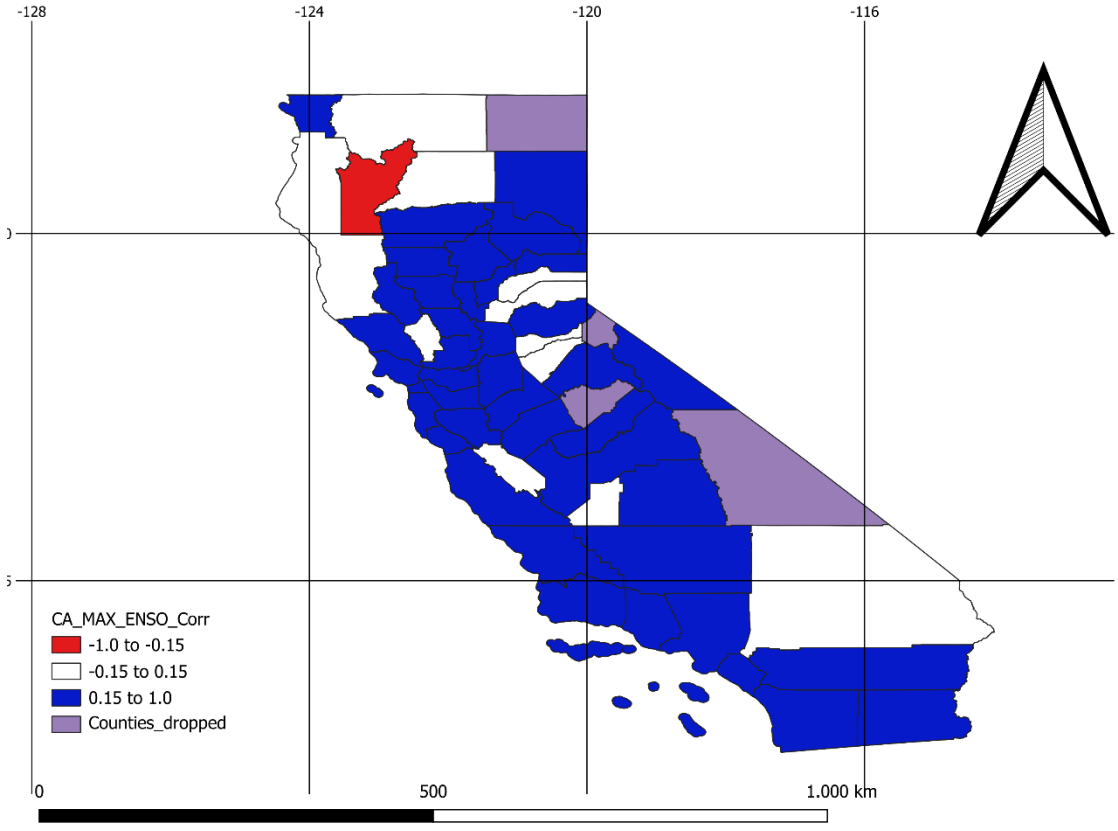


ΔΙΑΓΡΑΜΜΑ 7.11 ΔΙΑΓΡΑΜΜΑ ΚΑΤΑΝΟΜΗΣ ΣΥΣΧΕΤΙΣΗΣ ΔΕΙΚΤΗ ΕΛΑΧΙΣΤΟΥ ENSO (MIN_ENSO) ΣΕ ΟΛΕΣ ΤΙΣ ΠΟΛΙΤΕΙΕΣ ΤΩΝ Η.Π.Α.

Στα πλαίσια της πτυχιακής αυτής μελέτης επικεντρωθήκαμε στην Πολιτεία της Καλιφόρνια. Η Καλιφόρνια αποτελεί μια Πολιτεία που, όχι μόνο βρίσκεται στη 9η θέση (βλ. πίνακα 7.2) με τα περισσότερα αιτήματα, αλλά σημείωσε την υψηλότερη τιμή (0.35) συσχέτισης μεταξύ του μέγιστου δείκτη ENSO και των Αιτημάτων αποζημίωσης από κάθε άλλη Πολιτεία. Κατ' επέκταση η ανάλυση και η μελέτη της συγκεκριμένης Πολιτείας παρουσιάζει ιδιαίτερο ενδιαφέρον και ακολουθεί στη συνέχεια

ΠΙΝΑΚΑΣ 7.2- ΟΙ 10 ΠΟΛΙΤΕΙΕΣ ΤΩΝ ΗΝΩΜΕΝΩΝ ΠΟΛΙΤΕΙΩΝ ΜΕ ΤΟΝ ΥΨΗΛΟΤΕΡΟ ΑΡΙΘΜΟ ΑΙΤΗΣΕΩΝ ΑΠΟΖΗΜΙΩΣΗΣ ΛΟΓΩ ΠΛΗΜΜΥΡΩΝ, ΟΠΩΣ ΚΑΤΑΓΡΑΦΗΚΑΝ ΣΤΗ ΒΑΣΗ ΔΕΔΟΜΕΝΩΝ FEMA

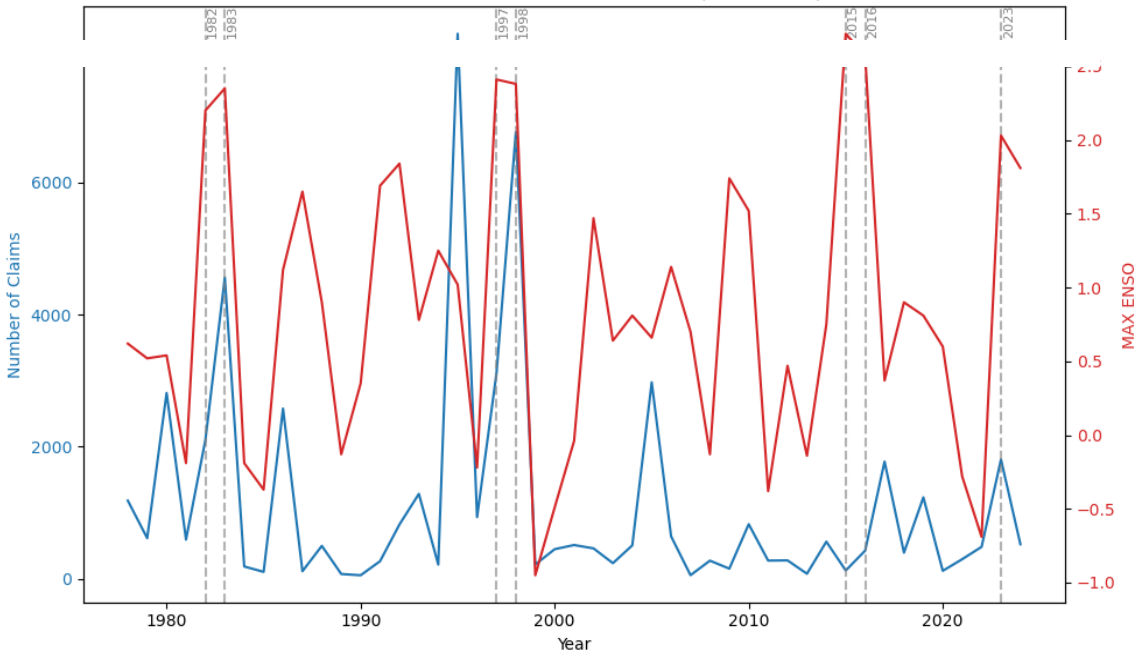
#	State	Number of Incidents
1	Louisiana	484,089
2	Florida	441,179
3	Texas	390,945
4	New Jersey	201,243
5	New York	174,828
6	North Carolina	108,558
7	Pennsylvania	76,666
8	Mississippi	64,133
9	California	53,107
10	Illinois	52455



ΕΙΚΟΝΑ 7.5: ΧΩΡΙΚΗ ΣΥΣΧΕΤΙΣΗ ΜΕΓΙΣΤΟΥ ΕΤΗΣΙΟΥ ΔΕΙΚΤΗ ENSO ΚΑΙ ΑΙΤΗΣΕΩΝ ΑΠΟΖΗΜΙΩΣΕΩΝ ΠΛΗΜΜΥΡΩΝ ΣΤΗΝ ΚΑΛΙΦΟΡΝΙΑ. (1980–2024)

Στην εικόνα 7.5 απεικονίζονται οι συσχετίσεις μεταξύ του μέγιστου ετήσιου δείκτη ENSO (El Niño–Southern Oscillation) και του αριθμού αιτημάτων αποζημιώσεων για

κάθε Κομητεία (county) της Πολιτείας της Καλιφόρνια. Εύκολα παρατηρεί κανείς ότι οι περισσότερες Κομητείες ξεπερνούν το άνω όριο του +0.15 (μπλε απόχρωση) ενώ όλες οι υπόλοιπες ανήκουν στην μέση κατηγορία (-0.15 έως 0.15). Εξαίρεση αποτελεί η Κομητεία Τρίνιτυ, στην οποία το μεγάλο πρόβλημα είναι οι χιονοπτώσεις και όχι οι καταστροφές λόγω ENSO, όπως επιβεβαιώνει η σχετική βιβλιογραφία (U.S. Geological Survey. (1998). Το παραπάνω εύρημα ενισχύει την άποψη, που διατυπώθηκε ανώτερα, σύμφωνα με την οποία ο αριθμός αιτημάτων αποζημίωσης δεν επηρεάζεται μονοσήμαντα (αν και σε μεγάλο βαθμό) από τα φαινόμενα ENSO αλλά και από πλήθος άλλων παραγόντων όπως τοπογραφικές και κλιματολογικές συνθήκες, γεωγραφική θέση κάθε περιοχής κ.α. Στην εικόνα 7.5 εμφανίζονται 4 Κομητείες (counties), οι οποίες εξαιρέθηκαν από τη διαδικασία, προς αποφυγή εσφαλμένων συμπερασμάτων, αφού εμφάνισαν πολύ μικρό δείγμα μετρήσεων αιτημάτων (κάτω από 10). Είναι οι Κομητείες Inyo, Mordoc, Mariposa και Alpine και διακρίνονται στο χάρτη με μωβ χρώμα.



ΔΙΑΓΡΑΜΜΑ 7.12: ΣΥΣΧΕΤΙΣΗ ΑΙΤΗΣΕΩΝ ΑΠΟΖΗΜΙΩΣΕΩΝ ΚΑΙ ΜΕΓΙΣΤΩΝ ΤΙΜΩΝ ΔΕΙΚΤΗ ENSO ΣΤΗΝ ΚΑΛΙΦΟΡΝΙΑ

Στο Διάγραμμα 7.12 απεικονίζεται η σχέση μεταξύ έντονων γεγονότων El Niño και του ετήσιου αριθμού αιτήσεων αποζημίωσης λόγω πλημμυρών στην Καλιφόρνια. Έυκολα διακρίνονται οι μέγιστες τιμές τόσο του δείκτη ENSO όσο και του αριθμού αιτήσεων κατά τα έτη 1982–83, 1997–98, 2015–16, και 2023, τα οποία είναι γνωστά για την ισχυρή εμφάνιση του φαινομένου El Niño. Η συσχέτιση ενισχύει την υπόθεση ότι οι κλιματικές αλλαγές μέσω του ENSO επηρεάζουν τη συχνότητα και σοβαρότητα των πλημμυρικών φαινομένων. Αν και δεν παρατηρείται πάντα αντίστοιχη αύξηση στις αιτήσεις για κάθε ENSO μέγιστο, πιθανώς λόγω τοπικών παραγόντων (π.χ. υποδομές, βροχοπτώσεις, εδαφολογικά χαρακτηριστικά), η γενική εικόνα υποδεικνύει σημαντική επίδραση.

Παρά το γεγονός ότι το φαινόμενο El Niño κορυφώθηκε έντονα την περίοδο 1997–98, ο αριθμός αιτήσεων αποζημίωσης στην Καλιφόρνια είχε ήδη φτάσει στο μέγιστο το 1995–96. Η περίοδος αυτή χαρακτηρίστηκε από έντονες ατμοσφαιρικές ροές και προκορεσμένα εδάφη, οδηγώντας σε σοβαρές πλημμύρες ανεξάρτητα από το ENSO (Dettinger, 1999). Αντίστοιχα, η περίοδος 2015–16, παρά την ισχυρή παρουσία του El Niño, δεν συνοδεύτηκε από αυξημένες βροχοπτώσεις στην Καλιφόρνια, μιας και η μετατόπιση του jet stream προς τον βορρά είχε ως αποτέλεσμα τα περισσότερα φαινόμενα να πλήξουν τις βορειότερες πολιτείες, αφήνοντας την Καλιφόρνια σχετικά ανεπηρέαστη, όπως αυτό αποτυπώνεται στο παραπάνω διάγραμμα (Swain et al,2016) . Συμπληρωματικά, αξίζει να γίνει μια αναφορά στην έρευνα του Murakami et al. (2025), που καταλήγει στο συμπέρασμα ότι τα φαινόμενα ENSO δεν αποτελούν αυτοτελώς επαρκή δείκτη κινδύνου, καθώς η παρουσία ξηρών μαζών αέρα (π.χ. σκόνη από τη Σαχάρα), η ατμοσφαιρική σταθερότητα, και άλλοι περιφερειακοί παράγοντες, μπορούν να αναστείλουν τη μεταφορά και ένταση των φαινομένων.

Το NOAA (National Oceanic and Atmospheric Administration, 2015) επιβεβαιώνει ότι η ένταση των φαινομένων ENSO δεν συνεπάγεται πάντοτε πλημμυρικές επιπτώσεις στην Καλιφόρνια, καθώς σημαντικό ρόλο παίζουν και άλλοι μετεωρολογικοί παράγοντες. Αξίζει φυσικά να σημειωθεί ότι η απόκτηση πολυετής εμπειρίας οδηγεί σε συνεχώς καλύτερη οργάνωση των αρμόδιων φορέων, που επιφέρει λιγότερες καταστροφές και κατ’ επέκταση αιτήματα για αποζημιώσεις.

7.4 Ανάλυση Συσχέτισης ENSO και Ακραίων Πλημμυρικών Φαινομένων

Στο πλαίσιο της παρούσας μελέτης, εξετάζεται η πιθανή συσχέτιση μεταξύ των φαινομένων ENSO και της εμφάνισης ακραίων πλημμυρικών γεγονότων. Για το σκοπό αυτό μελετήθηκαν οι χρονοσειρές παροχών της βάσης δεδομένων US-CAMELS (Newman et al., 2014) μέσω της μεθόδου υπερβάσης υπέρ κατωφλιού (peak-over-threshold method). Ακριβέστερα, για κάθε έναν από τους εξεταζόμενους σταθμούς μέτρησης, αντλήθηκε το συλλογικό ρίσκο S (Papoulakos, 2025), αλλά και το πλήθος υπερβάσεων. Αναλυτικότερα, υπολογίστηκαν οι συντελεστές συσχέτισης Pearson, μεταξύ δύο βασικών μεταβλητών πλημμυρικού κινδύνου όπως αυτές προέκυψαν από τις χρονοσειρές παροχής (το πλήθος υπερβάσεων κατωφλιού (N) και το συλλογικό ρίσκο (collective Risk)) και τριών βασικών δεικτών ποσοτικοποίησης των φαινομένων ENSO (η ετήσια μέση τιμή, η ετήσια μέγιστη και η ετήσια ελάχιστη τιμή). Η ανάλυση πραγματοποιήθηκε για τέσσερα διαδοχικά κατώφλια ποσοστιαίας τάξης παροχής στους μελετώμενους σταθμούς (90%, 95%, 98%, 99%), τα οποία αντιστοιχούν σε αυξανόμενα επίπεδα έντασης των καταγεγραμμένων φαινομένων.

ΠΙΝΑΚΑΣ 7.3- ΣΥΓΚΕΝΤΡΩΤΙΚΟΣ ΠΙΝΑΚΑΣ ΣΥΝΤΕΛΕΣΤΩΝ ΣΥΣΧΕΤΙΣΗΣ PEARSON ΜΕΤΑΞΥ ENSO ΔΕΙΚΤΩΝ (ΜΕΣΗ, ΜΕΓΙΣΤΗ ΚΑΙ ΕΛΑΧΙΣΤΗ ΤΙΜΗ) ΚΑΙ ΔΥΟ ΜΕΤΑΒΛΗΤΩΝ ΠΛΗΜΜΥΡΙΚΟΥ ΚΙΝΔΥΝΟΥ (ΠΛΗΘΟΣ ΥΠΕΡΒΑΣΕΩΝ ΠΑΡΟΧΗΣ ΚΑΙ ΣΥΛΛΟΓΙΚΟ ΡΙΣΚΟ ΥΠΕΡΒΑΣΕΩΝ ΠΑΡΟΧΗΣ), ΥΠΟΛΟΓΙΣΜΕΝΩΝ ΓΙΑ ΚΑΤΩΦΛΙΑ 90%, 95%, 98% ΚΑΙ 99%.

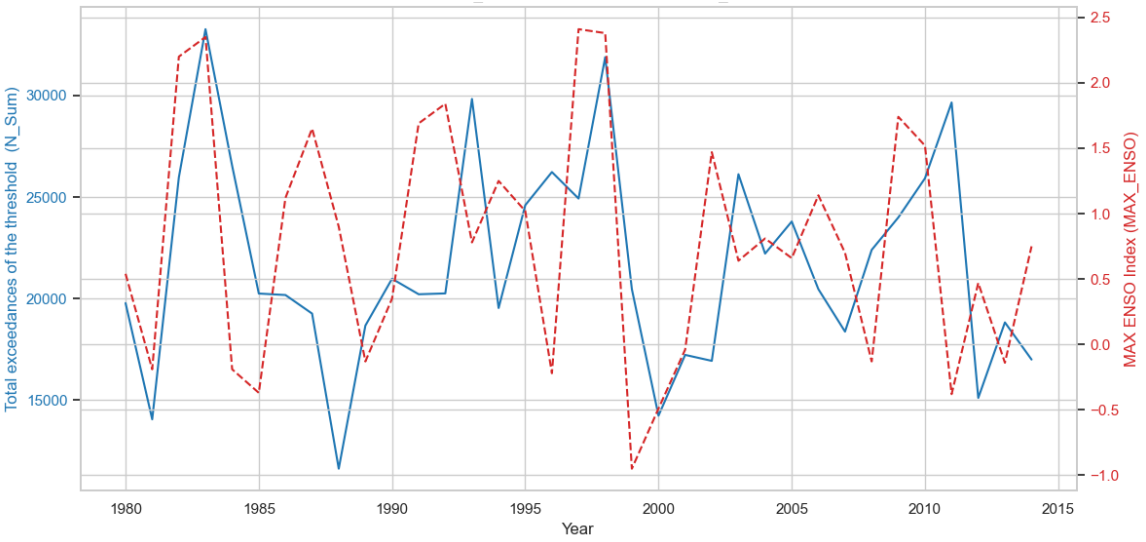
Κατώφλι Υπέρβασης	Μεταβλητή	Μεταβλητή ENSO	Συντελεστής Pearson
90	N_Sum	MAX_ENSO	0,3374
95	N_Sum	MAX_ENSO	0,3237
98	N_Sum	MAX_ENSO	0,2851
99	Coll_Risk	MAX_ENSO	0,2523

98	Coll_Risk	MAX_ENSO	0,2468
95	Coll_Risk	MAX_ENSO	0,2450
99	N_Sum	MAX_ENSO	0,2445
90	Coll_Risk	MAX_ENSO	0,2422
99	Coll_Risk	ENSO_avg	0,2026
95	Coll_Risk	ENSO_avg	0,1570
99	N_Sum	ENSO_avg	0,1501
90	Coll_Risk	ENSO_avg	0,1375
90	Coll_Risk	MIN_ENSO	-0,0582
90	N_Sum	MIN_ENSO	-0,0376
95	N_Sum	MIN_ENSO	-0,0365
95	Coll_Risk	MIN_ENSO	-0,0324
99	Coll_Risk	MIN_ENSO	0,0323
99	N_Sum	MIN_ENSO	-0,0187
98	N_Sum	MIN_ENSO	-0,0174

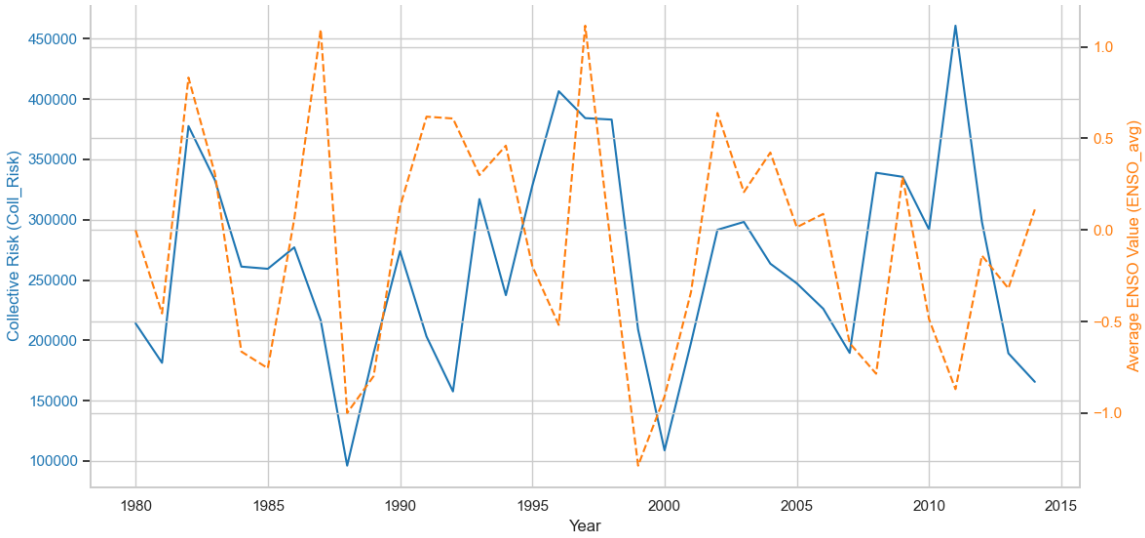
Αναλύοντας τον πίνακα 7.3. προκύπτουν τα εξής συμπεράσματα Οι ισχυρότερες θετικές συσχετίσεις παρατηρούνται μεταξύ της μέγιστης ετήσιας τιμής του ENSO (MAX_ENSO) και των μεταβλητών πλήθος γεγονότων και συλλογικό ρίσκο. Συγκεκριμένα, η μέγιστη συσχέτιση παρατηρείται για πλήθος γεγονότων (N_Sum) στο κατώφλι 90% (r = 0,3374), ενώ παρόμοια υψηλές τιμές καταγράφονται και για τα υπόλοιπα κατώφλια και δείκτες κινδύνου.

Η μέση ετήσια τιμή του ENSO (ENSO_avg) εμφανίζει μετρίως θετικές συσχετίσεις, με τις υψηλότερες τιμές να καταγράφονται για τη μεταβλητή του συλλογικού ρίσκου στο κατώφλι 99% (r = 0,2026) και του πλήθους γεγονότων στο κατώφλι 90% (r = 0,1874). Αντίθετα, ο δείκτης της ελάχιστης ετήσιας τιμής (MIN_ENSO), που αποτυπώνει τις αρνητικές φάσεις του φαινομένου, εμφανίζει χαμηλές ή αμελητέες συσχετίσεις, με τιμές κοντά στο μηδέν ή αρνητικές.

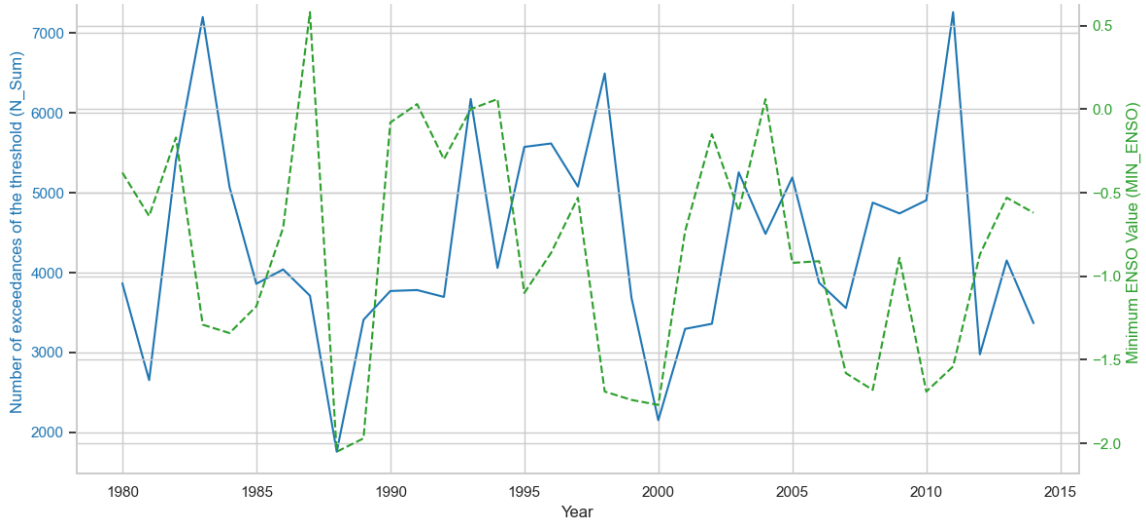
Ενδεικτικά αποτυπώνονται 3 διαγράμματα, τα οποία απεικονίζουν για κάθε μια από τις 3 μεταβλητές ENSO τη σχέση με τη μεγαλύτερη συσχέτιση, όπως αυτές αναφέρονται στον παραπάνω πίνακα 7.3



ΔΙΑΓΡΑΜΜΑ 7.13 ΔΙΑΓΡΑΜΜΑ ΧΡΟΝΟΣΕΙΡΑΣ ΤΩΝ ΥΠΕΡΒΑΣΕΩΝ (N_SUM) ΓΙΑ ΚΑΤΩΦΛΙ ΥΠΕΡΒΑΣΗΣ 90% ΣΥΓΚΡΙΤΙΚΑ ΜΕ ΤΙΣ ΜΕΓΙΣΤΕΣ ΤΙΜΕΣ ΤΟΥ ΔΕΙΚΤΗ ENSO (MAX_ENSO)

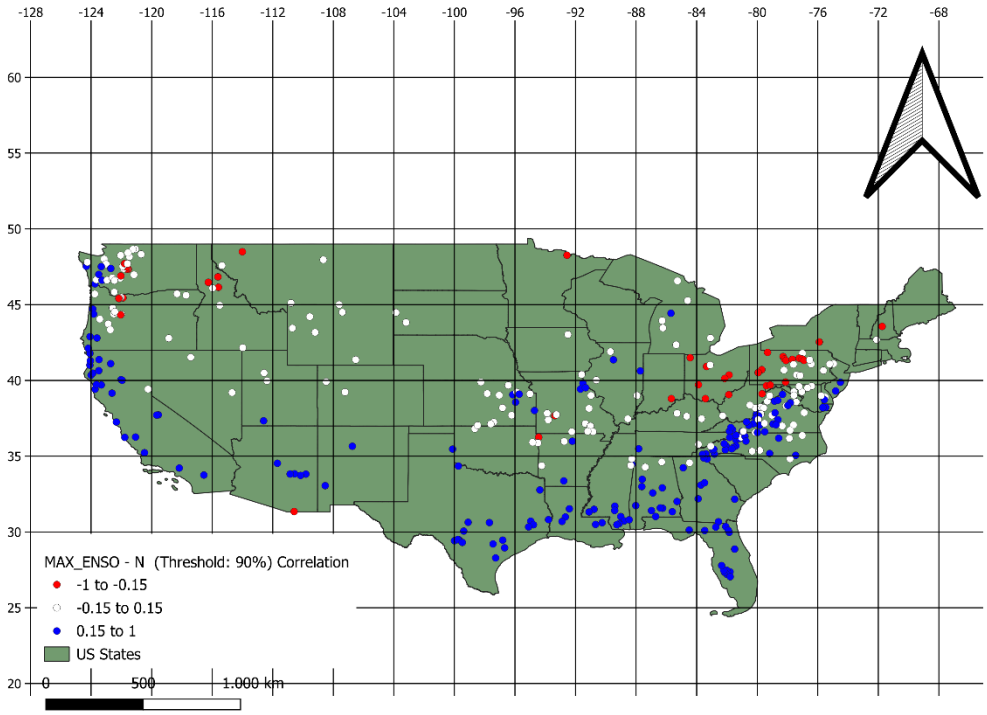


ΔΙΑΓΡΑΜΜΑ 7.14 ΔΙΑΓΡΑΜΜΑ ΧΡΟΝΟΣΕΙΡΑΣ ΤΟΥ ΣΥΛΛΟΓΙΚΟΥ ΡΙΣΚΟΥ (COLL_RISK) ΓΙΑ ΚΑΤΩΦΛΙ ΥΠΕΡΒΑΣΗΣ 99% ΣΥΓΚΡΙΤΙΚΑ ΜΕ ΤΙΣ ΜΕΣΕΣ ΤΙΜΕΣ ΤΟΥ ΔΕΙΚΤΗ ENSO (ENSO_AVG)

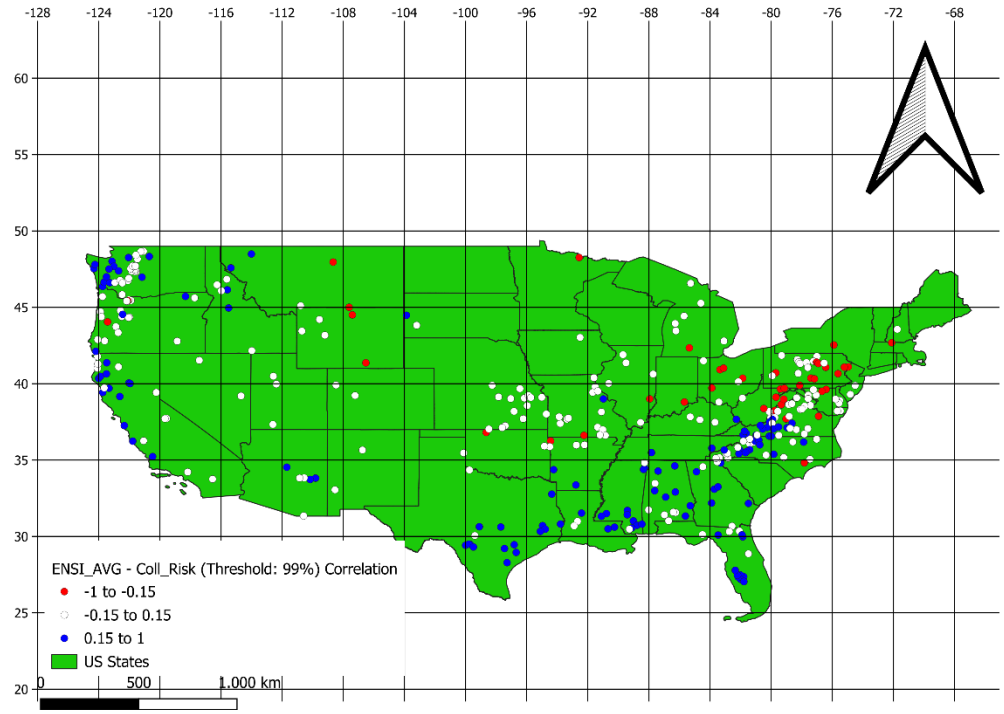


ΔΙΑΓΡΑΜΜΑ 7.15- ΔΙΑΓΡΑΜΜΑ ΧΡΟΝΟΣΕΙΡΑΣ ΤΩΝ ΥΠΕΡΒΑΣΕΩΝ (N_SUM) ΓΙΑ ΚΑΤΩΦΛΙ ΥΠΕΡΒΑΣΗΣ 98% ΣΥΓΚΡΙΤΙΚΑ ΜΕ ΤΙΣ ΕΛΑΧΙΣΤΕΣ ΤΙΜΕΣ ΤΟΥ ΔΕΙΚΤΗ ENSO (MIN_ENSO)

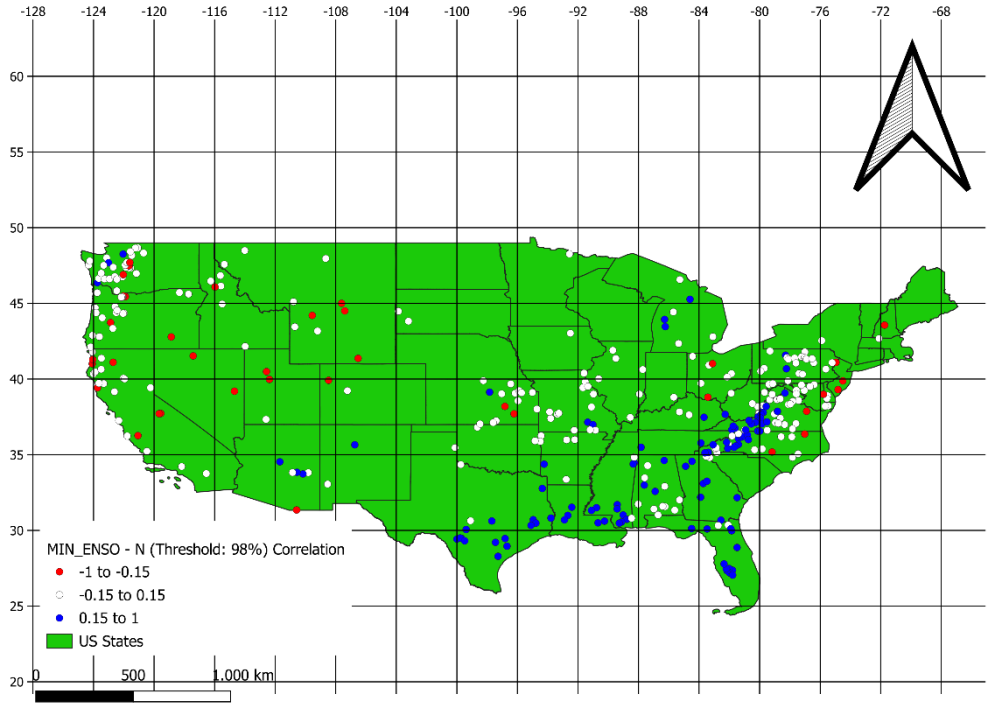
Εκτός από την κατανομή στο χρόνο όμως, μεγάλο ενδιαφέρον παρουσιάζει η χωρική κατανομή των στοιχείων παροχών. Αναλυτικότερα, στους παρακάτω πίνακες , για κάθε μία από τις παραπάνω χρονοσειρές παρουσιάζεται η κατανομή των συσχετίσεων των στοιχείων υπέρβασης παροχής κάθε σταθμού (N, Coll_Risk) με τους δείκτες ENSO (MAX_ENSO, MIN_ENSO, ENSO_AVG) για καθένα από τους 360 μελετώμενους σταθμούς διακεκριμένους σε 3 κατηγορίες ανάλογα με την τιμή της συσχέτισης.



ΕΙΚΟΝΑ 7.6 ΧΑΡΤΗΣ ΒΑΘΜΟΥ ΣΥΣΧΕΤΙΣΗΣ ΤΩΝ ΥΠΕΡΒΑΣΕΩΝ (N_SUM) ΓΙΑ ΚΑΤΩΦΛΙ ΥΠΕΡΒΑΣΗΣ 90% ΣΥΓΚΡΙΤΙΚΑ ΜΕ ΤΙΣ ΜΕΓΙΣΤΕΣ ΤΙΜΕΣ ΤΟΥ ΔΕΙΚΤΗ ENSO (MAX_ENSO) ΑΝΑ ΣΤΑΘΜΟ

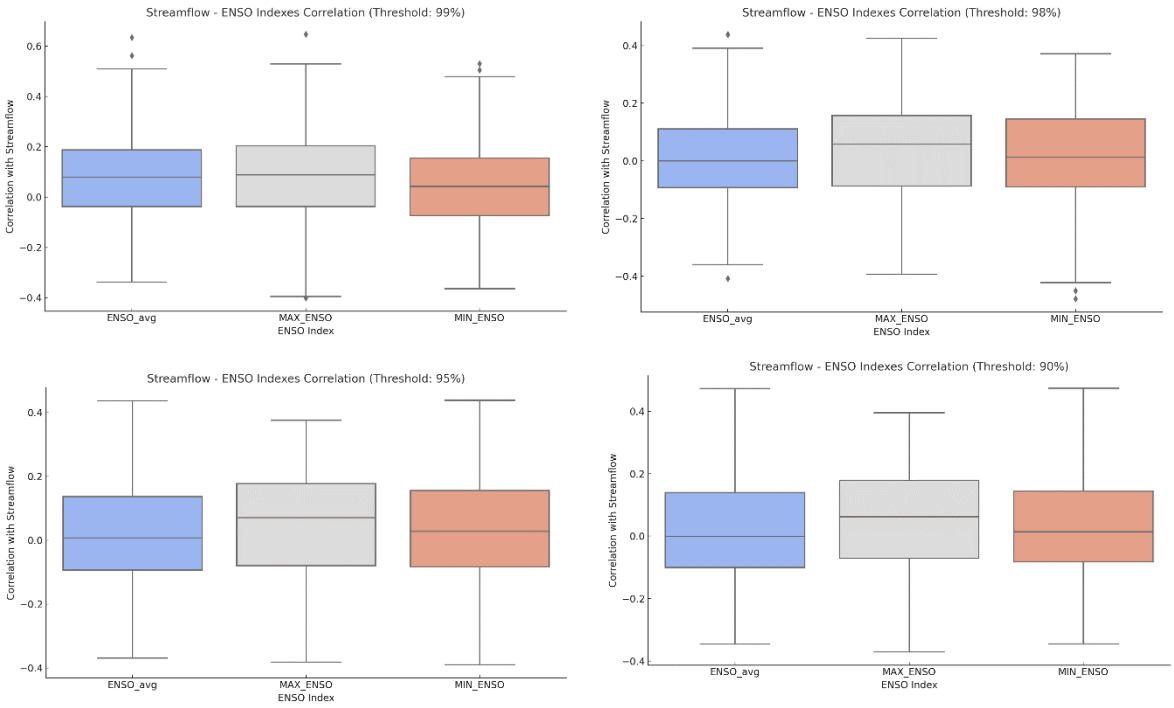


ΕΙΚΟΝΑ 7.7 ΧΑΡΤΗΣ ΒΑΘΜΟΥ ΣΥΣΧΕΤΙΣΗΣ ΤΟΥ ΣΥΛΛΟΓΙΚΟΥ ΡΙΣΚΟΥ (COLL_RISK) ΓΙΑ ΚΑΤΩΦΛΙ ΥΠΕΡΒΑΣΗΣ 99% ΣΥΓΚΡΙΤΙΚΑ ΜΕ ΤΙΣ ΜΕΣΕΣ ΤΙΜΕΣ ΤΟΥ ΔΕΙΚΤΗ ENSO (ENSO_AVG) ΑΝΑ ΣΤΑΘΜΟ



ΕΙΚΟΝΑ 7.8 – ΧΑΡΤΗΣ ΒΑΘΜΟΥ ΣΥΣΧΕΤΙΣΗΣ ΤΩΝ ΥΠΕΡΒΑΣΕΩΝ (N_SUM) ΓΙΑ ΚΑΤΩΦΛΙ ΥΠΕΡΒΑΣΗΣ 98% ΣΥΓΚΡΙΤΙΚΑ ΜΕ ΤΙΣ ΕΛΑΧΙΣΤΕΣ ΤΙΜΕΣ ΤΟΥ ΔΕΙΚΤΗ ENSO (MIN_ENSO) ΑΝΑ ΣΤΑΘΜΟ

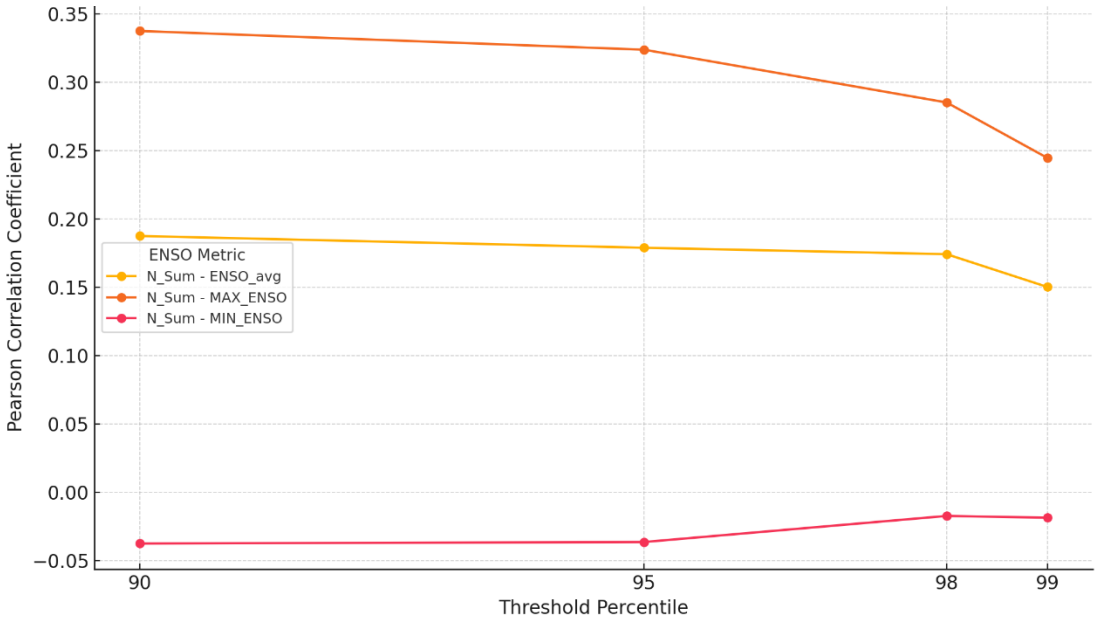
Παρατηρώντας και τους 3 χάρτες γίνεται εύκολα αντιληπτό ότι οι περισσότερες παράκτιες περιοχές εμφανίζουν μεγαλύτερες συσχετίσεις με τους δείκτες ENSO . Φυσικά κάτι τέτοιο είναι απολύτως δικαιολογημένο και πλήρως εναρμονισμένο με την ισχύουσα βιβλιογραφία. Στον πρώτο χάρτη μάλιστα, που αφορά τη συσχέτιση με το δείκτη μέγιστων ετήσιων τιμών ENSO (MAX_ENSO) και αφορά τα φαινόμενα El Nino είναι πολύ παραστατική εικόνα με τη συντριπτική πλειοψηφία των παράκτιων σταθμών να παρουσιάζουν μεγαλύτερο βαθμό συσχέτισης από 0.15. Συγκεκριμένα 171 από τους 360 σταθμούς εμφανίζουν σχετικό βαθμό συσχέτισης πάνω από 0.15, εύρημα, το οποίο αποδεικνύει την παραπάνω υπόθεση. Με σκοπό καλύτερη κατανόηση των παραπάνω αποτελεσμάτων και την ποσοτικοποίηση τους για μια πιο ξεκάθαρη εικόνα παρακάτω παρουσιάζεται ο πίνακας με τις τιμές και τα ποσοστά των συσχετίσεων καθώς και το αντίστοιχο διάγραμμα Boxplot 7.18



ΔΙΑΓΡΑΜΜΑ BOXPLOT 7.16 ΔΙΑΣΠΟΡΑ ΤΩΝ ΣΥΣΧΕΤΙΣΕΩΝ ΜΕΤΑΞΥ ΔΕΙΚΤΩΝ ENSO (ΜΕΣΟΣ ΟΡΟΣ, ΜΕΓΙΣΤΟ ΚΑΙ ΕΛΑΧΙΣΤΟ) ΚΑΙ ΤΩΝ ΔΙΑΦΟΡΕΤΙΚΩΝ ΔΕΙΚΤΩΝ ΥΠΕΡΒΑΣΗΣ ΟΡΙΩΝ ΠΑΡΟΧΗΣ (ΚΑΤΩΦΛΙΑ 90, 95, 98, 99%)

ΠΙΝΑΚΑΣ 7.4 ΚΑΤΑΝΟΜΗ ΤΩΝ ΣΤΑΘΜΩΝ ΣΕ ΤΡΕΙΣ ΚΑΤΗΓΟΡΙΕΣ ΤΙΜΩΝ ΣΥΣΧΕΤΙΣΗΣ ΔΕΙΚΤΩΝ ΥΠΕΡΒΑΣΗΣ ΟΡΙΩΝ ΠΑΡΟΧΗΣ ΜΕ ΔΕΙΚΤΕΣ ENSO ΓΙΑ ΤΑ ΜΕΛΕΤΟΥΜΕΝΑ ΣΕΝΑΡΙΑ

Category	Corr. Value Range	MIN_ENSO Count	MIN_ENSO Percentage	MAX_ENSO Count	MAX_ENSO Percentage	ENSO_avg Count	ENSO_avg Percentage
1	-1 to -0.15	37	10.3	40	11.1	41	11.4
2	-0.15 to 0.15	219	60.8	149	41.4	200	55.5
3	0.15 to 1	104	28.9	171	47.5	119	33.1
Sums		360	100	360	100	360	100



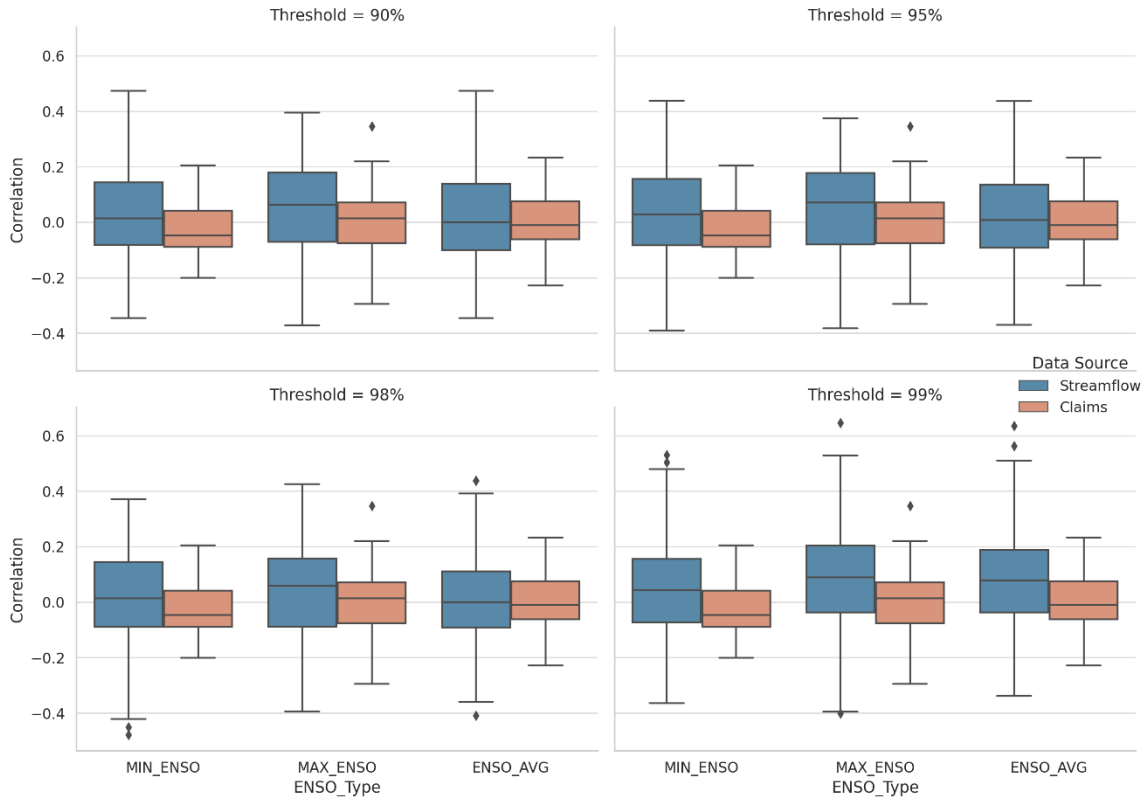
ΔΙΑΓΡΑΜΜΑ 7.17 ΜΕΤΑΒΟΛΗ ΤΟΥ ΣΥΝΤΕΛΕΣΤΗ ΣΥΣΧΕΤΙΣΗΣ PEARSON ΜΕΤΑΞΥ ΜΕΤΑΒΛΗΤΩΝ ΠΛΗΜΜΥΡΙΚΟΥ ΚΙΝΔΥΝΟΥ (ΑΡΙΘΜΟΣ ΣΥΜΒΑΝΤΩΝ) ΚΑΙ ΔΕΙΚΤΩΝ ENSO (ΜΕΣΟΣ ΟΡΟΣ, ΜΕΓΙΣΤΟ, ΕΛΑΧΙΣΤΟ), ΣΕ ΔΙΑΦΟΡΕΤΙΚΑ ΚΑΤΩΦΛΙΑ ΑΚΡΑΙΩΝ ΤΙΜΩΝ (90%, 95%, 98%, 99%)

Το Διάγραμμα 7.17. αποτυπώνει τη μεταβολή της συσχέτισης (Pearson) μεταξύ των 3 διαφορετικών δεικτών ENSO και της προαναφερθείσας πλημμυρικής μεταβλητής σε αυξανόμενα κατώφλια ακραίων τιμών (το πλήθος υπερβάσεων (N)). Παρατηρείται ότι η αύξηση του κατωφλιού (δηλ. επικεντρωνόμαστε σε πιο ακραίες πλημμύρες), επιφέρει μείωση της συσχέτισης της μεταβλητής του πλήθους υπερβάσεων κατωφλιού σε σχέση με όλες τις τιμές δεικτών ENSO. Οποιαδήποτε συσχέτιση με την ετήσια ελάχιστη τιμή ENSO εμφανίζει αρνητικές ή αμελητέες τιμές, για κάθε κατώφλι υπέρβασης.

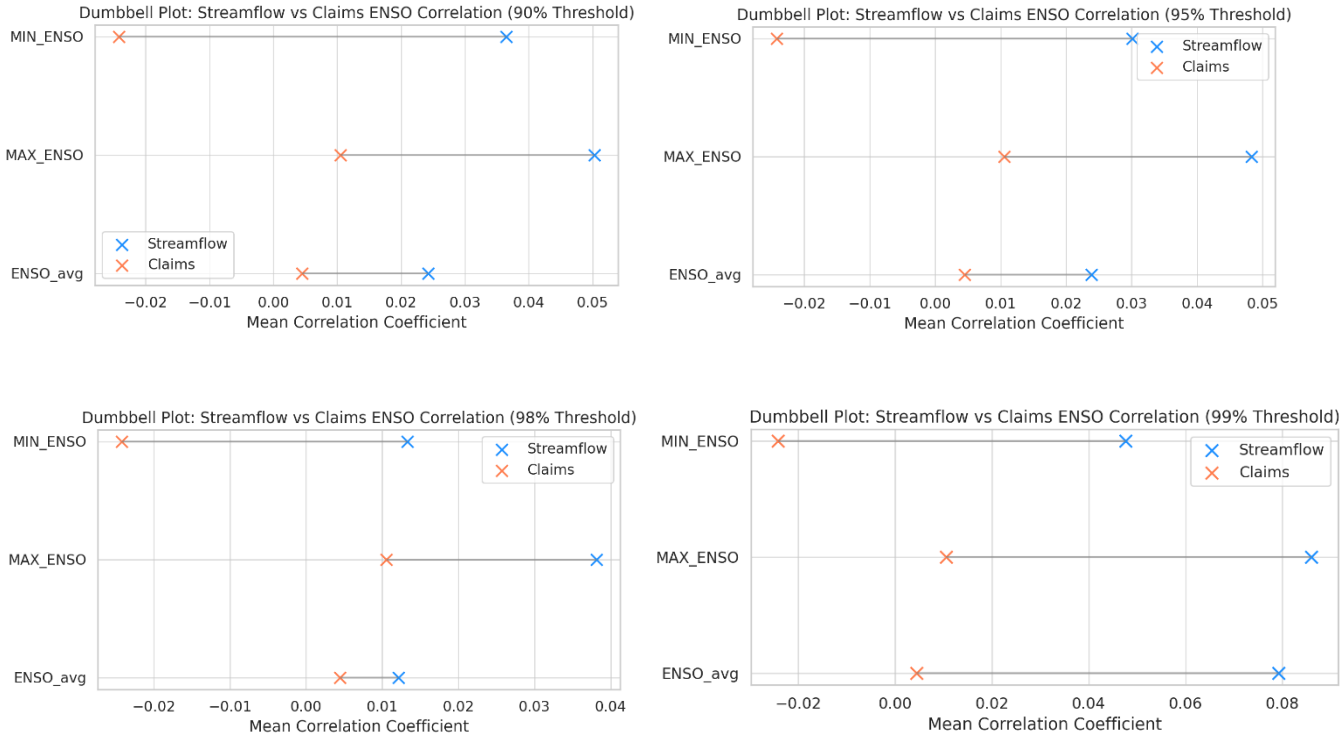
Τα παραπάνω αποτελέσματα συμφωνούν με προηγούμενες μελέτες που τεκμηριώνουν σημαντική επίδραση φαινομένων ENSO στις πλημμυρικές καταστροφές (Ward et al., 2014). Η χρήση κατωφλίων ποσοστιαίας τάξης για τον εντοπισμό ακραίων υδρολογικών φαινομένων υποστηρίζεται από τη βιβλιογραφία για την ανάλυση ακραίων τιμών (Coles, 2001)

7.5 Σύγκριση Συσχετίσεων δεικτών ENSO με Αιτήματα Αποζημίωσης και Δεικτών Υπέρβασης Ορίων παροχής

Σε αυτό το στάδιο επεξεργασίας των δεδομένων μας, έχοντας καταλήξει από τις προηγούμενες ενότητες σε ποσοτικοποιήσεις των συσχετίσεων βρισκόμαστε σε θέση να συγκρίνουμε τα αποτελέσματα μας.



ΔΙΑΓΡΑΜΜΑ 7.18 – ΣΥΓΚΡΙΣΗ ΤΩΝ ΔΙΑΓΡΑΜΜΑΤΩΝ BOXPLOT ΣΥΣΧΕΤΙΣΕΩΝ ΜΕΤΑΞΥ ΔΕΙΚΤΩΝ ENSO (ΜΕΣΟΣ ΟΡΟΣ, ΜΕΓΙΣΤΟ, ΕΛΑΧΙΣΤΟ), ΑΙΤΗΜΑΤΩΝ ΑΠΟΖΗΜΙΩΣΗΣ ΚΑΙ ΠΛΗΘΟΣ ΥΠΕΡΒΑΣΗΣ ΟΡΙΟΥ ΠΑΡΟΧΗΣ ΓΙΑ ΔΙΑΦΟΡΕΤΙΚΑ ΚΑΤΩΦΛΙΑ (90, 95, 98, 99%)



ΔΙΑΓΡΑΜΜΑ 7.19 – ΣΥΓΚΡΙΣΗ ΤΩΝ ΔΙΑΓΡΑΜΜΑΤΩΝ DUMBBELL ΣΥΣΧΕΤΙΣΕΩΝ ΜΕΤΑΞΥ ΔΕΙΚΤΩΝ ENSO (ΜΕΣΟΣ ΟΡΟΣ, ΜΕΓΙΣΤΟ, ΕΛΑΧΙΣΤΟ), ΑΙΤΗΜΑΤΩΝ ΑΠΟΖΗΜΙΩΣΗΣ ΚΑΙ ΠΛΗΘΟΣ ΥΠΕΡΒΑΣΗΣ ΟΡΙΟΥ ΠΑΡΟΧΗΣ ΓΙΑ ΔΙΑΦΟΡΕΤΙΚΑ ΚΑΤΩΦΛΙΑ (90, 95, 98, 99%)

Στο πλαίσιο της παρούσας μελέτης, συγκρίναμε τις συσχετίσεις μεταξύ τριών δεικτών ENSO (MIN_ENSO, MAX_ENSO, ENSO_avg) με δύο διαφορετικές μεταβλητές: τις μετρήσεις παροχής υδάτων (streamflow) και τα ασφαλιστικά αιτήματα λόγω πλημμυρών (claims), αντλώντας στοιχεία από 360 υδρολογικούς σταθμούς

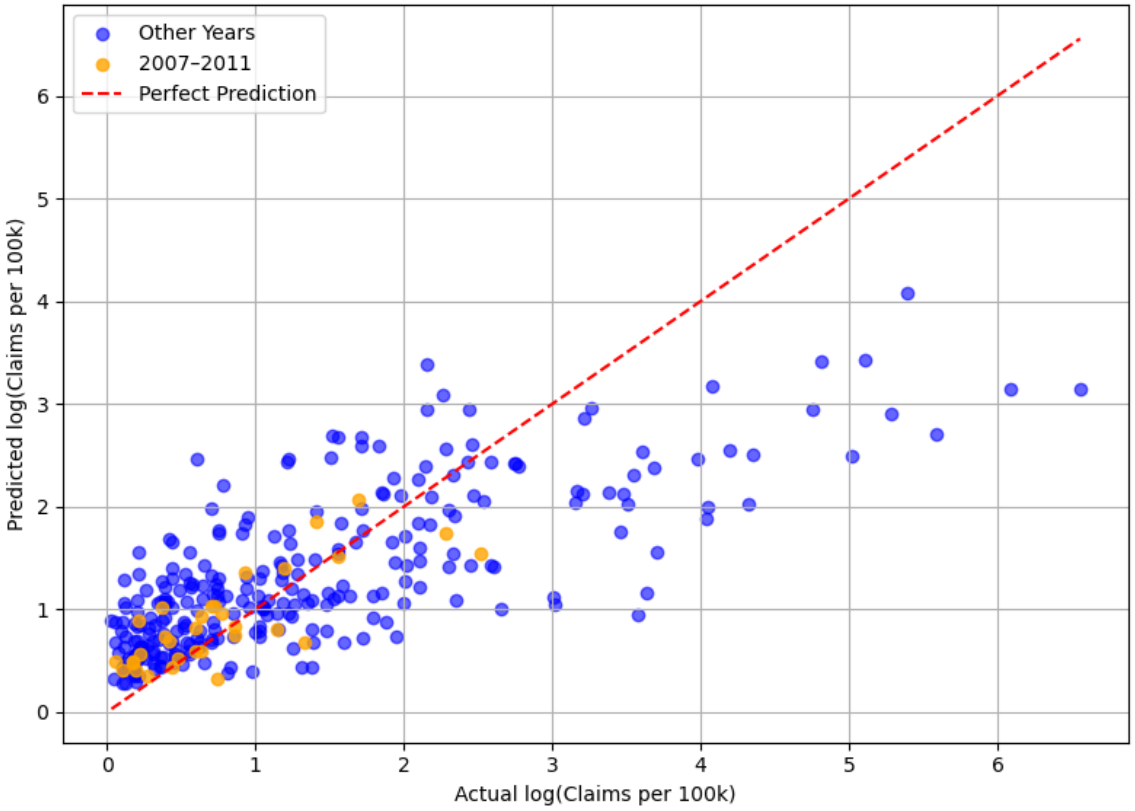
Οι συσχετίσεις με τα δεδομένα παροχής έχουν κατά κύριο λόγο μεγαλύτερες τιμές από τις αντίστοιχες με τα Αιτήματα Αποζημίωσης, ιδίως για τον δείκτη MAX_ENSO, υποδηλώνοντας ότι η επίδραση των φαινομένων ENSO εκδηλώνεται ευκολότερα στα δεδομένα της παροχής.

Οι τιμές των συσχετίσεων, που προκύπτουν από τα ασφαλιστικά αιτήματα μικρότερες, γεγονός που οφείλεται σε εξωτερικούς κοινωνικοοικονομικούς παράγοντες ή στην διαφοροποίηση της τοπικής πολιτικής ασφαλίσεων.

Κατά γενική ομολογία προκύπτει ότι η επίδραση του ENSO είναι ευκολότερα ανιχνεύσιμη στα υδρολογικά δεδομένα παρά στα καταγεγραμμένα οικονομικά δεδομένα των ζημιών.

7.6 Αποτελέσματα Μοντέλου Μηχανικής Μάθησης

Το τελικό μοντέλο Μηχανικής Μάθησης διαμορφώθηκε με βάση το πρότυπο του μοντέλου CatBoost regression. Για την αξιολόγηση της απόδοσης του, δημιουργήθηκε ένα διάγραμμα διασποράς με σκοπό τη σύγκριση των προβλεπόμενων και των πραγματικών τιμών, χρησιμοποιώντας ως μεταβλητή στόχο τις λογαριθμισμένες αιτήσεις ανά 100.000 κατοίκους (Διάγραμμα 7.20). Η κόκκινη διακεκομμένη γραμμή εκφράζει την ευθεία 1:1, που θα αντιπροσώπευε το ιδανικό μοντέλο και σχεδιάζεται για αντιπαραβολή με τα προβληθέντα αποτελέσματα του δικού μας μοντέλου. Προβάλλονται με διαφορετικό χρώμα οι προβλέψεις για τα έτη 2007-2011, που είναι η πενταετία, που διαλέξαμε σαν στόχο (για λόγους παρουσίασης στα πλαίσια της διπλωματικής μας εργασίας, το μοντέλο μπορεί να λειτουργήσει για οποιαδήποτε επιθυμητή χρονική περίοδο – στόχο) Οι προβλέψεις του μοντέλου σημειώνουν θετική τάση και συγκρατημένα καλή ευθυγράμμιση με τις πραγματικές τιμές, αν και παρατηρείται μια ελαφρά υποεκτίμηση στην περιοχή των υψηλών τιμών. Το διαμορφωθέν μοντέλο χαρακτηρίζεται από συντελεστή προσδιορισμού $R^2 = 0.638$, τιμή αρκετά ικανοποιητική.



ΔΙΑΓΡΑΜΜΑ 7.20 ΔΙΑΓΡΑΜΜΑ ΔΙΑΣΠΟΡΑΣ ΜΕΤΑΞΥ ΠΡΟΒΛΕΠΟΜΕΝΩΝ ΚΑΙ ΠΡΑΓΜΑΤΙΚΩΝ ΤΙΜΩΝ ΣΕ ΛΟΓΑΡΙΘΜΙΚΗ ΚΛΙΜΑΚΑ (LOG) ΤΩΝ ΑΙΤΗΣΕΩΝ ΑΝΑ 100.000 ΚΑΤΟΙΚΟΥΣ ΓΙΑ ΤΗΝ ΚΑΛΙΦΟΡΝΙΑ.

ΠΙΝΑΚΑΣ 7.5 - ΣΥΓΚΡΙΣΗ ΜΕΤΑΞΥ ΠΡΑΓΜΑΤΙΚΩΝ ΚΑΙ ΠΡΟΒΛΕΠΟΜΕΝΩΝ ΑΙΤΗΣΕΩΝ ΑΣΦΑΛΙΣΗΣ ΠΛΗΜΜΥΡΑΣ ΓΙΑ ΕΠΙΛΕΓΜΕΝΑ ΕΤΗ. ΟΙ ΤΙΜΕΣ ΠΕΡΙΛΑΜΒΑΝΟΥΝ ΤΟΣΟ ΤΙΣ ΑΙΤΗΣΕΙΣ ΑΝΑ 100.000 ΚΑΤΟΙΚΟΥΣ ΟΣΟ ΚΑΙ ΤΟΝ ΣΥΝΟΛΙΚΟ ΑΡΙΘΜΟ ΑΙΤΗΣΕΩΝ.

yearOfLoss	actual_per_100k	predicted_per_100k	actual_claims	predicted_claims
2007	0.61	1.67	54.0	220.96
2008	1.91	1.46	274.0	231.88
2009	0.85	1.88	153.0	334.07
2010	2.2	2.47	826.0	400.02
2011	2.13	1.35	275.0	246.77

Στον παραπάνω πίνακα 7.5. παρουσιάζεται μια ενδεικτική σύγκριση μεταξύ των πραγματικών και των προβλεπόμενων Αιτημάτων ασφάλισης έναντι πλημμυρικών καταστροφών για την πενταετία 2007-2011. Το μοντέλο μπορεί να ρυθμιστεί, ώστε

να προβλέψει Αιτήματα αποζημιώσεων για οποιαδήποτε χρονική περίοδο (με την εισαγωγή των απαραιτήτων μεταβλητών φυσικά). Στην περίπτωση μας, το χρονικό διάστημα-στόχος είναι η πενταετία αυτή (2007-2011) για λόγους παρουσίασης αποτελεσμάτων. Έχει επιλεχθεί ο τρόπος προβολής των τιμών τόσο σε κανονικοποιημένο επίπεδο (ανά 100.000 κατοίκους), όσο και ως απόλυτοι αριθμοί αιτημάτων με στόχο την καλύτερη εποπτεία του αναγνώστη. Οι παρατηρούμενες αποκλίσεις ανάμεσα στις τιμές είναι οι αναμενόμενες για ένα μοντέλο με συντελεστή προσδιορισμού $R^2 = 0.638$.

Η βάση δεδομένων μας έχει δυναμική 1572 καταχωρήσεων. Στο παράδειγμα μας, προκείμενου να εκπαιδεύσουμε με τον καλύτερο δυνατό τρόπο το διαμορφωθέν μοντέλο το τροποποιήσαμε με ένα πλήθος 1418 αρχικών δεδομένων, που περιλαμβάνουν όλα τα αιτήματα, που σημειώθηκαν εντός των συνόρων της Πολιτείας της Καλιφόρνια από το έτος 1978 έως το 2024 (με εξαίρεση τα στοιχεία-αιτήματα της πενταετίας-στόχου 2007-2011). Περισσότερα για τα χαρακτηριστικά του μοντέλου και τη διαδικασία μοντελοποίησης έχουν ήδη αναλυθεί στις προηγούμενες ενότητες (4.2 Εξαγωγή Χαρακτηριστικών (Feature Engineering) και Μοντελοποίηση με Μηχανική Μάθηση (Machine Learning)) αλλά έχει ενδιαφέρον ένας ενδεικτικός πίνακας των στοιχείων εισαγωγής με τα οποία εκπαιδεύτηκε το μοντέλο:

ΠΙΝΑΚΑΣ 7.6 – ΑΠΟΣΠΑΣΜΑ ΤΥΧΑΙΩΝ ΣΗΜΕΙΩΝ ΕΚ ΤΩΝ ΔΕΔΟΜΕΝΩΝ ΕΙΣΑΓΩΓΗΣ ΚΑΙ ΕΚΠΑΙΔΕΥΣΗΣ ΤΟΥ ΜΟΝΤΕΛΟΥ

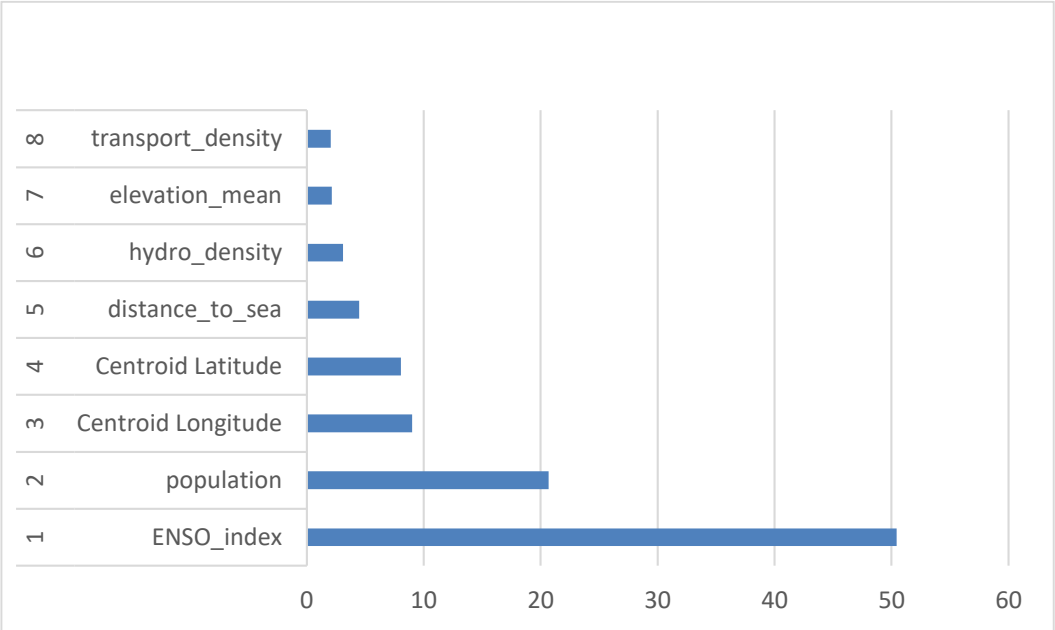
X												Y
county Code	county Name	yearOf Loss	state	hydro density	transport density	ENSO_avg	distance_km	INTPTLAT	INTPTLON	population	elev_mean	Num_Claim
6019	Fresno	1995	CA	2,7820	1,4960	-0,1975	184,5314	36,7610	-119,6550	1017162	184,3267	68
6055	Napa	1986	CA	2,8385	1,0301	-0,8692	39,6856	38,5071	-122,3259	133216	191,6744	164
6061	Placer	2000	CA	3,2502	2,4328	-0,9092	131,6999	39,0620	-120,7227	423561	213,4431	10
6081	San Mateo	2016	CA	3,5176	3,9410	0,3325	7,3819	37,4147	-122,3715	726353	87,1748	20
6005	Amador	2017	CA	4,0254	1,6884	-0,2000	85,1437	38,4435	-120,6539	41811	220,5851	10

Κατά την τελική μορφή του μοντέλου αποτιμήθηκε και η τελική λίστα βαρύτητας/σπουδαιότητα μεταβλητών/χαρακτηριστικών (feature importance) μέσω PredictionValuesChange Αναλυτικότερα, μέσα από τις τιμές αυτές υποδεικνύεται ο βαθμός επίδρασης κάθε μεταβλητής στην απόδοση του διαμορφωθέντος μοντέλου. Όσο μεγαλύτερη είναι η τιμή, τόσο περισσότερο επηρεάζει τις προβλέψεις. Η αποτίμηση αυτή κρίνεται βαρύνουσας σημασίας για την κατανόηση των

σημαντικότερων παραγόντων που συμβάλλουν στα Αιτήματα ασφάλισης και περιγράφεται στην Εικόνα 7.9 και τον Πίνακα 7.7. Οι δείκτης απόδοσης του μοντέλου περιγράφονται στον Πίνακα 7.8.

ΠΙΝΑΚΑΣ 7.7. ΚΑΤΑΤΑΞΗ ΣΠΟΥΔΑΙΟΤΗΤΑΣ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ

	Feature	Importance
1	ENSO_index	50.44
2	population	20.69
3	Centroid Longitude	9.02
4	Centroid Latitude	8.05
5	distance_to_sea	4.48
6	hydro_density	3.11
7	elevation_mean	2.15
8	transport_density	2.06



ΕΙΚΟΝΑ 7.9 - ΣΠΟΥΔΑΙΟΤΗΤΑ ΤΩΝ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ-ΠΑΡΑΜΕΤΡΩΝ ΤΟΥ ΔΙΑΜΟΡΦΩΘΕΝΤΟΣ ΜΟΝΤΕΛΟΥ CATBOOST

ΠΙΝΑΚΑΣ 7.8 – ΤΙΜΕΣ ΔΕΙΚΤΩΝ ΑΠΟΔΟΣΗΣ ΤΕΛΙΚΟΥ Μ ΟΝΤΕΛΟΥ ΜΗΧ. ΕΚΜΑΘΗΣΗΣ

Συμβολισμός	Δείκτης	Τιμή
R^2	Συντελεστής Προσδιορισμού	0,638
RMSE	Ριζική Μέση Τυπική Απόκλιση	0,81
MAE	Μέσο Απόλυτο Σφάλμα	0,59

8. Συμπεράσματα & Προτάσεις για Μελλοντική Έρευνα

8.1 Κύρια Συμπεράσματα

Από την έρευνα αυτή προκύπτουν τα εξής συμπεράσματα.

Αναλύοντας τα αποτελέσματα της ερευνητικής διαδικασίας παρατηρήθηκε ότι οι συσχετίσεις (Pearson correlation) μεταξύ των δεικτών ENSO και των Αιτημάτων ασφαλιστικών αποζημιώσεων από πλημμύρες ήταν χαμηλές. Συγκεκριμένα το 66% των Η.Π.Α. είχε βαθμό συσχέτισης κάτω από 0.10 με τον ετήσιο μ.ο. του δείκτη ENSO (ENSO_AVG) (33 πολιτείες) ενώ με τον αντίστοιχο μέγιστο ετήσιο δείκτη (MAX_ENSO) το 56% (28 πολιτείες). Και στις δυο περιπτώσεις οι υπόλοιπες πολιτείες είχαν πάνω από 0.1. Αυτό οφείλεται σε πολλούς παράγοντες. Αρχικά, οι δείκτες ENSO είναι δείκτες παγκοσμίου κλίμακας, και δεν είναι σχεδιασμένοι για να ποσοτικοποιήσουν την επιρροή συγκεκριμένα στις περιοχές των ΗΠΑ. Επιπλέον, τα αιτήματα ασφαλιστικών αποζημιώσεων δεν εξαρτώνται μονοσήμαντα από τα μετεωρολογικά φαινόμενα, αλλά και από κοινωνικούς και θεσμικούς παράγοντες (πχ βαθμός αστικοποίησης, ασφαλιστικές πρακτικές, ενημέρωση κατοίκων κ.α.). Τέλος, η επίδραση των φαινομένων ENSO δεν προκαλεί πάντα πλημμύρες και καταστροφές πάρα μόνο στα μέγιστα σημεία της.

Στο προηγούμενο Κεφάλαιο (Κεφάλαιο 7) έγιναν εκτενείς συγκρίσεις ανάμεσα στις συσχετίσεις των δεικτών ENSO με τους τα καταγεγραμμένα συνολικά αιτήματα αποζημίωσης αλλά και με τις μεταβλητές παροχής αντίστοιχα. Όπως ευκολά παρατηρεί κάνεις τόσο από τα διαγράμματα Boxplot όσο και τα διαγράμματα Dumbbell οι δείκτες ENSO εμφανίζουν μεγαλύτερη συσχέτιση με τα δεδομένα παροχών. Η διαφορά αυτή οφείλεται σε ποικίλους παράγοντες. Αρχικά τα δεδομένα παροχής εκφράζουν κυρίως το φυσικό φαινόμενο στο οποίο είναι ίσως ευκολότερη η αναγνώριση της επίδρασης ενός κλιματολογικού δείκτη. Αντίθετα, τα Αιτήματα Αποζημίωσης εξαρτώνται από κοινωνικοοικονομικούς παράγοντες, όπως ο βαθμός αστικοποίησης, εκπαίδευσης πληθυσμού, το βιοτικό επίπεδο, η ύπαρξη ασφαλιστικής κουλτούρας κλπ. Χαρακτηριστικό παράδειγμα αποτελεί η Καλιφόρνια, που αποτελεί μια εξελιγμένη κοινωνία, οπού η τοπική πολιτική ασφαλιστικής

κάλυψης αλλά και η κουλτούρα των πολιτών οδηγούν σε αυξημένα Αιτήματα Αποζημιώσεων. Αντίθετα σε κάποια άλλη Πολιτεία, μπορεί να είναι εμφανή τα αποτελέσματα των φαινομένων ENSO αλλά να μην κατατεθεί ποτέ κάποιο αίτημα αποζημίωσης, με αποτέλεσμα τα νούμερα αυτά να μη δίνουν ακριβή εικόνα και τη μειωμένη συσχέτιση (με τους δείκτες ENSO) σε σχέση με τον Αριθμό Αιτημάτων Αποζημίωσης. Συμπερασματικά, τα υδρολογικά δεδομένα δίνουν μία σαφέστερα καλύτερη εικόνα για τις συνέπειες των φαινομένων ENSO.

8.2 Προτάσεις για Βελτιώσεις & Επέκταση της Μεθοδολογίας

Στο τελικό στάδιο της ολοκλήρωσης της παρούσας πτυχιακής εργασίας ιδιαίτερο ενδιαφέρον παρουσιάζουν οι παρατηρήσεις και ιδέες που σημειώθηκαν κατά την διάρκεια της εξέλιξης της για τη συνέχεια του έργου της επιστημονικής κοινότητας.

Η βασική πρόταση για μελλοντικές ερευνητικές κινήσεις σε σχέση με τα φαινόμενα ENSO θα ήταν η μελέτη και ανάλυση τους σε σχέση με άλλα υδρομετεωρολογικά δεδομένα, που δεν δύναται να διαστρεβλωθούν από τον ανθρώπινο και κοινωνικό παράγοντα. Έτσι θα είναι πιο εύκολα εμφανής η επιρροή των φαινομένων στους μηχανισμούς της φύσης.

Το Μοντέλο Μηχανικής Εκμάθησης, που αναπτύχθηκε, παρ’ ότι αναπτύχθηκε συμπληρωματικά της κεντρικής ιδέας της πτυχιακής εργασίας εμφάνισε ικανοποιητική αξιοπιστία. Η επέκταση του στο σύνολο των Η.Π.Α. για την οποία βέβαια απαιτούνται μεγάλοι υπολογιστικοί πόροι σίγουρα θα εμφανίσει πολύ ενδιαφέροντα αποτελέσματα.

Τέλος, εκτός από τη γενίκευση του μοντέλου στις ΗΠΑ, μια ακόμα πρόταση θα ήταν η προσθήκη δεδομένων παροχής στο Μοντέλο μηχανικής Εκμάθησης για να ληφθούν υπόψη χωρικά σημαντικές συσχετίσεις μεταξύ παροχής και αιτημάτων προς αποζημίωση, όπως έχουν αναδειχθεί στη βιβλιογραφία (Papoulakos, K., Iliopoulou, T., Dimitriadis, P., Tsaknias, D. and Koutsoyiannis, D., 2025).

Βιβλιογραφία

Jonkman, S. N. (2005). Global perspectives on loss of human life caused by floods. *Natural Hazards*, 34(2), 151–175. https://doi.org/10.1007/s11069-004-8891-3

Cai, W., Santoso, A., Wang, G., et al. (2020). Increasing frequency of extreme El Niño events due to greenhouse warming. Nature Reviews Earth & Environment, 1(4), 198–210. https://doi.org/10.1038/s43017-020-0031-7

Cayan, D.R., Redmond, K.T., & Riddle, L.G. (1999). ENSO and hydrologic extremes in the Western United States. Journal of Climate, 12(9), 2881–2893. https://doi.org/10.1175/1520-0442(1999)012<2881:EAHEIT>2.0.CO;2

Coles, S. (2001). An Introduction to Statistical Modeling of Extreme Values. Springer. https://doi.org/10.1007/978-1-4471-3675-0

Dettinger, M.D. (1999). Climate change, atmospheric rivers, and floods in California. Journal of the American Water Resources Association, 35(6), 1417–1429. https://doi.org/10.1111/j.1752-1688.1999.tb04226.x

França, F., Ferreira, J., Vaz-de-Mello, F.Z., et al. (2020). El Niño impacts on human-modified tropical forests: Consequences for dung beetle diversity and associated ecological processes. Biotropica, 52(1), 252–262. https://doi.org/10.1111/btp.12756

Phillips, O. L., Aragão, L. E. O. C., Lewis, S. L., Fisher, J. B., Lloyd, J., López-González, G., Malhi, Y., Monteagudo, A., Peacock, J., Quesada, C. A., et al. (2009). Drought sensitivity of the Amazon rainforest. Science, 323(5919), 1344–1347. https://doi.org/10.1126/science.1164033

Kobayashi, N., Nakagawa, M., & Kanzaki, M. (2022).Long-term effects of El Niño on seedling survival in a seasonal tropical forest in northern Thailand. Ecological Research, 37, 652–664. https://doi.org/10.1111/1440-1703.12310

Huang, B., Thorne, P. W., Banzon, V. F., Boyer, T., Chepurin, G., Lawrimore, J. H., Menne, M. J., Smith, T. M., Vose, R. S., Zhang, H.-M., & Liu, W. (2017). *NOAA Extended Reconstructed Sea Surface Temperature (ERSST), Version 5.*

NOAA National Centers for Environmental Information.
https://doi.org/10.7289/V5T72FNM

Hirahara, S., Ishii, M., & Fukuda, Y. (2014).Centennial-scale sea surface temperature analysis and its uncertainty.Journal of Climate, 27(1), 57–75.
https://doi.org/10.1175/JCLI-D-12-00837.1

U.S. Census Bureau. (2023).TIGER/Line Shapefiles. United States Census Bureau.
https://www.census.gov/geographies/mapping-files/time-series/geo/tiger-line-file.html

U.S. Geological Survey (USGS). (2023)National Hydrography Dataset (NHD). U.S. Department of the Interior.https://www.usgs.gov/national-hydrography

OpenStreetMap contributors. (2023)Planet dump retrieved from https://planet.openstreetmap.org.https://www.openstreetmap.org

U.S. Geological Survey (USGS). (2023).National Transportation Dataset (NTD). U.S. Department of the Interior.https://www.usgs.gov/national-transportation-dataset

U.S. Geological Survey (USGS). (2023).3D Elevation Program (3DEP) – 1/3 arc-second historical elevation data.U.S. Department of the Interior.
https://www.usgs.gov/3d-elevation-program

U.S. Census Bureau. (2023).Population Estimates Program (PEP).
United States Department of Commerce.
https://www.census.gov/programs-surveys/popest.html

U.S. Geological Survey. (2006). *Sierra Nevada Headwaters Stream Mapping Project*. U.S. Department of the Interior.California Resources Agency. (1999). CalWater 2.2 Watershed Boundaries Dataset and Planning Documents. Sacramento, CA.

Santa Barbara County Public Works Department. (2015). Hydrology Report – Watershed Management for Stream Systems.

U.S. Geological Survey. (2009). Urban Stream Density Estimates for Major U.S. Cities. National Water-Quality Assessment (NAWQA) Program.

North Coast Regional Water Quality Control Board. (2012). North Coast Watershed Assessment Reports. California Environmental Protection Agency.

Caltrans PRD 2021 – Santa Barbara County (Urban Fringe): County of Santa Barbara. (2021). *Transportation Element – Caltrans PRD 2021*. https://www.countyofsb.org/2408/Transportation

Caltrans PRD 2021 – Los Angeles (Urban Roads): Los Angeles Metro. (2021). Countywide Transportation Plan – Caltrans PRD. https://www.metro.net/about/reports/2021-countywide-transportation-plan/

Urban Modeling Estimate – San Diego (SANDAG): San Diego Association of Governments (SANDAG). (n.d.). Transportation Planning and Modeling. https://www.sandag.org/index.asp?fuseaction=home.page

San Francisco City Estimate – Road Area Calculation: San Francisco County Transportation Authority (SFCTA). (n.d.). Projects & Data Explorer. https://www.sfcta.org/

Caltrans PRD + Sparse County Avg – San Bernardino: San Bernardino County Transportation Authority (SBCTA). (n.d.). Transportation Data and Planning. https://sbcta.com/

Fresno County – Ag-heavy reference: Fresno County Public Works and Planning Department. (n.d.). Transportation Planning. https://www.co.fresno.ca.us/departments/public-works-planning

Alameda County – Urban Studies + Caltrans PRD: Alameda County Transportation Commission (ACTC). (n.d.). Transportation Planning and Performance. https://www.alamedactc.org/

Imperial County – Rural Road Stats: Imperial County Public Works Department. (n.d.). Roadway Planning. https://www.co.imperial.ca.us/publicworks/Freund, M.B.,

Henley, B.J., Karoly, D.J., et al. (2019). Higher frequency of Central Pacific El Niño events in recent decades. Nature Geoscience, 12, 450–455. https://doi.org/10.1038/s41561-019-0353-3

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning: With Applications in R (Vol. 103). Springer. https://doi.org/10.1007/978-1-4614-7138-7

Hamlet, A.F., & Lettenmaier, D.P. (2007). Effects of 20th century warming and climate variability on flood risk in the western US. Water Resources Research, 43(6). https://doi.org/10.1029/2006WR005099

Chen, T., & Guestrin, C. (2016) XGBoost: A scalable tree boosting system.In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794).https://doi.org/10.1145/2939672.2939785

Tepetidis, N., Koutsoyiannis, D., Iliopoulou, T., & Dimitriadis, P. (2024). Investigating the performance of the Informer model for streamflow forecasting. Water, 16(20), Article 20734441. https://doi.org/10.3390/w16202041

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: unbiased boosting with categorical features.In Advances in Neural Information Processing Systems (NeurIPS), 31, 6638–6648. https://proceedings.neurips.cc/paper_files/paper/2018/file/14491b756b3a51daac41c24863285549-Paper.pdf

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., et al. (2011).Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

Huang, B., Thorne, P.W., Banzon, V.F., et al. (2017). NOAA Extended Reconstructed Sea Surface Temperature (ERSST), Version 5. NOAA NCEI. https://doi.org/10.7289/V5T72FNM

Ishii, M., Shouji, A., Sugimoto, S., & Matsumoto, T. (2005). Objective Analyses of Sea-Surface Temperature and Marine Meteorological Variables for the 20th Century using ICOADS and the Kobe Collection. International Journal of Climatology, 25, 865–879.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013).An Introduction to Statistical Learning: With Applications in R (Vol. 103). Springer. https://doi.org/10.1007/978-1-4614-7138-7

Kaas, R., Goovaerts, M., Dhaene, J., & Denuit, M. (2008). Modern actuarial risk theory: Using R. Springer.

Kunkel, K.E., Andsager, K., & Easterling, D.R. (2003). Long-term trends in extreme precipitation events over the conterminous United States and Canada. Journal of Climate, 16(13), 2125–2147. https://doi.org/10.1175/2768.1

Murakami, H., et al. (2025). Forecasting Challenges in 2024: Systematic Overprediction of North Atlantic Tropical Cyclone Activity in Seasonal Forecasts. Geophysical Research Letters (submitted).

NOAA Physical Sciences Laboratory (PSL). (2023).NOAA Physical Sciences Laboratory Datasets & Research Tools.National Oceanic and Atmospheric Administration. https://psl.noaa.gov

NOAA. El Niño / Southern Oscillation (ENSO), Southern Oscillation Index (SOI), National Centers for Environmental Information (NCEI). Διαθέσιμο στο: https://www.ncei.noaa.gov/access/monitoring/enso/

Papoulakos, K., Iliopoulou, T., Dimitriadis, P., Tsaknias, D., and Koutsoyiannis, D. et al. Spatiotemporal clustering of streamflow extremes and relevance to flood insurance claims: a stochastic investigation for the contiguous USA. Nat Hazards 121, 447–484 (2025). https://doi.org/10.1007/s11069-024-06766-z

Ropelewski, C.F., & Halpert, M.S. (1986). North American precipitation and temperature patterns associated with the El Niño/Southern Oscillation (ENSO). Monthly Weather Review, 114(12), 2352–2362. https://doi.org/10.1175/1520-0493(1986)114<2352:NAPATP>2.0.CO;2

Serinaldi, F., & Kilsby, C.G. (2016). The importance of preconditioning in extreme rainfall analysis: an application to the UK floods in winter 2013/2014. Hydrology and Earth System Sciences, 20(2), 465–483.

Smith, J.A., Baeck, M.L., & Zhang, Y. (2013). Extreme rainfall and flooding from landfalling tropical cyclones in the United States. Water Resources Research, 49(6), 3694–3707. https://doi.org/10.1002/wrcr.20298

Swain, D.L., Tsiang, M., Haugen, M., et al. (2016). The extraordinary California drought of 2013–2014: Character, context, and the role of climate change. Bulletin of the American Meteorological Society, 97(12), S3–S7. https://doi.org/10.1175/BAMS-D-16-0147.1

Trenberth, K.E. (1997). The definition of El Niño. Bulletin of the American Meteorological Society, 78(12), 2771–2777. https://doi.org/10.1175/1520-0477(1997)078<2771:TDOENO>2.0.CO;2

NOAA Climate.gov. (2015).
Not all El Niños are created equal: Why El Niño may not mean wet for California. National Oceanic and Atmospheric Administration. https://www.climate.gov/news-features/blogs/enso/not-all-el-ni%C3%B1os-are-created-equal

U.S. Geological Survey (USGS). Floods of 1997 in Northern California and Southern Oregon. Fact Sheet 111-98. https://pubs.usgs.gov/fs/fs-11-98/

Ward, P.J., Jongman, B., Weiland, F.S., et al. (2014). Strong influence of El Niño Southern Oscillation on flood risk around the world. Proceedings of the National Academy of Sciences, 111(44), 15659–15664. https://doi.org/10.1073/pnas.1409822111

10. Python Scripts

10.1 Τελική Διαμόρφωση Μοντέλου Μηχανικής Εκμάθησης CatBoost

```
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.metrics import r2_score, mean_squared_error, mean_absolute_error
from catboost import CatBoostRegressor

df = pd.read_csv(r"C:/Users/chris/Desktop/Πτυχιακη_final_after
crash/California_ML/Merging/claims_CA_with_population_and_elevation.csv")

df['claims_per_100k'] = df['Num_Claims'] / (df['population'] / 100000)
df['log_claims_per_100k'] = np.log1p(df['claims_per_100k'])

features = [
    'hydro_density', 'transport_density', 'distance_km',
    'INTPTLAT', 'INTPTLON', 'ENSO_avg', 'population', 'elev_mean'
]
target = 'log_claims_per_100k'

df = df.dropna(subset=features + [target])
X = df[features]
y = df[target]

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

model = CatBoostRegressor(
    verbose=0,
    random_state=42,
    iterations=200,
    depth=4,
    learning_rate=0.05
)
model.fit(X_train, y_train)

y_pred = model.predict(X_test)
print("\n✅ Final CatBoost Model (California - All Features):")
print("R²:", round(r2_score(y_test, y_pred), 3))
print("RMSE:", round(mean_squared_error(y_test, y_pred, squared=False), 2))
print("MAE:", round(mean_absolute_error(y_test, y_pred), 2))
```

10.2 Εκτύπωση αποτελεσμάτων Προβλέψεων σε κείμενο και Διάγραμμα του μοντέλου για τα έτη 2007-2011

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.metrics import r2_score, mean_squared_error, mean_absolute_error
from catboost import CatBoostRegressor

df = pd.read_csv(r"C:/Users/chris/Desktop/Πτυχιακη_final_after
crash/California_ML/Merging/claims_CA_with_population_and_elevation.csv")

df['claims_per_100k'] = df['Num_Claims'] / (df['population'] / 100000)
df['log_claims_per_100k'] = np.log1p(df['claims_per_100k'])

features = [
    'hydro_density', 'transport_density', 'distance_km',
    'INTPTLAT', 'INTPTLON', 'ENSO_avg', 'population', 'elev_mean'
]
target = 'log_claims_per_100k'

df = df.dropna(subset=features + [target])
X = df[features]
y = df[target]

# Save year info before splitting
years = df['yearOfLoss']

X_train, X_test, y_train, y_test, years_train, years_test = train_test_split(
    X, y, years, test_size=0.2, random_state=42
)

model = CatBoostRegressor(
    verbose=0,
    random_state=42,
    iterations=200,
    depth=4,
    learning_rate=0.05
)
model.fit(X_train, y_train)

y_pred = model.predict(X_test)
print("\n✅ Final CatBoost Model (California - All Features):")
print("R²:", round(r2_score(y_test, y_pred), 3))
print("RMSE:", round(mean_squared_error(y_test, y_pred, squared=False), 2))
```

```
print("MAE:", round(mean_absolute_error(y_test, y_pred), 2))

mask_2007_2011 = years_test.isin([2007, 2008, 2009, 2010, 2011])

y_test_2007_2011 = y_test[mask_2007_2011]
y_pred_2007_2011 = y_pred[mask_2007_2011]

y_test_other = y_test[~mask_2007_2011]
y_pred_other = y_pred[~mask_2007_2011]

plt.figure(figsize=(8, 6))
plt.scatter(y_test_other, y_pred_other, alpha=0.6, label="Other Years", color="blue")
plt.scatter(y_test_2007_2011, y_pred_2007_2011, alpha=0.8, label="2007–2011",
color="orange")
plt.plot([y.min(), y.max()], [y.min(), y.max()], 'r--', label="Perfect Prediction")
plt.xlabel("Actual log(Claims per 100k)")
plt.ylabel("Predicted log(Claims per 100k)")
plt.title("CatBoost Predictions vs Actuals (California)")
plt.legend()
plt.grid(True)
plt.tight_layout()
plt.show()
```

10.3 Εκτύπωση αποτελεσμάτων Σπουδαιότητας Χαρακτηριστικών Εισόδου

```
import pandas as pd
import matplotlib.pyplot as plt
from catboost import CatBoostRegressor, Pool

model_path = r"C:/Users/chris/Desktop/Πτυχιακη_final_after
crash/California_ML/Modeling/catboost_final_model.cbm"
model = CatBoostRegressor()
model.load_model(model_path)
feature_names = [
    'hydro_density', 'transport_density', 'distance_km',
    'INTPTLAT', 'INTPTLON', 'ENSO_avg', 'population', 'elev_mean'
]

importances = model.get_feature_importance()
feature_importance_df = pd.DataFrame({
    'Feature': feature_names,
    'Importance': importances
}).sort_values(by='Importance', ascending=False)
print("\n📊 Feature Importances:")
print(feature_importance_df.round(2))
```

```
plt.figure(figsize=(8, 5))
plt.barh(feature_importance_df['Feature'], feature_importance_df['Importance'],
color='#4B8BBE')
plt.gca().invert_yaxis()
plt.title("Feature Importance (CatBoost)")
plt.xlabel("Importance Score")
plt.grid(axis='x', linestyle='--', alpha=0.5)
plt.tight_layout()
plt.show().
```

10.4 Φιλτράρισμα στοιχείων από βάση δεδομένων FEMA

```
import os
import pandas as pd
import time
from tqdm import tqdm

input_path = r"C:\Users\chris\Desktop\Πτυχιακη_final_after crash\FimaNfipClaims.csv"
output_raw = r"C:\Users\chris\Desktop\Πτυχιακη_final_after
crash\California\Claims\claims_CA.csv"
output_grouped = r"C:\Users\chris\Desktop\Πτυχιακη_final_after
crash\California\Claims\claims_CA_grouped.csv"
output_grouped_money = r"C:\Users\chris\Desktop\Πτυχιακη_final_after
crash\California\Claims\claims_CA_grouped_money.csv"
chunk_size = 100000
state_filter = 'CA'

county_name_lookup = {
   6001.0: "Alameda", 6003.0: "Alpine", 6005.0: "Amador", 6007.0: "Butte", 6009.0:
"Calaveras",
   6011.0: "Colusa", 6013.0: "Contra Costa", 6015.0: "Del Norte", 6017.0: "El Dorado",
   6019.0: "Fresno", 6021.0: "Glenn", 6023.0: "Humboldt", 6025.0: "Imperial", 6027.0:
"Inyo",
   6029.0: "Kern", 6031.0: "Kings", 6033.0: "Lake", 6035.0: "Lassen", 6037.0: "Los Angeles",
   6039.0: "Madera", 6041.0: "Marin", 6043.0: "Mariposa", 6045.0: "Mendocino", 6047.0:
"Merced",
   6049.0: "Modoc", 6051.0: "Mono", 6053.0: "Monterey", 6055.0: "Napa", 6057.0:
"Nevada",
   6059.0: "Orange", 6061.0: "Placer", 6063.0: "Plumas", 6065.0: "Riverside", 6067.0:
"Sacramento",
   6069.0: "San Benito", 6071.0: "San Bernardino", 6073.0: "San Diego", 6075.0: "San
Francisco",
   6077.0: "San Joaquin", 6079.0: "San Luis Obispo", 6081.0: "San Mateo", 6083.0: "Santa
Barbara",
   6085.0: "Santa Clara", 6087.0: "Santa Cruz", 6089.0: "Shasta", 6091.0: "Sierra", 6093.0:
"Siskiyou",
   6095.0: "Solano", 6097.0: "Sonoma", 6099.0: "Stanislaus", 6101.0: "Sutter", 6103.0:
"Tehama",
```

```
    6105.0: "Trinity", 6107.0: "Tulare", 6109.0: "Tuolumne", 6111.0: "Ventura", 6113.0:
"Yolo", 6115.0: "Yuba"
}

file_size_mb = os.path.getsize(input_path) / (1024 * 1024)
avg_chunk_size_mb = 8
estimated_total_chunks = int(file_size_mb / avg_chunk_size_mb)

start_time = time.time()
output_rows = []
processed_chunks = 0
total_rows = 0

print(f"🚀 Starting extraction... Estimated chunks: ~{estimated_total_chunks}")

for chunk in tqdm(pd.read_csv(input_path, chunksize=chunk_size, low_memory=False),
desc="Processing", unit="chunk"):
    filtered = chunk[chunk['state'] == state_filter]
    filtered = filtered.dropna(subset=['latitude', 'longitude'])
    output_rows.append(filtered)
    total_rows += len(filtered)
    processed_chunks += 1

    elapsed = time.time() - start_time
    avg_time = elapsed / processed_chunks
    remaining = max(estimated_total_chunks - processed_chunks, 1)
    time_remaining = avg_time * remaining

    print(f"□ Chunk {processed_chunks} processed — Est. remaining: {time_remaining:.1f}
sec")

df_ca_claims = pd.concat(output_rows)
df_ca_claims.to_csv(output_raw, index=False)

print(f"\n✅ Done! Extracted {total_rows} CA claims.")
print(f"📁 Saved to: {output_raw}")

grouped_base = df_ca_claims.groupby(['countyCode', 'yearOfLoss']).agg(
    Num_Claims=('id', 'count'),
    Total_PolicyCount=('policyCount', 'sum')
).reset_index()

grouped_base['Mean_Claims_Per_Policy'] = grouped_base['Num_Claims'] /
grouped_base['Total_PolicyCount']
grouped_base['countyName'] = grouped_base['countyCode'].map(county_name_lookup)
grouped_base['state'] = state_filter
```

```
cols_base = ['countyCode', 'countyName', 'yearOfLoss', 'state', 'Num_Claims',
'Total_PolicyCount', 'Mean_Claims_Per_Policy']
grouped_base = grouped_base[cols_base]

grouped_base.to_csv(output_grouped, index=False)
print(f"✅ Saved grouped claims to:\n{output_grouped}")

grouped_money = df_ca_claims.groupby(['countyCode', 'yearOfLoss']).agg(
    Num_Claims=('id', 'count'),
    Total_PolicyCount=('policyCount', 'sum'),
    Total_Building_Paid=('amountPaidOnBuildingClaim', 'sum'),
    Total_Contents_Paid=('amountPaidOnContentsClaim', 'sum'),
    Total_ICC_Paid=('amountPaidOnIncreasedCostOfComplianceClaim', 'sum'),
    Net_Building_Payment=('netBuildingPaymentAmount', 'sum'),
    Net_Contents_Payment=('netContentsPaymentAmount', 'sum'),
    Net_ICC_Payment=('netIccPaymentAmount', 'sum')
).reset_index()

grouped_money['Mean_Claims_Per_Policy'] = grouped_money['Num_Claims'] /
grouped_money['Total_PolicyCount']
grouped_money['countyName'] =
grouped_money['countyCode'].map(county_name_lookup)
grouped_money['state'] = state_filter

cols_money = ['countyCode', 'countyName', 'yearOfLoss', 'state',
          'Num_Claims', 'Total_PolicyCount', 'Mean_Claims_Per_Policy',
          'Total_Building_Paid', 'Total_Contents_Paid', 'Total_ICC_Paid',
          'Net_Building_Payment', 'Net_Contents_Payment', 'Net_ICC_Payment']

grouped_money = grouped_money[cols_money]
grouped_money.to_csv(output_grouped_money, index=False)

print(f"✅ Saved grouped money claims to:\n{output_grouped_money}")
```

10.5 Αρχική Ένωση Διαφορετικών Χαρακτηριστικών με σκοπό τη Δημιουργία Βάσης δεδομένων για το Μοντέλο Μηχανικής Εκμάθησης

(**) σε κάθε στάδιο προσθήκης νέου χαρακτηριστικού χρησιμοποιούνταν ένα αντίστοιχο script για τη νέα μεταβλητή εισόδου

```
import pandas as pd
import os

base_path = r"C:/Users/chris/Desktop/Πτυχιακη_final_after crash/California_ML"
```

```
claims_path = f"{base_path}/Claims/claims_CA_grouped.csv"
hydro_path = f"{base_path}/outputs_hydro/hydro_density_output.csv"
transport_path = f"{base_path}/outputs_transp/transport_density_ALAND.csv"
enso_path = f"{base_path}/ENSO/ENSO_1950_2024_Yearly_Averages.csv"
output_path = f"{base_path}/Merging/claims_CA_merged_for_ML.csv"

df_claims = pd.read_csv(claims_path)
df_hydro = pd.read_csv(hydro_path)
df_transport = pd.read_csv(transport_path)
df_enso = pd.read_csv(enso_path)

df_hydro = df_hydro.rename(columns={"hydro_km2": "hydro_density"})

df_merged = df_claims.merge(df_hydro[['countyCode', 'hydro_density']], on='countyCode',
how='left')

df_merged = df_merged.merge(df_transport[['countyCode', 'density_km2']],
on='countyCode', how='left')
df_merged = df_merged.rename(columns={'density_km2': 'transport_density'})

df_enso = df_enso.rename(columns={'year': 'yearOfLoss'})
df_merged = df_merged.merge(df_enso, on='yearOfLoss', how='left')

os.makedirs(os.path.dirname(output_path), exist_ok=True)
df_merged.to_csv(output_path, index=False)

print(f"✅ Merged ML dataset saved to:\n{output_path}")
```

11. QGIS Scripts

11.1 Υπολογισμός Πυκνότητας Υδρογραφικού Δικτύου

```
import processing
import time
import csv
from qgis.PyQt.QtCore import QVariant

print("🔷 Starting hydrographic density calculation for California...")

counties_path = "C:/Users/chris/Desktop/Πτυχιακη_final_after
crash/Counties_Boundaries_Census_tiger/tl_2023_us_county.shp"
nhd_path = "C:/Users/chris/Desktop/Πτυχιακη_final_after crash/California_ML/Database
(hydro)/NHD_H_California_State_GPKG.gpkg|layername=NHDFlowline"
```

```
output_gpkg = "C:/Users/chris/Desktop/Πτυχιακη_final_after
crash/California_ML/outputs_hydro/hydro_density_output.gpkg"
output_csv = "C:/Users/chris/Desktop/Πτυχιακη_final_after
crash/California_ML/outputs_hydro/hydro_density_output.csv"

counties_layer = QgsVectorLayer(counties_path, "All_Counties", "ogr")
if not counties_layer.isValid():
    raise Exception("❌ Counties layer failed to load.")

ca_counties = processing.run("native:extractbyattribute", {
    'INPUT': counties_layer,
    'FIELD': "STATEFP",
    'OPERATOR': 0,
    'VALUE': '06',
    'OUTPUT': QgsProcessing.TEMPORARY_OUTPUT
})['OUTPUT']

ca_counties_proj = processing.run("native:reprojectlayer", {
    'INPUT': ca_counties,
    'TARGET_CRS': QgsCoordinateReferenceSystem("EPSG:5070"),
    'OUTPUT': QgsProcessing.TEMPORARY_OUTPUT
})['OUTPUT']

flowline_layer = QgsVectorLayer(nhd_path, "CA_Flowlines", "ogr")
if not flowline_layer.isValid():
    raise Exception("❌ NHDFlowline layer failed to load.")

flowline_proj = processing.run("native:reprojectlayer", {
    'INPUT': flowline_layer,
    'TARGET_CRS': QgsCoordinateReferenceSystem("EPSG:5070"),
    'OUTPUT': QgsProcessing.TEMPORARY_OUTPUT
})['OUTPUT']

flowline_fixed = processing.run("native:fixgeometries", {
    'INPUT': flowline_proj,
    'OUTPUT': QgsProcessing.TEMPORARY_OUTPUT
})['OUTPUT']

clipped_lines = processing.run("native:clip", {
    'INPUT': flowline_fixed,
    'OVERLAY': ca_counties_proj,
    'OUTPUT': QgsProcessing.TEMPORARY_OUTPUT
})['OUTPUT']

lengths_layer = processing.run("native:sumlinelengths", {
    'POLYGONS': ca_counties_proj,
    'LINES': clipped_lines,
    'LEN_FIELD': 'LENGTH',
```

```
    'COUNT_FIELD': 'COUNT',
    'OUTPUT': QgsProcessing.TEMPORARY_OUTPUT
})['OUTPUT']

lengths_layer.dataProvider().addAttributes([
    QgsField("aland_km2", QVariant.Double),
    QgsField("hydro_km2", QVariant.Double)
])
lengths_layer.updateFields()

features = list(lengths_layer.getFeatures())
start = time.time()

with edit(lengths_layer):
    for feat in features:
        aland = feat["ALAND"] / 1e6 if feat["ALAND"] else 0
        hydro_len_km = feat["LENGTH"] / 1000 if feat["LENGTH"] else 0
        density = hydro_len_km / aland if aland > 0 else 0
        feat["aland_km2"] = aland
        feat["hydro_km2"] = density
        lengths_layer.updateFeature(feat)

print("✅ Hydrographic density calculated.")

QgsVectorFileWriter.writeAsVectorFormat(lengths_layer, output_gpkg, "UTF-8",
lengths_layer.crs(), "GPKG")
print(f"✅ Saved to: {output_gpkg}")

export_fields = ["NAME", "aland_km2", "LENGTH", "hydro_km2"]
with open(output_csv, mode='w', newline='', encoding='utf-8') as csvfile:
    writer = csv.writer(csvfile)
    writer.writerow(export_fields)
    for feat in lengths_layer.getFeatures():
        row = [feat[field] for field in export_fields]
        writer.writerow(row)

print(f"📄 Exported CSV to: {output_csv}")
iface.addVectorLayer(output_gpkg, "Hydro Density CA", "ogr")
```

11.2 Υπολογισμός Πυκνότητας Οδικού Δικτύου

```
import processing
import time
from qgis.PyQt.QtCore import QVariant
import csv

print("🔷 Starting transport density calculation for California...")
```

```
counties_path = "C:/Users/chris/Desktop/Πτυχιακη_final_after
crash/Counties_Boundaries_Census_tiger/tl_2023_us_county.shp"
output_gpkg = "C:/Users/chris/Desktop/Πτυχιακη_final_after
crash/California_ML/outputs_transp/transport_density_output.gpkg"
output_csv = "C:/Users/chris/Desktop/Πτυχιακη_final_after
crash/California_ML/outputs_transp/transport_density_output.csv"

counties_layer = QgsVectorLayer(counties_path, "All_Counties", "ogr")
if not counties_layer.isValid():
    raise Exception("❌ Counties layer failed to load.")
print(f"✅ Loaded counties: {counties_layer.featureCount()} features")

ca_counties = processing.run("native:extractbyattribute", {
    'INPUT': counties_layer,
    'FIELD': "STATEFP",
    'OPERATOR': 0,  # "="
    'VALUE': '06',
    'OUTPUT': QgsProcessing.TEMPORARY_OUTPUT
})['OUTPUT']
print(f"✅ Filtered to CA counties: {ca_counties.featureCount()} features")

ca_counties_proj = processing.run("native:reprojectlayer", {
    'INPUT': ca_counties,
    'TARGET_CRS': QgsCoordinateReferenceSystem("EPSG:5070"),
    'OUTPUT': QgsProcessing.TEMPORARY_OUTPUT
})['OUTPUT']
print("✅ CA counties reprojected to EPSG:5070")

roads_layer = QgsProject.instance().mapLayersByName("Trans_RoadSegment")[0]
if not roads_layer.isValid():
    raise Exception("❌ Roads layer failed to load.")
print(f"✅ Road network loaded from QGIS session: {roads_layer.featureCount()} features")

roads_proj = processing.run("native:reprojectlayer", {
    'INPUT': roads_layer,
    'TARGET_CRS': QgsCoordinateReferenceSystem("EPSG:5070"),
    'OUTPUT': QgsProcessing.TEMPORARY_OUTPUT
})['OUTPUT']

roads_fixed = processing.run("native:fixgeometries", {
    'INPUT': roads_proj,
    'OUTPUT': QgsProcessing.TEMPORARY_OUTPUT
})['OUTPUT']

clipped_roads = processing.run("native:clip", {
    'INPUT': roads_fixed,
    'OVERLAY': ca_counties_proj,
```

```
    'OUTPUT': QgsProcessing.TEMPORARY_OUTPUT
})['OUTPUT']
print(f"✅ Clipped roads to CA counties.")

lengths_layer = processing.run("native:sumlinelengths", {
    'POLYGONS': ca_counties_proj,
    'LINES': clipped_roads,
    'LEN_FIELD': 'LENGTH',
    'COUNT_FIELD': 'COUNT',
    'OUTPUT': QgsProcessing.TEMPORARY_OUTPUT
})['OUTPUT']
print("✅ Total road lengths calculated per county.")

lengths_layer.dataProvider().addAttributes([
    QgsField("area_km2", QVariant.Double),
    QgsField("density_km2", QVariant.Double)
])
lengths_layer.updateFields()

features = list(lengths_layer.getFeatures())
total = len(features)
start_time = time.time()

print("⏳ Calculating area and road density with ETA...")

with edit(lengths_layer):
    for idx, feat in enumerate(features, 1):
        geom = feat.geometry()
        if geom and not geom.isEmpty():
            area_km2 = geom.area() / 1e6
            road_km = feat["LENGTH"] / 1000 if feat["LENGTH"] else 0
            density = road_km / area_km2 if area_km2 > 0 else 0
            feat["area_km2"] = area_km2
            feat["density_km2"] = density
            lengths_layer.updateFeature(feat)

        elapsed = time.time() - start_time
        avg_time = elapsed / idx
        remaining = avg_time * (total - idx)
        print(f"  ➤ {idx}/{total} | Elapsed: {elapsed:.1f}s | Remaining: {remaining:.1f}s",
end="\r")

print("\n✅ Area and road density calculated.")

QgsVectorFileWriter.writeAsVectorFormat(lengths_layer, output_gpkg, "UTF-8",
lengths_layer.crs(), "GPKG")
print(f"✅ Results saved to: {output_gpkg}")
```

```
export_fields = ["NAME", "area_km2", "LENGTH", "density_km2"]
with open(output_csv, mode='w', newline='', encoding='utf-8') as csv_file:
    writer = csv.writer(csv_file)
    writer.writerow(export_fields)
    for feat in lengths_layer.getFeatures():
        row = [feat[field] for field in export_fields]
        writer.writerow(row)
print(f"✅ Results exported to CSV at: {output_csv}")
```

SCRIPT 2 (ΜΕ ΦΟΡΤΩΜΕΝΑ ΤΑ GPKG ΠΟΥ ΠΡΟΚΥΠΤΟΥΝ ΑΠΟ ΤΟ ΠΑΝΩ SCRIPT)

```
from qgis.PyQt.QtCore import QVariant
import csv
import time

layer = QgsProject.instance().mapLayersByName("transport_density_output")[0]
if not layer.isValid():
    raise Exception("❌ Loaded layer is invalid.")

print(f"✅ Layer loaded: {layer.featureCount()} features")

field_names = [f.name() for f in layer.fields()]
if "aland_km2" not in field_names:
    layer.dataProvider().addAttributes([
        QgsField("aland_km2", QVariant.Double),
        QgsField("density_km2", QVariant.Double)
    ])
    layer.updateFields()

features = list(layer.getFeatures())
total = len(features)
start = time.time()

print("🔄 Recalculating density using ALAND...")

with edit(layer):
    for idx, feat in enumerate(features, 1):
        aland_m2 = feat["ALAND"]
        aland_km2 = aland_m2 / 1e6 if aland_m2 else 0
        road_km = feat["LENGTH"] / 1000 if feat["LENGTH"] else 0
        density = road_km / aland_km2 if aland_km2 > 0 else 0

        feat["aland_km2"] = aland_km2
        feat["density_km2"] = density
        layer.updateFeature(feat)

        elapsed = time.time() - start
        print(f"  ➤ {idx}/{total} | Elapsed: {elapsed:.1f}s", end="\r")
```

```
print("\n✅ Density calculated using ALAND.")

output_csv = "C:/Users/chris/Desktop/Πτυχιακη_final_after
crash/California_ML/outputs_transp/transport_density_ALAND.csv"
fields_to_export = ["NAME", "aland_km2", "LENGTH", "density_km2"]

with open(output_csv, mode='w', newline='', encoding='utf-8') as file:
    writer = csv.writer(file)
    writer.writerow(fields_to_export)
    for feat in layer.getFeatures():
        row = [feat[field] for field in fields_to_export]
        writer.writerow(row)

print(f"📄 CSV exported to:\n  {output_csv}")
```

11.3 Δημιουργία Χαρτών Η.Π.Α. με γραφική απεικόνιση Συσχετίσεων ανάμεσα σε Μεγέθη και κατηγοριοποίηση με βάση αριθμητικές τιμές

```
from qgis.core import *
from qgis.PyQt.QtGui import QColor
from qgis.utils import iface
import os

shapefile_path = r'C:/Users/chris/Desktop/Πτυχιακη_final_after
crash/States_Boundaries_Census/tl_2023_us_state.shp'
csv_path = r'C:/Users/chris/Desktop/Πτυχιακη_final_after crash/00- Χάρτες -
Correlation/USA ( states )/state_level_combined/correlation_by_state_pearson.csv'

output_path_avg = r'C:/Users/chris/Desktop/Πτυχιακη_final after crash/00- Χάρτες -
Correlation/USA ( states )/choropleth_state_AVG_ENSO_Correlation.gpkg'
output_path_max = r'C:/Users/chris/Desktop/Πτυχιακη_final after crash/00- Χάρτες -
Correlation/USA ( states )/choropleth_state_MAX_ENSO_Correlation.gpkg'
output_path_min = r'C:/Users/chris/Desktop/Πτυχιακη_final after crash/00- Χάρτες -
Correlation/USA ( states )/choropleth_state_MIN_ENSO_Correlation.gpkg'

states_layer = iface.addVectorLayer(shapefile_path, "US_States", "ogr")
if not states_layer or not states_layer.isValid():
    raise Exception("❌ Failed to load the states shapefile.")
print("✅ States shapefile loaded.")

csv_uri = f"file:///{csv_path.replace(os.sep, '/')}" + "?delimiter=,&crs=EPSG:4326"
csv_layer = QgsVectorLayer(csv_uri, "State Correlation Data", "delimitedtext")
if not csv_layer.isValid():
```

```
    raise Exception("❌ Failed to load the CSV file.")
print("✅ CSV loaded successfully.")

join = QgsVectorLayerJoinInfo()
join.setJoinFieldName("STATE_NAME")
join.setTargetFieldName("NAME")
join.setJoinLayer(csv_layer)
join.setUsingMemoryCache(True)
join.setPrefix("")
states_layer.addJoin(join)
print("🔗 Join complete.")

avg_layer = states_layer.clone()
avg_layer.setName("ENSO_Avg_Correlation")

max_layer = states_layer.clone()
max_layer.setName("MAX_ENSO_Correlation")

min_layer = states_layer.clone()
min_layer.setName("MIN_ENSO_Correlation")

avg_layer.setSubsetString('"Pearson_Corr_avg" IS NOT NULL')
max_layer.setSubsetString('"MAX_ENSO_Corr" IS NOT NULL')
min_layer.setSubsetString('"MIN_ENSO_Corr" IS NOT NULL')

def apply_graduated_style(layer, field_name):
    classes = [
        QgsRendererRange(-1.0, -0.25, QgsSymbol.defaultSymbol(layer.geometryType()), "-1 to
-0.25"),
        QgsRendererRange(-0.25, 0.25, QgsSymbol.defaultSymbol(layer.geometryType()), "-0.25
to 0.25"),
        QgsRendererRange(0.25, 1.0, QgsSymbol.defaultSymbol(layer.geometryType()), "0.25 to
1")
    ]
    renderer = QgsGraduatedSymbolRenderer(field_name, classes)
    renderer.setMode(QgsGraduatedSymbolRenderer.Custom)
    layer.setRenderer(renderer)

apply_graduated_style(avg_layer, 'Pearson_Corr_avg')
apply_graduated_style(max_layer, 'MAX_ENSO_Corr')
apply_graduated_style(min_layer, 'MIN_ENSO_Corr')
print("🎨 Symbology applied.")

def apply_labels(layer, field_name):
    label_settings = QgsPalLayerSettings()
    label_settings.fieldName = '"STATE_NAME" || \' | Corr: \' || format_number("{}",
2)'.format(field_name)
    label_settings.placement = QgsPalLayerSettings.Line
```

```
    label_settings.enabled = True
    labeling = QgsVectorLayerSimpleLabeling(label_settings)
    layer.setLabeling(labeling)
    layer.setLabelsEnabled(True)
    layer.triggerRepaint()

apply_labels(avg_layer, "Pearson_Corr_avg")
apply_labels(max_layer, "MAX_ENSO_Corr")
apply_labels(min_layer, "MIN_ENSO_Corr")
print("🏷️ Labels applied.")

QgsProject.instance().addMapLayer(avg_layer)
QgsProject.instance().addMapLayer(max_layer)
QgsProject.instance().addMapLayer(min_layer)

def export_layer(layer, path):
    options = QgsVectorFileWriter.SaveVectorOptions()
    options.driverName = "GPKG"
    options.fileEncoding = "UTF-8"
    options.layerName = layer.name()
    result = QgsVectorFileWriter.writeAsVectorFormatV3(
        layer,
        path,
        QgsProject.instance().transformContext(),
        options
    )
    if result == QgsVectorFileWriter.NoError:
        print(f"💾 Saved: {path}")
    else:
        print(f"❌ Error saving {layer.name()}. Code: {result}")

export_layer(avg_layer, output_path_avg)
export_layer(max_layer, output_path_max)
export_layer(min_layer, output_path_min)
```