ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΠΟΛΙΤΙΚΩΝ ΜΗΧΑΝΙΚΩΝ

ΤΟΜΕΑΣ ΜΕΤΑΦΟΡΩΝ ΚΑΙ ΣΥΓΚΟΙΝΩΝΙΑΚΗΣ ΥΠΟΔΟΜΗΣ

Ανάλυση Οδικών Ατυχημάτων Στα Ελληνικά Νησιά

Διπλωματική Εργασία

Φιλιππίδης Νικήτας Μαράτος

Επιβλέπων: Γιώργος Γιαννής, Καθηγητής Ε.Μ.Π

Αθήνα, Οκτώβριος 2024

Ευχαριστίες

Θέλω να ευχαριστήσω θερμά τον κ. Γιώργο Γιαννή, Καθηγητή της Σχολής Πολιτικών Μηχανικών Ε.Μ.Π., για την ευκαιρία που μου έδωσε να ασχοληθώ με το θέμα της παρούσας Διπλωματικής Εργασίας, αλλά και για τον βοηθητικό ρόλο που παρείχε κατά τη διάρκεια της εκπόνησης της.

Θέλω να ευχαριστήσω, επίσης, τον κ. Δημήτρη Νικολάου, Διδάκτορα του ΕΜΠ και την κ. Κατερίνα Φώλλα, υποψήφια Διδάκτορα του ΕΜΠ, καθώς αυτό το εγχείρημα θα ήταν ανεκπλήρωτο χωρίς τις δικές τους συμβουλές.

Τέλος, οφείλω ένα μεγάλο ευχαριστώ στους γονείς μου, που ποτέ δεν σταμάτησαν να πιστεύουν στις δυνατότητες μου σε όλη τη διάρκεια των σπουδών μου.

Αθήνα, Οκτώβριος 2024

Φιλιππίδης Νικήτας Μαράτος

Ανάλυση Οδικών Ατυχημάτων Στα Νησιά Της Ελλάδας

Φιλιππίδης Νικήτας Μαράτος

Επιβλέπων: Γιώργος Γιαννής, Καθηγητής Ε.Μ.Π

Σύνοψη

Στόχος της παρούσας Διπλωματικής Εργασίας είναι η ανάλυση των οδικών ατυχημάτων στα ελληνικά νησιά. Η ανάλυση αυτή θα επιτρέψει τον προσδιορισμό της επιρροής του τουρισμού στην οδική ασφάλεια των νησιών, μέσω της συσχέτισης των αφίξεων τουριστών με τα οδικά ατυχήματα και τον αριθμό των νεκρών σε αυτά. Για τον σκοπό αυτό, συλλέχθηκαν δεδομένα μηνιαίων αφίξεων, ατυχημάτων και θανάτων για 38 νησιά της Ελλάδας και αφότου αυτά χωρίστηκαν σε 4 ομάδες με βάση τη γεωγραφική θέση τους, δημιουργήθηκαν σχετικά γραφήματα τεσσάρων νησιών, ώστε να προκύψουν αρχικά συμπεράσματα σχετικά με τις συσχετίσεις αφίξεων – οδικών ατυχημάτων στις περιοχές αυτές. Στη συνέχεια, αυτές οι συσχετίσεις ερευνήθηκαν περαιτέρω με την ανάπτυξη στατιστικών μοντέλων Generalized Linear Model και Random Forest με τη βοήθεια της γλώσσας προγραμματισμού R. Το βασικότερο πόρισμα που προέκυψε από την ανάλυση ήταν ότι οι αφίξεις σε κάποιο νησί συσχετίζονται θετικά με τον αριθμό των ατυχημάτων και των νεκρών που καταγράφονται εκείνη τη χρονική περίοδο. Επίσης, ενώ οι αφίξεις συσχετίστηκαν με τα ατυχήματα σε κάθε ομάδα νησιών που εξετάστηκε, τα αποτελέσματα διέφεραν, πιθανώς εξαιτίας των καλύτερων οδικών συνθηκών που επικρατούν σε νησιά που δέχονται μεγάλο αριθμό αφίξεων, αλλά παρουσιάζουν μικρό αριθμό ατυχημάτων.

Λέξεις Κλειδιά: Οδικά Ατυχήματα, Τουρισμός, Ελληνικά Νησιά, Generalized Linear Model, Random Forest

Road Crashes Analysis In Greek Islands

Filippidis Nikitas Maratos

Supervisor: George Yannis, Professor of N.T.U.A.

Abstract

The objective of the present Diploma Thesis is to analyze road accidents on Greek islands. This analysis will allow for the identification of tourism's impact on road safety on the islands by correlating tourist arrivals with road crash and related fatalities. On that purpose, data on monthly arrivals, crashes, and fatalities were collected for 38 Greek islands, which were then divided into four groups based on their geographical location. Relevant charts were created for four representative islands to provide initial insights into the correlation between arrivals and accidents, as well as arrivals and fatalities in these areas. Subsequently, these correlations were further examined through the development of Generalized Linear Models and Random Forest models using the R programming language. The main finding from the analysis was that arrivals on an island are positively correlated with the number of accidents and fatalities recorded during that period. Additionally, while arrivals were correlated with accidents across each group of islands examined, the results differed, possibly due to better road conditions on islands with high arrival numbers but low accident rates.

Keywords: Road Accidents, Tourism, Greek Islands, Generalized Linear Model, Random Forest

Περίληψη

Η παρούσα Διπλωματική Εργασία στοχεύει να διερευνήσει τα οδικά ατυχήματα στα ελληνικά νησιά και πιο συγκεκριμένα **την επιρροή του τουρισμού στον αριθμό των παθόντων στα οδικά ατυχήματα**. Για τον σκοπό αυτόν δημιουργήθηκαν μοντέλα συσχέτισης ατυχημάτων – αφίξεων τουριστών και μοντέλα συσχέτισης θανάτων-αφίξεων τουριστών.

Για την επίτευξη του στόχου της Διπλωματικής Εργασίας, **συλλέχθηκαν δεδομένα οδικών ατυχημάτων και αφίξεων τουριστών** σε λιμάνια και αεροδρόμια 38 ελληνικών νησιών σε μηνιαία βάση. Τα απαραίτητα για την μελέτη δεδομένα αντλήθηκαν από την Ελληνική Στατιστική Υπηρεσία (ΕΛΣΤΑΤ) για τις αφίξεις σε λιμάνια και από το Ινστιτούτο των Συνδέσμων Ελληνικών Τουριστικών Επιχειρήσεων (ΙΝΣΕΤΕ) για τις αφίξεις τουριστών σε αεροδρόμια. Επιπλέον, μηνιαία στοιχεία ατυχημάτων με τραυματίες και νεκρούς σε οδικά ατυχήματα αντλήθηκαν από τη βάση των αναλυτικών δεδομένων τροχαίων ατυχημάτων του ΕΜΠ, όπως έχουν καταγραφεί στην ΕΛΣΤΑΤ. Όλα τα δεδομένα αφορούν στην περίοδο 2009-2018.

Έπειτα, **αναπτύχθηκαν δύο διαφορετικά μοντέλα για τη συσχέτιση των οδικών ατυχημάτων με τις αφίξεις των τουριστών: α) Generalized Linear Model (GLM) και β) Random Forest**. Το κάθε μοντέλο προέβλεψε τον αριθμό των οδικών ατυχημάτων και θανάτων σε αυτά με βάση τις τουριστικές αφίξεις για τα τελευταία δύο έτη (2017 και 2018), διαβάζοντας τις χρονοσειρές των οκτώ πρώτων ετών της βάσης δεδομένων. Με την χρήση των δεικτών MAE (Mean Absolute Error) και RMSE (Root Mean Squared), τέθηκε δυνατή η σύγκριση των δύο μεθοδολογιών, προκειμένου να ελεγχθεί η ακρίβεια των προβλέψεων. Σημειώνεται ότι τα εξεταζόμενα νησιά χωρίστηκαν σε τέσσερις γεωγραφικές ομάδες (Κυκλάδες, Δωδεκάνησα, Ιόνιο, Κεντρικό και Βόρειο Αιγαίο), για τις οποίες αναπτύχθηκαν ξεχωριστά μοντέλα.

Τα αποτελέσματα που προέκυψαν παρουσιάζονται συγκεντρωτικά στον Πίνακα 6.1. Και οι δυο τύποι μοντέλων έδειξαν ότι για κάθε ομάδα νησιών, οι αφίξεις συσχετίζονται θετικά με τα ατυχήματα και τους θανάτους που σημειώνονται στις υπό μελέτη περιοχές. Αυτό το γεγονός υποδεικνύει ότι η αύξηση των αφίξεων και του τουρισμού, συνεπάγεται αύξηση στα οδικά ατυχήματα και τους θανάτους κατά τις τουριστικές περιόδους.

Ειδικότερα, τα μοντέλα GLM έδειξαν ότι στην περιοχή του Κεντρικού και Βορείου Αιγαίου, η **επιρροή των τουριστικών αφίξεων στον αριθμό των ατυχημάτων είναι μεγαλύτερη** σε σχέση με τα νησιά του Ιονίου, των Δωδεκανήσων και των Κυκλάδων, το οποίο ενδεχομένως οφείλεται

πέρα από τις επικρατούσες συνθήκες (οδική υποδομή, ετοιμότητα πρώτων βοηθειών, κτλ.), και στο γεγονός ότι σε αυτά τα νησιά οι επισκέψεις ξένων τουριστών είναι λιγότερες.

Συγκρίνοντας τα **μοντέλα GLM αφίξεων-θανάτων**, προέκυψε ότι δεν υπάρχει συσχέτιση για την ομάδα νησιών του Κεντρικού και Βορείου Αιγαίου, ενώ στις άλλες τρεις περιοχές που εξετάστηκαν, υπήρχε μία μικρή συσχέτιση αφίξεων – θανάτων, και πιο συγκεκριμένα στις περιοχές των Δωδεκανήσων και των Κυκλάδων. Σημειώνεται ότι ο αριθμός των νεκρών σε οδικά ατυχήματα που καταγράφεται στα νησιά σε μηνιαία βάση είναι μικρός, γεγονός που επηρεάζει την σημαντικότητα των στατιστικών μοντέλων.

Αντίστοιχα μοντέλα αφίξεων-ατυχημάτων και αφίξεων-θανάτων αναπτύχθηκαν με τη μέθοδο **Random Forest**. Σε σχέση με τα μοντέλα για τον αριθμό των θανάτων σε οδικά ατυχήματα, προέκυψε μόνο ένα ικανοποιητικό μοντέλο, για την ομάδα των Δωδεκανήσων με αποδεκτό σφάλμα πρόβλεψης. Για τα δε μοντέλα συσχέτισης αφίξεων ατυχημάτων, προέκυψαν αποδεκτά μοντέλα για όλες τις ομάδες νησιών. Επισημαίνεται παρ’ όλα αυτά, ότι για τις Κυκλάδες και το Κεντρικό και Βόρειο Αιγαίο, το ποσοστό διακύμανσης που εξηγείται είναι σχετικά χαμηλό. Τα μικρότερα σφάλματα πρόβλεψης υπολογίστηκαν για τις Κυκλάδες και τα Δωδεκάνησα, με το Κεντρικό και Βόρειο Αιγαίο και το Ιόνιο να ακολουθούν.

Γενικότερα, συγκρίνοντας τις δύο μεθόδους ως προς τα σφάλματα προβλέψεων, πιο χαμηλοί δείκτες MAE και RMSE παρατηρούνται για τα μοντέλα αφίξεων-ατυχημάτων με τη μέθοδο **Random Forest, η οποία φαίνεται να προβλέπει πιο ικανοποιητικά**.

	Generalized Linear Model (GLM)					
	Fatalities			Crashes		
Ομάδα Νησιών	**Συντελεστής**	**McFadden R^2**	**AICc**	**Συντελεστής**	**McFadden R^2**	**AICc**
Όλα τα νησιά	0.00977 < p = 0.001	0.12	3011.21	0.0148 < p = 0.001	0.10	9580.49
Δωδεκάνησα	0.00909 < p = 0.001	0.16	954.66	0.0123 < p = 0.001	0.12	2638.03
Κυκλάδες	0.01122 < p = 0.001	0.11	628.61	0.0182 < p = 0.001	0.11	2491.14
Ιόνιο	0.00714 < p = 0.001	0.10	634.42	0.0095 < p = 0.001	0.11	1853.99
Κεντρικό/Βόρειο Αιγαίο	0.03354 < p = 0.001	0.06	708.54	0.0574 < p = 0.001	0.10	2227.65

Ομάδα Νησιών	**Fatalities MAE**	**Fatalities RMSE**	**Crashes MAE**	**Crashes RMSE**
Όλα τα νησιά	0.41	1.67	19.66	177.93
Δωδεκάνησα	0.64	2.42	15.40	87.42
Κυκλάδες	0.16	0.34	3.40	15.83
Ιόνιο	0.59	1.07	5.78	18.60
Κεντρικό/Βόρειο Αιγαίο	0.27	0.51	4.35	17.52

	Random Forest			
	Fatalities		Crashes	
Ομάδα Νησιών	**Mean of Squared Residual**	**% of Var explained**	**Mean of Squared Residual**	**% of Var explained**
Όλα τα νησιά	0.286	5.96	4.976	46.71
Δωδεκάνησα	0.454	23.16	5.977	63.15
Κυκλάδες	0.099	-26.53	2.185	16.53
Ιόνιο	0.471	2.84	8.015	53.14
Κεντρικό/Βόρειο Αιγαίο	0.245	-21.47	4.658	15.18

Ομάδα Νησιών	**Fatalities MAE**	**Fatalities RMSE**	**Crashes MAE**	**Crashes RMSE**
Όλα τα νησιά	0.27	0.62	1.41	2.75
Δωδεκάνησα	0.29	0.65	1.06	2.10
Κυκλάδες	0.16	0.40	0.96	1.91
Ιόνιο	0.40	0.72	1.92	3.42
Κεντρικό/Βόρειο Αιγαίο	0.24	0.50	1.73	3.26

Πίνακας 6.1: Συγκεντρωτικός Πίνακας Αποτελεσμάτων

Ακολουθούν τα βασικά συμπεράσματα που προέκυψαν από την ανάλυση των αποτελεσμάτων της στατιστικής ανάλυσης.

1. Ο **τουρισμός, και γενικότερα οι αφίξεις σε κάποιο νησί, συσχετίζονται θετικά με τον αριθμό των ατυχημάτων και των νεκρών** σε οδικά ατυχήματα που καταγράφονται σε εκείνη την χρονική περίοδο.

2. Παρ΄ όλο που οι αφίξεις συσχετίστηκαν με τα ατυχήματα σε όλες τις ομάδες νησιών που εξετάστηκαν, τα αποτελέσματα διέφεραν. Είναι πιθανό σε **νησιά που εμφάνιζαν μικρό αριθμό ατυχημάτων και μεγάλο αριθμό αφίξεων**, να επικρατούν καλύτερες οδικές συνθήκες, είτε διότι το οδικό δίκτυο είναι καταλληλότερο είτε διότι η συμπεριφορά των οδηγών (ντόπιων και τουριστών) είναι καλύτερη.

3. Στην περιοχή του **Κεντρικού και Βορείου Αιγαίου**, η επιρροή των τουριστικών αφίξεων στον αριθμό των ατυχημάτων είναι μεγαλύτερη σε σχέση με τα νησιά του Ιονίου, των Δωδεκανήσων και των Κυκλάδων. Αυτό ενδεχομένως οφείλεται τόσο στις επικρατούσες συνθήκες (πιθανώς χειρότερη οδική υποδομή, ακατάλληλη συμπεριφορά οδηγών, ελλιπής αστυνόμευση, κτλ.), τόσο και στο γεγονός ότι σε αυτά τα νησιά οι επισκέψεις ξένων τουριστών είναι λιγότερες. Είναι γνωστό και από τη βιβλιογραφία ότι οι ξένοι τουρίστες τείνουν να προσαρμόζονται δυσκολότερα σε ένα άγνωστο σε αυτούς οδικό περιβάλλον, επομένως, πιθανώς στα συγκεκριμένα νησιά, λόγω χαμηλότερου ξένου τουρισμού, να μην έχουν ληφθεί τα κατάλληλα μέτρα για τη σωστή προσαρμογή τους.

4. Οι **αφίξεις επηρεάζουν σε μικρότερο βαθμό τους θανάτους** που καταγράφονται εκείνη την περίοδο, με τα μοντέλα να δείχνουν ότι σε κάποιες περιπτώσεις δεν υπάρχει συσχέτιση αυτών των δύο μεταβλητών. Αυτό πιθανώς οφείλεται στο γεγονός ότι ο αριθμός των νεκρών σε οδικά ατυχήματα που καταγράφονται σε μηνιαία βάση στα νησιά είναι σημαντικά μικρός, το οποίο δεν επιτρέπει την ανάπτυξη στατιστικά σημαντικών μοντέλων. Παρ’ όλα αυτά, η μικρή συσχέτιση αφίξεων και θανάτων πιθανώς υποδεικνύει την ύπαρξη και άλλων παραγόντων που συμβάλλουν στη σοβαρότητα των ατυχημάτων, όπως επικίνδυνες συμπεριφορές (υψηλές ταχύτητες, οδήγηση υπό την επήρεια αλκοόλ, κτλ.), οδικές υποδομές, παροχή πρώτων βοηθειών και περίθαλψης μετά το ατύχημα, κτλ., οι οποίοι θα μπορούσαν να διερευνηθούν.

5. Τέλος, το γεγονός ότι τα ατυχήματα αυξάνονται με τις αφίξεις, άλλα όχι οι θάνατοι, οδηγεί στο συμπέρασμα ότι, πιθανώς στην πλειοψηφία των περιπτώσεων, στα **ατυχήματα αυτά εμπλέκονται οδηγοί που δυσκολεύονται να αφομοιώσουν τις κυκλοφοριακές συνθήκες της περιοχής**, ωστόσο ο τρόπος οδήγησης τους δεν είναι τόσο επικίνδυνος ώστε να οδηγήσει σε πολύ σοβαρό ατύχημα.

Περιεχόμενα

Κεφάλαιο 1: Εισαγωγή

1.1 Γενική Ανασκόπηση

Η χώρα μας διαθέτει μια πληθώρα κατοικήσιμων νησιών, χαρακτηριστικό που την τοποθετεί στις κορυφαίες θέσεις παγκοσμίως στην συγκεκριμένη λίστα. Η ομορφιά πολλών από αυτών είναι ο λόγος που ένας μεγάλος όγκος τουριστών, που προέρχεται είτε εντός Ελλάδας, είτε εκτός, επιλέγει να τα επισκεφτεί τους θερινούς μήνες. Παρατηρείται, μάλιστα, ότι ο πληθυσμός αυτών των νησιών αυξάνεται ραγδαία εκείνη την περίοδο και οι δήμοι τους έχουν να προσαρμοστούν σε αυτήν την πραγματικότητα κάθε χρόνο, τέτοια εποχή.

Σύμφωνα με έρευνες του ΕΜΠ, τα ατυχήματα που λαμβάνουν χώρα στη νησιωτική Ελλάδα, είτε αυτά περιέχουν ελαφρά τραυματίες, είτε νεκρούς, είναι λιγότερα από αυτά της ηπειρωτικής. Οι στενότεροι οδοί, η χαμηλότερη κυκλοφορία, ακόμα και ο μικρότερος πληθυσμός συντελούν στο να φαντάζει λογικό αυτό το γεγονός. Ωστόσο, οι κυκλοφοριακές απαιτήσεις του χειμώνα διαφέρουν με αυτές του θέρους, αναλόγως, προφανώς, την επισκεψιμότητα κάθε τόπου. Είναι σπάνιο σε ένα κατοικήσιμο ελληνικό νησί να μην αυξηθεί κατακόρυφα ο πληθυσμός του, μάλιστα πολλά εξ αυτών παρουσιάζουν τριπλασιασμό των κατοίκων τους στους μήνες του Ιουλίου και του Αυγούστου. Έρευνες δείχνουν ότι τα τροχαία ατυχήματα αυξάνονται εξίσου.

1.2 Στόχος διπλωματικής Εργασίας

Στόχος της παρούσας Διπλωματικής Εργασίας είναι **η ανάλυση των οδικών ατυχημάτων στα ελληνικά νησιά.**

Συγκεκριμένα, διερευνήθηκε **η συμβολή της αυξημένης κίνησης λόγω τουρισμού, τόσο στα ατυχήματα με τραυματίες, όσο και στα θανατηφόρα ατυχήματα**. Επίσης, δημιουργήθηκαν μοντέλα συσχέτισης ατυχημάτων – αφίξεων, τα οποία προέκυψαν από τα συλλεχθέντα δεδομένα.

Για την επίτευξη των στόχων που έχουν τεθεί, χρησιμοποιήθηκαν οι βάσεις δεδομένων της Ελληνικής Στατιστικής Αρχής (ΕΛΣΤΑΤ) για την καταγραφή των αφίξεων σε λιμάνια και αεροδρόμια 38 ελληνικών νησιών, για τα οποία δεν υπάρχει πρόσβαση με διαφορετικό τρόπο,

έτσι ώστε να υπάρχει μια ακριβή εικόνα των επισκεπτών. Προκειμένου να προκύψουν ασφαλή συμπεράσματα που θα συσχετιστούν με την εποχικότητα, οι αφίξεις αναφέρονται στον **μηνιαίο** αριθμό εισιτηρίων που χρησιμοποιήθηκαν για πρόσβαση στο κάθε νησί. Ακόμα, χρησιμοποιήθηκε η βάση των αναλυτικών δεδομένων τροχαίων ατυχημάτων του ΕΜΠ για την καταγραφή των μηνιαίων, εξίσου, ατυχημάτων, όπως καταγράφηκαν από την Ελληνική Στατιστική Υπηρεσία (ΕΛΣΤΑΤ).

Με βάση τα συλλεχθέντα στοιχεία, πραγματοποιήθηκε περιγραφική ανάλυση για κάθε δείκτη και κάθε νησί μεμονωμένα. Επιπλέον, δόθηκε βάση στην ανάπτυξη και σύγκριση μοντέλων μέσω υπολογιστικού λογισμικού, τα οποία συσχετίζουν τα οδικά ατυχήματα και τους παθόντες σε αυτά με τις αφίξεις των τουριστών.

1.3 Μεθοδολογία

Σε αυτή την ενότητα παρουσιάζεται επιγραμματικά η μεθοδολογία που ακολουθήθηκε για την εκπόνηση της Διπλωματικής Εργασίας.

Αρχικά, καθορίστηκε ο **στόχος** της Διπλωματικής Εργασίας έτσι ώστε να προσδιοριστεί η πορεία της εργασίας.

Στη συνέχεια, ακολούθησε η **βιβλιογραφική ανασκόπηση**, στην οποία αναλύονται έρευνες που σχετίζονται με το αντικείμενο του τουρισμού στα νησιά και στην επιρροή αυτού στην κυκλοφορία και στα οδικά ατυχήματα.

Ακολουθεί η **συλλογή και η επεξεργασία των στοιχείων**. Τα δεδομένα των αφίξεων στα λιμάνια των νησιών αντλήθηκαν από την ΕΛΣΤΑΤ, τα αντίστοιχα δεδομένα των αεροδρομίων από την INSETE , ενώ τα στοιχεία των οδικών ατυχημάτων και δυστυχημάτων αντλήθηκαν από τη βάση δεδομένων του Τομέα Μεταφορών και Συγκοινωνιακής Υποδομής του ΕΜΠ.

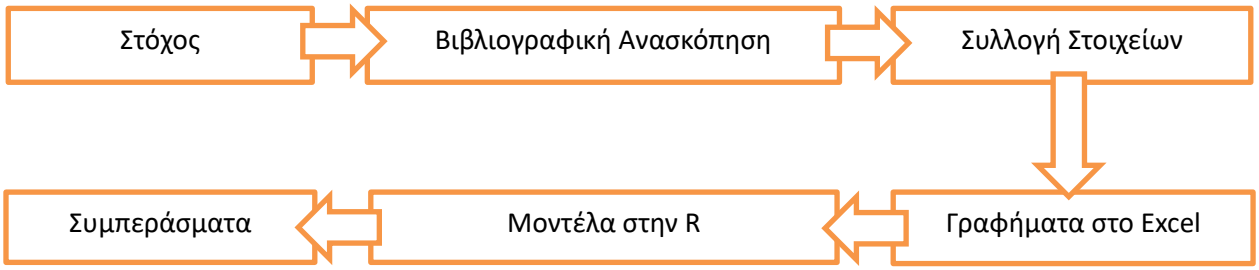
Για την πραγματοποίηση αυτού του τμήματος της διπλωματικής εργασίας, δημιουργήθηκε μία πρώτη **βάση δεδομένων** σε ένα αρχείο **Microsoft Excel**, στην οποία εισήχθησαν και αθροίστηκαν τα δεδομένα από την ΕΛΣΤΑΤ και από την INSETE. Τα μηνιαία αυτά δεδομένα, αφορούν την περίοδο από 01-01-2009 εώς 31-12-2018. Ο συνδυασμός των μηνιαίων αφίξεων με τον μόνιμο πληθυσμό έδωσε γραφήματα, από τα οποία προέκυψαν χρήσιμα συμπεράσματα για την κατάσταση κάθε νησιού.

Στη συνέχεια, δημιουργήθηκε μια δεύτερη βάση δεδομένων στην οποία, με τη βοήθεια των δεδομένων του ΕΜΠ, οργανώθηκαν σε στήλες η ημερομηνία, το νησί, τα μηνιαία ατυχήματα, τα μηνιαία δυστυχήματα και οι αφίξεις. Το συγκεκριμένο φύλλο Excel εισήχθη στην **γλώσσα**

προγραμματισμού R, η οποία με τη σειρά της υπολόγισε την ύπαρξη συσχέτισης μεταξύ των ατυχημάτων και των αφίξεων, αλλά και μεταξύ των δυστυχημάτων και των αφίξεων.

Έπειτα, αναπτύχθηκαν δύο διαφορετικά μοντέλα στην R για την πρόβλεψη των οδικών ατυχημάτων: α) **Generalized Linear Model (GLM)** και β) **Random Forest**. Το καθένα τους προέβλεψε τον αριθμό των οδικών ατυχημάτων και παθόντων σε αυτά με βάση τις τουριστικές αφίξεις για τα τελευταία δύο έτη (2017 και 2018), διαβάζοντας τις χρονοσειρές των οκτώ πρώτων ετών της βάσης δεδομένων. Με την χρήση των δεικτών **MAE** (Mean Absolute Error) και **RMSE** (Root Mean Squared), τέθηκε δυνατή η σύγκριση των δύο μοντέλων, προκειμένου να ελεγχθεί η ακρίβεια των προβλέψεων.

Τέλος, παρατίθενται τα **συμπεράσματα** που προέκυψαν από το σύνολο της διαδικασίας της Διπλωματικής Εργασίας.



1.4 Δομή Διπλωματικής Εργασίας

Παρακάτω παρουσιάζεται η δομή της Διπλωματικής Εργασίας με τα επιμέρους κεφάλαια που την απαρτίζουν.

Το **Κεφάλαιο 1** αποτελεί την εισαγωγή στο επιστημονικό πεδίο της ανάλυσης των οδικών ατυχημάτων των ελληνικών νησιών. Επιπλέον, καθορίζεται ο στόχος της εργασίας και παρουσιάζεται η μεθοδολογία που ακολουθήθηκε.

Το **Κεφάλαιο 2** αποτελεί τη βιβλιογραφική ανασκόπηση όπου παρουσιάζονται μελέτες παρόμοιου χαρακτήρα και καταγράφονται ορισμένα χρήσιμα συμπεράσματα.

Το **Κεφάλαιο 3** παρουσιάζει το θεωρητικό υπόβαθρο των στατιστικών μοντέλων που χρησιμοποιήθηκαν για τους σκοπούς της παρούσας Διπλωματικής Εργασίας.

Στο **Κεφάλαιο 4** παρουσιάζεται η διαδικασία συλλογής και επεξεργασίας των δεδομένων που χρησιμοποιήθηκαν στα μαθηματικά μοντέλα. Αρχικά, περιγράφεται η μέθοδος συλλογής στοιχείων και έπειτα η ανάλυσή τους με χρήση του προγράμματος Microsoft Excel και του στατιστικού λογισμικού R.

Το **Κεφάλαιο 5** αποτελεί την εφαρμογή της μεθοδολογίας, κατά την οποία παρουσιάζονται τα αποτελέσματα τα οποία προέκυψαν από τις αναλύσεις.

Στο **Κεφάλαιο 6** παρουσιάζονται τα τελικά συμπεράσματα που προέκυψαν από τα προηγούμενα κεφάλαια.

Στο **Κεφάλαιο 7** παρατίθενται οι βιβλιογραφικές αναφορές οι οποίες χρησιμοποιήθηκαν για την εκπόνηση της Διπλωματικής Εργασίας.

Κεφάλαιο 2: Βιβλιογραφική Ανασκόπηση

2.1 Εισαγωγή

Στο κεφάλαιο αυτό παρουσιάζεται η βιβλιογραφική ανασκόπηση που πραγματοποιήθηκε στο πλαίσιο της Διπλωματικής Εργασίας. Συγκεκριμένα, παρατίθενται έρευνες που αφορούν την επιρροή του τουρισμού στα οδικά ατυχήματα, μαζί με τις μεθοδολογίες που χρησιμοποιήθηκαν και τα συμπεράσματα που προέκυψαν.

2.2 Έρευνες

Σε αυτήν την ενότητα παρουσιάζονται έρευνες διεθνούς βιβλιογραφίας, με σκοπό να ενισχυθούν τα ευρήματα της παρούσας Διπλωματικής Εργασίας.

Το 2019, πραγματοποιήθηκε μια έρευνα σχετική με **τις επιπτώσεις του τουρισμού στα οδικά ατυχήματα στην Ελλάδα** (Bellos et. al, 2020).

Η έρευνα αυτή συλλέγει και αναλύει δεδομένα για όλα τα τροχαία ατυχήματα που καταγράφηκαν στην Ελλάδα την περίοδο 2011-2015, σύμφωνα με τη βάση δεδομένων της ΕΛΣΤΑΤ. Περιλαμβάνει μόνο ατυχήματα με τραυματισμούς ή θανάτους και όχι υλικές ζημιές. Για την ανάλυση εξετάστηκαν 39,720 ατυχήματα, με πληροφορίες για την ημερομηνία και τοποθεσία τους.

Χρησιμοποιήθηκε **αρνητική διωνυμική παλινδρόμηση**, μια στατιστική μέθοδος κατάλληλη για δεδομένα με υπερδιασπορά, για να αναλυθεί η συχνότητα των ατυχημάτων. Οι ανεξάρτητες μεταβλητές περιλαμβάνουν τον σκοπό του ταξιδιού (τουριστικός ή μη) και την περιοχή (τουριστική ή μη), ενώ εξαιρέθηκαν δεδομένα για τις μεγάλες πόλεις και τους μήνες μετάβασης (Μάιος και Οκτώβριος).

Η ανάλυση έδειξε ότι **τα ατυχήματα κορυφώνονται κατά την τουριστική περίοδο** (Ιούνιος-Σεπτέμβριος) και ότι οι τουριστικές περιοχές, κυρίως τα νησιά, παρουσιάζουν περισσότερα ατυχήματα, με τους τουρίστες να έχουν αυξημένο κίνδυνο λόγω της άγνωστης κυκλοφοριακής

κατάστασης. Τέλος, η έρευνα προτείνει πρακτικές παρεμβάσεις για τη βελτίωση της οδικής ασφάλειας, όπως αυστηρότερους ελέγχους αλκοόλ και καλύτερη σήμανση στις τουριστικές περιοχές, ιδιαίτερα κατά τους καλοκαιρινούς μήνες.

Το 2019, επίσης πραγματοποιήθηκε μια έρευνα που επικεντρώνεται στη σύνδεση **του τουρισμού και των οδικών ατυχημάτων** (Nikolaou, et al., 2019), με δεδομένα της ΕΛΣΤΑΤ για τραυματισμούς σε τροχαία ατυχήματα, εξαιρώντας αυτά με μόνο υλικές ζημιές.

Η ανάλυση χωρίστηκε σε δύο στάδια: το πρώτο ανά τύπο περιοχής (τουριστικές έναντι μη τουριστικών) και το δεύτερο ανά εθνικότητα των τραυματιών ή νεκρών (Έλληνες, ξένοι τουρίστες και μετανάστες) και χρησιμοποιήθηκε η μέθοδος εξαγόμενης έκθεσης σε κίνδυνο. Η έρευνα καταλήγει σε ενδιαφέροντα αποτελέσματα.

Ο αριθμός των θυμάτων τροχαίων ατυχημάτων **αυξάνεται το καλοκαίρι**, με σημαντική κορύφωση τον Αύγουστο στις τουριστικές περιοχές. Τα θύματα στις περιοχές αυτές είναι σχεδόν διπλάσια σε σύγκριση με άλλους μήνες. Επίσης, στις τουριστικές περιοχές, νεότερα άτομα (ηλικίας 15-44) εμπλέκονται συχνότερα σε ατυχήματα και τα δίτροχα αποτελούν την πλειοψηφία των τραυματιών (52%). Αντίστοιχα στις μη τουριστικές περιοχές, οι μεγαλύτεροι σε ηλικία άνθρωποι (45+) εμφανίζουν μεγαλύτερη συμμετοχή στα ατυχήματα και οι οδηγοί επιβατικών αυτοκινήτων αποτελούν την πλειοψηφία των τραυματιών (54%). Τέλος, η μελέτη αποκαλύπτει ότι οι ξένοι τουρίστες έχουν μεγαλύτερη πιθανότητα να εμπλακούν σε ατυχήματα, ιδιαίτερα στις μη τουριστικές περιοχές.

Συμπερασματικά, η μελέτη προτείνει μέτρα όπως αυστηρότερους ελέγχους στους δρόμους, καλύτερη σήμανση και στοχευμένες καμπάνιες οδικής ασφάλειας για τη μείωση των ατυχημάτων. Επιπλέον, αναδεικνύονται διαφορές στην οδική ασφάλεια μεταξύ τουριστικών και μη τουριστικών περιοχών και προσφέρονται χρήσιμες πληροφορίες στους υπεύθυνους λήψης αποφάσεων για τη βελτίωση των πολιτικών οδικής ασφάλειας στην Ελλάδα.

Το 2023, πραγματοποιήθηκε μια έρευνα η οποία είχε ως σκοπό την εύρεση μοτίβων ή σχέσεων ανάμεσα σε διάφορους παράγοντες που σχετίζονται με **τραυματισμούς από τροχαία ατυχήματα**, τόσο σε **νησιωτικές** όσο και σε **ηπειρωτικές** περιοχές της Ελλάδας (Ziakopoulos, et al., 2023).

Στην παρούσα μελέτη, αναλύθηκαν δεδομένα από τη βάση SANTRA, που περιλάμβανε 41,541 τραυματισμούς από τροχαία ατυχήματα, χωρισμένα σε μικρούς τραυματισμούς και θανάσιμους ή σοβαρούς τραυματισμούς. Επίσης, χρησιμοποιήθηκε ο αλγόριθμος **apriori** ώστε να ανακαλυφθούν κανόνες συσχέτισης. Τα αποτελέσματα έδειξαν τη συχνότητα με την οποία συνδέονται οι παράγοντες, όπως ο καθαρός καιρός, το αστικό οδικό περιβάλλον, οι άνδρες οδηγοί και οι ευάλωτοι χρήστες του δρόμου στα τροχαία ατυχήματα. Αυτές οι συσχετίσεις παρουσιάζονται σε υψηλές συχνότητες, συνήθως πάνω από 70% ή 80% του συνόλου των

τραυματισμών, και παρέχουν πληροφορίες για το πώς αναμένονται ορισμένα μοτίβα σε τραυματισμούς λόγω της υψηλής έκθεσης αυτών των παραγόντων.

Ωστόσο, ενδιαφέρον παρουσιάζει το γεγονός ότι το περιβάλλον των νησιών **δεν εμφάνισε σημαντικές διαφορές** στους κανόνες συσχέτισης σε σύγκριση με την ηπειρωτική Ελλάδα, υποδεικνύοντας ότι οι όποιες διαφορές απαιτούν περαιτέρω λεπτομερή ανάλυση.

Το 2004, δημοσιεύτηκε μια μελέτη που αναδεικνύει την επίδραση των τουριστών στα τροχαία ατυχήματα στην περιοχή της Κεντρικής Σκωτίας, επισημαίνοντας **διαφορές** μεταξύ ατυχημάτων που εμπλέκουν **ντόπιους** και **ξένους** οδηγούς (Walker & Page, 2004).

Η μελέτη περιλαμβάνει την ανάλυση δεδομένων από τροχαία ατυχήματα στη Σκωτία, τα οποία έχουν ληφθεί από τη βάση δεδομένων "STATS19" της Μονάδας Διερεύνησης Τροχαίων Ατυχημάτων της Κεντρικής Σκωτίας (CSRAIU). Τα δεδομένα καλύπτουν 2,841 τροχαία ατυχήματα με συνολικά 4,842 οχήματα και 7,384 θύματα στην περιοχή ευθύνης της Αστυνομίας της Κεντρικής Σκωτίας. Η βάση δεδομένων περιέχει πεδίο με ταχυδρομικούς κώδικες, που χρησιμοποιείται για να προσδιοριστεί αν οι οδηγοί είναι **“ντόπιοι”** ή **“επισκέπτες”**. Οι κάτοικοι εντός της περιοχής ευθύνης της Κεντρικής Σκωτίας χαρακτηρίζονται ως “ντόπιοι”, ενώ όσοι ζουν εκτός περιοχής χαρακτηρίζονται ως “επισκέπτες”.

Η έρευνα δείχνει ότι οι ξένοι οδηγοί εμπλέκονται στο 28% των ατυχημάτων, με υψηλότερο ποσοστό σοβαρών ή θανατηφόρων ατυχημάτων σε σύγκριση με τους ντόπιους. Τα ατυχήματα αυτά αυξάνονται κατά τις περιόδους διακοπών, με κορύφωση τους καλοκαιρινούς μήνες. Ακόμα, οι επισκέπτες τείνουν να χρησιμοποιούν περισσότερο τις κύριες οδούς λόγω άγνοιας των τοπικών οδών, κάτι που αυξάνει την πιθανότητα ατυχημάτων σε κύριες οδούς ή αυτοκινητόδρομους.

Ενδιαφέρον παρουσιάζει ότι η μελέτη υπογραμμίζει την ανάγκη για στοχευμένα μέτρα οδικής ασφάλειας, ιδιαίτερα για τους επισκέπτες, όπως εκστρατείες "Keep Left" (“Μείνετε Αριστερά”) σε πολλές γλώσσες, γεγονός που καταδεικνύει **τη δυσκολία προσαρμογής των ξένων οδηγών** στις κυκλοφοριακές συνθήκες της χώρας.

Το 1996, στη Νέα Ζηλανδία εκδόθηκε μία έρευνα στην οποία εξετάζεται αν τα ατυχήματα των τουριστών είναι **τυχαία γεγονότα** ή αν μπορούν να **προβλεφθούν** και να αποτραπούν (Page & Meyer, 1996).

Σύμφωνα με τους ερευνητές, οι τραυματισμοί των ταξιδιωτών δεν είναι τυχαία γεγονότα και μπορούν να προληφθούν με κατάλληλη εκπαίδευση και ενημέρωση. Ωστόσο, η πρόληψη των ατυχημάτων δεν διαδίδεται ευρέως στη βιομηχανία τουρισμού λόγω έλλειψης συνεργασίας μεταξύ επαγγελματιών υγείας, τουριστικών πρακτόρων και άλλων φορέων. Ένα ακόμα

σημαντικό εύρημα τους είναι ότι τα τροχαία ατυχήματα αποτελούν την **κύρια αιτία θανάτου** για Αμερικανούς που ταξιδεύουν στο εξωτερικό.

Οι επαγγελματίες υγείας μπορούν να παίξουν ρόλο στην ενημέρωση των ταξιδιωτών με πακέτα πληροφοριών για την πρόληψη ατυχημάτων, περιλαμβάνοντας ενεργητικές και παθητικές στρατηγικές. Διάφορα μέτρα, όπως η προβολή ενημερωτικών βίντεο ή η διανομή φυλλαδίων, θα μπορούσαν να βοηθήσουν στη μείωση των τροχαίων ατυχημάτων. Παρόλα αυτά, η βιομηχανία τουρισμού στην Νέα Ζηλανδία δεν θεωρεί τα τουριστικά ατυχήματα ως σοβαρό πρόβλημα, ελλείψει επαρκών δεδομένων.

Στην περίπτωση της μελέτης, εφαρμόστηκε η **μέθοδος X^2** για να συγκριθεί η κατανομή των αιτήσεων αποζημίωσης των διεθνών τουριστών με εκείνη του συνολικού πληθυσμού της Νέας Ζηλανδίας. Ο στόχος είναι να διαπιστωθεί εάν οι τουρίστες υποβάλλουν αιτήσεις με συχνότητα και διάρκεια παρόμοια με τους ντόπιους ή εάν υπάρχουν σημαντικές διαφορές. Πιο συγκεκριμένα, οι αιτήσεις αποζημίωσης από τους τουρίστες, καθώς και οι αιτήσεις του συνολικού πληθυσμού, ταξινομήθηκαν ανά κατηγορία (π.χ., ταξίδια, αθλητισμός, αναψυχή) και εξετάστηκαν σε σύγκριση με τα ετήσια δεδομένα.

Η δοκιμή X^2 εξετάζει εάν υπάρχει στατιστικά σημαντική διαφορά μεταξύ της κατανομής των αιτήσεων των τουριστών και των ντόπιων. Αυτή η ανάλυση επιτρέπει στους ερευνητές να συμπεράνουν εάν οι τουρίστες και οι ντόπιοι έχουν παρόμοια πρότυπα κατανομής στις αιτήσεις, κάτι που μπορεί να δείχνει διαφορετική έκθεση σε κίνδυνο ή διαφορετικές τάσεις υποβολής αιτήσεων αποζημίωσης.

Η ανάλυση επισημαίνει την ανάγκη για πιο λεπτομερή έρευνα σχετικά με τα τουριστικά ατυχήματα, καθώς τα υπάρχοντα δεδομένα παρέχουν μόνο μια αρχική εικόνα της κατάστασης. Τέλος, προτείνεται η βελτίωση της συνεργασίας μεταξύ των εμπλεκόμενων φορέων για τη δημιουργία ενός ασφαλούς περιβάλλοντος για τους επισκέπτες, με στόχο τη μείωση των ατυχημάτων και τη βελτίωση της εικόνας του τουρισμού στην οικονομία της χώρας.

2.3 Σύνοψη

Από τη διεθνή βιβλιογραφία προέκυψαν τα ακόλουθα συμπεράσματα:

- Τα ατυχήματα αυξάνονται σημαντικά όταν το ταξίδι πραγματοποιείται για τουρισμό ή αναψυχή σε σύγκριση με άλλους σκοπούς, ιδιαίτερα κατά την τουριστική περίοδο.
- Οι ξένοι τουρίστες εμπλέκονται σε λιγότερα ατυχήματα από τους ντόπιους σε απόλυτους αριθμούς, πιθανότατα λόγω μειωμένης έκθεσης στον κίνδυνο. Ωστόσο, η πιθανότητα

εμπλοκής τους σε ατυχήματα αυξάνεται σε τουριστικές περιοχές κατά την τουριστική περίοδο.

- Η τουριστική περίοδος (Ιούνιος-Σεπτέμβριος) σχετίζεται με αυξημένα ατυχήματα, ενδεχομένως λόγω αυξημένου κυκλοφοριακού φόρτου και άλλων παραγόντων όπως κατανάλωση αλκοόλ.
- Οι τουριστικές περιοχές εμφανίζουν αυξημένο αριθμό ατυχημάτων, κυρίως λόγω της οδήγησης σε άγνωστα περιβάλλοντα από επισκέπτες.
- Τα συνηθισμένα αίτια για τα ατυχήματα των τουριστών περιλαμβάνουν την οδήγηση από λάθος πλευρά του δρόμου, ή γενικότερα την οδήγηση σε άγνωστο περιβάλλον.

Επιπλέον, οι έρευνες για την Ελλάδα φανερώνουν ότι:

- Οι νέοι ηλικίας 15-24 και τα δίκυκλα παρουσιάζουν αυξημένα ποσοστά ατυχημάτων σε τουριστικές περιοχές.
- Οι ευάλωτοι χρήστες, όπως πεζοί και οδηγοί δίκυκλων, εμφανίζουν υψηλή ευαισθησία σε τραυματισμούς, κυρίως σε αστικά περιβάλλοντα.
- Στα νησιά, παρατηρείται υψηλότερος λόγος βαριά τραυματιών σε σχέση με την ηπειρωτική χώρα, γεγονός που οφείλεται στις ιδιαίτερες συνθήκες του οδικού περιβάλλοντος (στενοί δρόμοι, περιορισμένη ορατότητα).

Οι ερευνητές κατέληξαν σε αυτά τα συμπεράσματα, χρησιμοποιώντας χρήσιμες μεθόδους, με τις σημαντικότερες εξ αυτών να είναι η Αρνητική Διωνυμική παλινδρόμηση, η Μέθοδος Εξαγόμενης Έκθεσης σε Κίνδυνο (Induced Exposure Method) και ο Αλγόριθμος Apriori.

Οι μελέτες που έχουν διενεργηθεί μέχρι στιγμής για την διερεύνηση της επιρροής του τουρισμού στα οδικά ατυχήματα στην Ελλάδα, επικεντρώνονται κυρίως στην ανάλυση των δεδομένων ατυχημάτων και των χαρακτηριστικών τους. Η παρούσα Διπλωματική Εργασία με βάση τις αφίξεις των τουριστών επιχειρεί να ποσοτικοποιήσει την επιρροή του τουρισμού στα οδικά ατυχήματα μέσω μοντέλων συσχέτισης.

Κεφάλαιο 3: Θεωρητικό Υπόβαθρο

3.1 Εισαγωγή

Στον παρόν κεφάλαιο παρατίθεται η περιγραφή του θεωρητικού υποβάθρου σχετικά με τα στατιστικά μοντέλα που αναπτύχθηκαν στην παρούσα Διπλωματική Εργασία. Επίσης, αναφέρονται έννοιες της στατιστικής που χρησιμοποιούνται στο κεφάλαιο 5 κατά την εφαρμογή της μεθοδολογίας, μαζί με τους απαραίτητους μαθηματικούς τύπους.

3.2 Βασικές Κατανομές Στατιστικής

3.2.1 Poisson

Η κατανομή **Poisson** είναι μια διακριτή πιθανή κατανομή που περιγράφει τον αριθμό των γεγονότων που συμβαίνουν σε ένα σταθερό χρονικό διάστημα ή χώρο, όταν αυτά τα γεγονότα είναι σπάνια και ανεξάρτητα το ένα από το άλλο. Χρησιμοποιείται συχνά για να μοντελοποιήσει δεδομένα που περιγράφουν πόσες φορές συμβαίνει ένα συγκεκριμένο γεγονός.

Τα γεγονότα θεωρούνται **ανεξάρτητα** μεταξύ τους, δηλαδή η εμφάνιση ενός γεγονότος δεν επηρεάζει την πιθανότητα εμφάνισης άλλου. Ο μέσος ρυθμός εμφάνισης των γεγονότων παραμένει **σταθερός** στο δεδομένο χρονικό διάστημα ή χώρο. Ο αριθμός των γεγονότων που συμβαίνουν είναι **ακέραιος** και μπορεί να πάρει τιμές όπως 0, 1, 2...

Η πιθανότητα να συμβούν k γεγονότα σε ένα σταθερό διάστημα είναι:

$$P (X = k) = e^{-\lambda}\lambda^{k}/k!$$

όπου:

- X είναι η τυχαία μεταβλητή που αντιπροσωπεύει τον αριθμό των γεγονότων
- k είναι ο αριθμός των γεγονότων (0, 1, 2, ...)

- λ είναι η μέση τιμή (μέσος ρυθμός εμφάνισης των γεγονότων) στο διάστημα
- e είναι η βάση των φυσικών λογαρίθμων (περίπου ίση με 2.718)

Στην κατανομή Poisson η **Μέση Τιμή (Expectation)** και η **Διακύμανση (Variance)** είναι ίσες και ισούνται με λ.

3.2.2 Negative Binomial

Η κατανομή **Negative Binomial** ή αρνητική διωνυμική είναι μια διακριτή πιθανή κατανομή που χρησιμοποιείται για να περιγράψει τον αριθμό των αποτυχιών πριν συμβούν ένας προκαθορισμένος αριθμός επιτυχιών σε μια σειρά ανεξάρτητων δοκιμών Bernoulli (δοκιμών με δύο δυνατά αποτελέσματα: επιτυχία ή αποτυχία). Είναι χρήσιμη όταν η διακύμανση είναι μεγαλύτερη από τη μέση τιμή.

Τα αποτελέσματα των δοκιμών θεωρούνται **ανεξάρτητα** το ένα από το άλλο. Η κατανομή προσδιορίζει τον αριθμό των αποτυχιών που συμβαίνουν μέχρι να επιτευχθούν ένας συγκεκριμένος αριθμός επιτυχιών r. Σε κάθε δοκιμή, υπάρχει σταθερή πιθανότητα p για επιτυχία. Η πιθανότητα να συμβούν k αποτυχίες πριν από r επιτυχίες είναι:

$$P(X = k) = \binom{k + r - 1}{k} p^r (1 - p)^k$$

όπου:

- το X είναι η τυχαία μεταβλητή που αντιπροσωπεύει τον αριθμό των αποτυχιών
- το r είναι ο προκαθορισμένος αριθμός επιτυχιών
- το k είναι ο αριθμός των αποτυχιών (0, 1, 2, ...)
- το p είναι η πιθανότητα επιτυχίας σε κάθε δοκιμή
- το $\binom{k + r - 1}{k}$ είναι ο συνδυαστικός συντελεστής

Η **μέση τιμή** της κατανομής ισούται με E(X) = r (1 – p)/p και η **διακύμανση** ισούται με

$Var(X) = r\,(1 - p)/p^2$.

3.3 Συσχέτιση μεταβλητών

Η συσχέτιση μεταξύ δύο μεταβλητών δείχνει το βαθμό και την κατεύθυνση της σχέσης τους. Χρησιμοποιείται για να γίνει κατανοητό πώς μία μεταβλητή αλλάζει καθώς αλλάζει μια άλλη.

Όταν μια μεταβλητή αυξάνεται και η άλλη αυξάνεται επίσης, θεωρείται ότι υπάρχει **θετική συσχέτιση**. Αντίθετα, αν μια μεταβλητή αυξάνεται και η άλλη μειώνεται, η συσχέτιση χαρακτηρίζεται ως **αρνητική**. Αν δεν υπάρχει εμφανής σχέση μεταξύ των δύο μεταβλητών και αλλαγές σε μία δεν φαίνεται να σχετίζονται με αλλαγές στην άλλη, η συσχέτιση θεωρείται **μηδενική**.

Βέβαια, η συσχέτιση δεν συνεπάγεται **αιτιότητα**. Δηλαδή, μια ισχυρή συσχέτιση δεν σημαίνει απαραίτητα ότι η μια μεταβλητή προκαλεί την άλλη. Είναι πιθανή η ύπαρξη μίας τρίτης μεταβλητής που επηρεάζει και τις δύο.

3.4 Generalized Linear Model

Η **GLM – Generalized Linear Model** είναι μια ευέλικτη στατιστική μέθοδος που επεκτείνει τα παραδοσιακά γραμμικά μοντέλα (όπως την απλή γραμμική παλινδρόμηση), επιτρέποντάς τους να προσαρμόζονται σε δεδομένα που δεν πληρούν τις απαιτήσεις της κανονικότητας και της ομοσκεδαστικότητας. Είναι ιδιαίτερα χρήσιμο όταν τα δεδομένα δεν ακολουθούν την κανονική κατανομή και όταν το ενδιαφέρον είναι να προσομοιωθεί μια σχέση μεταξύ μιας ή περισσότερων ανεξάρτητων μεταβλητών και μιας εξαρτημένης μεταβλητής.

Τα **στοιχεία** που χαρακτηρίζουν την GLM είναι τα ακόλουθα:

1. **Συνάρτηση Κατανομής (Distribution)**: Το GLM υποστηρίζει διαφορετικές κατανομές για την εξαρτημένη μεταβλητή, όπως η κανονική, η δυαδική (binomial), η Πουασόν (Poisson) και άλλες κατανομές από την εκθετική οικογένεια (exponential family). Αυτό επιτρέπει την ανάλυση δεδομένων όπως δυαδικές εκβάσεις, μετρήσεις ή καταμετρημένα δεδομένα.
2. **Γραμμικός Συνδυασμός (Linear Predictor)**: Η συνάρτηση που συνδέει τις ανεξάρτητες μεταβλητές με την εξαρτημένη μεταβλητή είναι ένας γραμμικός συνδυασμός της μορφής:

$$n = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

όπου n είναι ο γραμμικός προβλεπτής, $\beta_0 + \beta_1, \ldots, \beta_p$ είναι οι συντελεστές των ανεξάρτητων μεταβλητών $x_1, x_2, \ldots, x_p$.

3. **Συνάρτηση Σύνδεσης (Link Function)**: Αυτή η συνάρτηση συνδέει τη μέση τιμή της εξαρτημένης μεταβλητής με τον γραμμικό συνδυασμό των ανεξάρτητων μεταβλητών. Για παράδειγμα:
    - Η λογαριθμική συνάρτηση σύνδεσης (log link) χρησιμοποιείται συνήθως για δεδομένα με κατανομή Πουασόν.
    - Η λογιστική συνάρτηση σύνδεσης (logit link) χρησιμοποιείται για δυαδικά δεδομένα (logistic regression).
    - Η ταυτοτική συνάρτηση σύνδεσης (identity link) χρησιμοποιείται για την κανονική κατανομή.

Πλεονεκτήματα του GLM:

- **Ευελιξία**: Το GLM μπορεί να εφαρμοστεί σε πολλά είδη δεδομένων που δεν μπορούν να αναλυθούν με παραδοσιακά γραμμικά μοντέλα.
- **Προσαρμογή σε μη γραμμικές σχέσεις**: Μέσω της συνάρτησης σύνδεσης, μπορεί να μοντελοποιήσει πολύπλοκες σχέσεις μεταξύ των μεταβλητών.
- **Αντιμετώπιση ετεροσκεδαστικότητας**: Δεδομένα που παρουσιάζουν μεταβαλλόμενη διακύμανση μπορούν να μοντελοποιηθούν καλύτερα με τη σωστή επιλογή κατανομής και συνάρτησης σύνδεσης.

Χρήσεις του GLM:

- **Logistic Regression** για δυαδικά δεδομένα, όπου ενδιαφέρει η πιθανότητα εμφάνισης ενός γεγονότος.
- **Poisson Regression** για δεδομένα καταμέτρησης, όπως πλήθος συμβάντων ανά μονάδα χρόνου.
- **Gamma Regression** για μοντελοποίηση δεδομένων με θετικές, συνεχείς αποκρίσεις.

Γενικότερα, τα GLM αποτελούν ένα ισχυρό εργαλείο για την ανάλυση και μοντελοποίηση δεδομένων, ιδιαίτερα όταν τα δεδομένα αποκλίνουν από τις παραδοχές των απλών γραμμικών μοντέλων.

3.5 Random Forest

Η **Random Forest** είναι μια μέθοδος μηχανικής μάθησης που χρησιμοποιείται για προβλήματα ταξινόμησης και παλινδρόμησης. Πρόκειται για έναν τύπο ενισχυμένης μάθησης (ensemble learning), όπου πολλά ανεξάρτητα μοντέλα συνδυάζονται για να βελτιώσουν την ακρίβεια των προβλέψεων. Η βασική ιδέα είναι να δημιουργηθούν πολλαπλά δέντρα αποφάσεων και να συνδυαστούν τα αποτελέσματά τους για να παραχθεί ένα πιο ισχυρό και ακριβές μοντέλο.

Ο **τρόπος λειτουργίας** της Random Forest είναι ο ακόλουθος:

1. **Δημιουργία πολλών «δέντρων αποφάσεων»**: Η Random Forest δημιουργεί πολλά διαφορετικά «δέντρα αποφάσεων» χρησιμοποιώντας δειγματοληψία με επανατοποθέτηση (bootstrap sampling) από τα δεδομένα εκπαίδευσης. Κάθε δέντρο "εκπαιδεύεται" σε ένα τυχαίο υποσύνολο των δεδομένων και χρησιμοποιεί ένα τυχαίο υποσύνολο από τις μεταβλητές σε κάθε διακλάδωση του.
2. **Δημιουργία διαφορετικών δέντρων**: Κατά τη δημιουργία κάθε δέντρου, η Random Forest επιλέγει τυχαία μόνο ένα υποσύνολο χαρακτηριστικών για κάθε κόμβο του δέντρου. Αυτό μειώνει τη συσχέτιση μεταξύ των δέντρων και βοηθά στην αποφυγή υπερεκπαίδευσης (overfitting).
3. **Συνδυασμός προβλέψεων**:
    - Στην **ταξινόμηση**, κάθε δέντρο "ψηφίζει" για μια κλάση και η τελική πρόβλεψη είναι η κλάση που λαμβάνει τις περισσότερες ψήφους (majority vote).
    - Στην **παλινδρόμηση**, η τελική πρόβλεψη είναι ο μέσος όρος των προβλέψεων από όλα τα δέντρα.

Πλεονεκτήματα του Random Forest:

- **Ανθεκτικότητα στο υπερεκπαίδευση (overfitting)**: Λόγω της τυχαίας δειγματοληψίας και της δημιουργίας πολλαπλών δέντρων, το Random Forest είναι λιγότερο πιθανό να υπερεκπαιδεύσει σε σχέση με ένα μόνο δέντρο αποφάσεων.

- **Υψηλή ακρίβεια**: Η μέθοδος συχνά έχει καλύτερη ακρίβεια από απλά μοντέλα, καθώς συνδυάζει τις προβλέψεις πολλών δέντρων.
- **Αξιολόγηση σημαντικότητας μεταβλητών**: Το Random Forest μπορεί να παρέχει πληροφορίες για τη σχετική σημασία των μεταβλητών, δείχνοντας ποιες μεταβλητές επηρεάζουν περισσότερο την πρόβλεψη.

Χρήσεις του Random Forest:

- **Ταξινόμηση**: Για προβλήματα όπου πρέπει να προβλεφθεί μια κατηγορία (π.χ., διάγνωση ασθενειών, ταξινόμηση ειδών).
- **Παλινδρόμηση**: Για προβλήματα πρόβλεψης συνεχών τιμών (π.χ., πρόβλεψη τιμής ακινήτων, πρόβλεψη ζήτησης προϊόντων).
- **Ανάλυση δεδομένων με υψηλή διάσταση**: Μπορεί να διαχειριστεί σύνολα δεδομένων με πολλούς ανεξάρτητους μεταβλητούς, δίνοντας έμφαση στις πιο σημαντικές.

Γενικά, το Random Forest είναι μια ισχυρή και αξιόπιστη μέθοδος που χρησιμοποιείται ευρέως στη μηχανική μάθηση για προβλήματα όπου η ακρίβεια και η αξιοπιστία είναι κρίσιμες.

3.6 Κριτήρια Αποδοχής Μοντέλων

3.6.1 McFadden R^2

Ο συντελεστής McFadden R^2 είναι ένας τρόπος μέτρησης της προσαρμογής ενός μοντέλου λογιστικής παλινδρόμησης (logistic regression). Είναι παρόμοιος με το R^2 που χρησιμοποιείται σε γραμμικά μοντέλα, αλλά έχει τροποποιηθεί για να ταιριάζει καλύτερα σε μη γραμμικά μοντέλα, όπως η λογιστική παλινδρόμηση. Ο McFadden R^2 ορίζεται ως εξής:

$$\mathbf{R^2_{McFadden} = 1 - \ln(L_{\mu o \nu \tau \acute{\varepsilon} \lambda o \upsilon})/\ln(L_{\mu \eta \delta \varepsilon \nu \iota \kappa o \acute{\upsilon}\ \mu o \nu \tau \acute{\varepsilon} \lambda o \upsilon})}$$

Όπου:

- $\ln(L_{\mu o \nu \tau \acute{\varepsilon} \lambda o \upsilon})$ είναι ο λογάριθμος της μέγιστης πιθανότητας (likelihood) για το προσαρμοσμένο μοντέλο.
- $\ln(L_{\mu \eta \delta \varepsilon \nu \iota \kappa o \acute{\upsilon}\ \mu o \nu \tau \acute{\varepsilon} \lambda o \upsilon})$ είναι ο λογάριθμος της μέγιστης πιθανότητας για το μοντέλο που περιλαμβάνει μόνο τη σταθερά (χωρίς άλλες ανεξάρτητες μεταβλητές).

Ο McFadden R^2 παίρνει τιμές από 0 έως 1, αλλά τυπικά είναι χαμηλότερος από το παραδοσιακό R^2 των γραμμικών μοντέλων. Τιμές από 0.2 έως 0.4 θεωρούνται ικανοποιητικές για λογιστικά μοντέλα, υποδεικνύοντας ότι το μοντέλο παρέχει μια καλή προσαρμογή στα δεδομένα. Στην παρούσα εργασία, μοντέλα με McFadden R^2 μεγαλύτερο ή ίσο του 0.1 θα θεωρούνται αποδεκτά, σε συνδυασμό με τα υπόλοιπα κριτήρια αποδοχής ων μοντέλων.

3.6.2 AICc

Το AICc (Corrected Akaike Information Criterion) είναι μια τροποποιημένη έκδοση του κριτηρίου πληροφορίας του Akaike (AIC), που χρησιμοποιείται για την αξιολόγηση και τη σύγκριση μοντέλων. Η διόρθωση (correction) αυτή λαμβάνει υπόψη το μέγεθος του δείγματος, ειδικά όταν αυτό είναι μικρό, και στοχεύει στη μείωση της μεροληψίας του AIC.

Το AIC μετρά την ποιότητα ενός στατιστικού μοντέλου με βάση την ισορροπία μεταξύ της προσαρμογής του μοντέλου στα δεδομένα και της πολυπλοκότητάς του. Ο τύπος για το AIC είναι:

$$\mathbf{AIC = -2\ln(L) + 2k}$$

όπου:

- ln(L) είναι το λογάριθμο της μέγιστης πιθανότητας (likelihood) του μοντέλου
- k είναι ο αριθμός των παραμέτρων του μοντέλου

Η διόρθωση του AICc γίνεται ως εξής:

$$\mathbf{AICc = AIC + 2k(k + 1)/(n - k - 1)}$$

όπου:

Το AICc είναι πιο αξιόπιστο όταν το μέγεθος του δείγματος είναι μικρό (δηλαδή όταν n/k είναι μικρό). Για μεγάλα δείγματα, το AICc συγκλίνει προς το AIC, οπότε η διαφορά τους γίνεται αμελητέα.

Το AICc χρησιμοποιείται για τη **σύγκριση διαφορετικών μοντέλων**: το μοντέλο με τη μικρότερη τιμή AICc θεωρείται το καλύτερο, καθώς υποδεικνύει την καλύτερη ισορροπία μεταξύ της ακρίβειας πρόβλεψης και της απλότητας.

3.6.3 Δείκτες MAE και RMSE

1. Mean Absolute Error (MAE)

Ο δείκτης **Mean Absolute Error** (MAE) μετρά τη μέση απόλυτη διαφορά μεταξύ των προβλεπόμενων και των πραγματικών τιμών. Εκφράζει, δηλαδή, πόσο απέχουν κατά μέσο όρο οι προβλέψεις από τις πραγματικές τιμές, ανεξαρτήτως του πρόσημου της διαφοράς.

Ο τύπος για τον MAE είναι:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

όπου:

- n είναι το πλήθος των παρατηρήσεων
- y_i είναι η πραγματική τιμή της i-ης παρατήρησης
- $y_i\tilde{}$ είναι η προβλεπόμενη τιμή της i-ης παρατήρησης

Ο MAE προτιμάται όταν τα σφάλματα θεωρούνται ισοδύναμα σε όλη την κλίμακα. Ένα χαμηλό MAE υποδεικνύει ότι το μοντέλο έχει καλή ακρίβεια.

2. Root Mean Squared Error (RMSE)

Ο δείκτης **Root Mean Squared Error** (RMSE) μετρά την τετραγωνική ρίζα του μέσου όρου των τετραγωνικών διαφορών μεταξύ των προβλεπόμενων και των πραγματικών τιμών. Λόγω του τετραγωνισμού των διαφορών, το RMSE δίνει μεγαλύτερο βάρος σε μεγαλύτερα σφάλματα.

Ο τύπος για τον RMSE είναι:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i-1}^{n}(y_i - \hat{y}_i)^2}$$

όπου:

- n είναι το πλήθος των παρατηρήσεων
- y_i είναι η πραγματική τιμή της i-ης παρατήρησης
- $y_i\tilde{}$ είναι η προβλεπόμενη τιμή της i-ης παρατήρησης

Ο RMSE είναι χρήσιμο όταν θέλουμε να δώσουμε μεγαλύτερη βαρύτητα στα μεγάλα σφάλματα και μας βοηθά να κατανοήσουμε πόσο σοβαρές είναι οι αποκλίσεις των προβλέψεων του μοντέλου από τις πραγματικές τιμές.

Κεφάλαιο 4: Συλλογή & Επεξεργασία Στοιχείων

4.1 Εισαγωγή

Όπως προαναφέρθηκε στο Κεφάλαιο 1, σκοπός της παρούσας Διπλωματικής Εργασίας είναι η ανάλυση των οδικών ατυχημάτων στα ελληνικά νησιά. Σε αρχικό στάδιο, επιδιώκεται η διερεύνηση της επιρροής του τουρισμού στην αύξηση των ατυχημάτων τους θερινούς μήνες όπως προκύπτει και από την βιβλιογραφία που παρουσιάστηκε στο κεφάλαιο 2. Στο παρόν κεφάλαιο περιγράφεται ο τρόπος συλλογής των απαραίτητων στοιχείων για την επίτευξη των στόχων της Διπλωματικής Εργασίας, καθώς και η κατάλληλη επεξεργασία τους.

Γίνεται μία αρχική επεξήγηση των στοιχείων που συλλέχθηκαν και της σημασίας τους και εν συνεχεία, ακολουθεί η πρώτη τους επεξεργασία με το Microsoft Excel, κατά την οποία διαμορφώνονται κατάλληλα για να εισαχθούν στην γλώσσα προγραμματισμού R.

4.2 Συλλογή Δεδομένων

Προκειμένου να προκύψει ένα ικανοποιητικό αποτέλεσμα, το οποίο θα καταδεικνύει με σαφή τρόπο τη συσχέτιση του τουρισμού με τα οδικά ατυχήματα, χρειάστηκε να ληφθεί ένα μεγάλο δείγμα ελληνικών νησιών. Συλλέχθηκε ένα σύνολο 38 ελληνικών νησιών, τα οποία κατανέμονται γεωγραφικά στις ακόλουθες ομάδες: (1) **Δωδεκάνησα** (2) **Κυκλάδες** (3) **Ιόνιο** (3) **Κεντρικό και Βόρειο Αιγαίο**. Οι ομάδες αυτές αποτελούνται από τα παρακάτω νησιά:

Δωδεκάνησα	Κυκλάδες	Ιόνιο	Κεντρικό/ Βόρειο Αιγαίο
1) Αστυπάλαια	1) Αμοργός	1) Ιθάκη	1) Άγιος Ευστράτιος
2) Κάλυμνος	2) Αντίπαρος	2) Κέρκυρα	2) Αλλόνησος
3) Κάσος	3) Άνδρος	3) Κεφαλονιά	3) Ικαρία
4) Κάρπαθος	4) Θήρα	4) Παξοί	4) Λήμνος
5) Κως	5) Ίος	5) Ζάκυνθος	5) Μυτιλήνη
6) Λέρος	6) Κύθνος		6) Σάμος
7) Νίσυρος	7) Μήλος		7) Σκιάθος
8) Πάτμος	8) Μύκονος		8) Σκόπελος
9) Ρόδος	9) Νάξος		9) Χίος
10) Τήλος	10) Πάρος		
	11) Σέριφος		
	12) Σίκινος		
	13)Σύρος		
	14) Φολέγανδρος		

Εικόνα 4.1: Το σύνολο των νησιών

Στην παρούσα Διπλωματική Εργασία συλλέχθηκαν μηνιαία δεδομένα για τη **δεκαετία 2009-2018**, με σκοπό να ληφθεί υπόψη η εποχικότητα του φαινομένου στις αναλύσεις.

Τα δεδομένα των **αφίξεων** στα λιμάνια των παραπάνω νησιών αντλήθηκαν από την ανοικτή βάση δεδομένων της Ελληνικής Στατιστικής Υπηρεσίας (**ΕΛΣΤΑΤ)**, ενώ τα δεδομένα των αφίξεων στα αεροδρόμια (για όσα νησιά διαθέτουν αερολιμένα), αντλήθηκαν από τον επίσημο διαδικτυακό ιστότοπο του Ινστιτούτου των Συνδέσμων Ελληνικών Τουριστικών Επιχειρήσεων (**ΙΝΣΕΤΕ**). Καθώς η οδική πρόσβαση στα παραπάνω νησιά δεν καθίσταται δυνατή, αθροίζοντας τα δεδομένα των λιμανιών και των αεροδρομίων, προκύπτουν οι συνολικές αφίξεις ανά μήνα. Στο σημείο σημειώνεται ότι οι αφίξεις τουριστών μέσω θαλαμηγών (yacht) δεν ελήφθησαν υπόψιν στις συνολικές αφίξεις ανά νησί για την παρούσα εργασία.

Από τον **Τομέα Μεταφορών και Συγκοινωνιακής Υποδομής του ΕΜΠ** ελήφθησαν τα μηνιαία δεδομένα των οδικών ατυχημάτων με τουλάχιστον έναν **τραυματία** για τη δεκαετία 2009-2018, τα οποία προέρχονται από την ΕΛΣΤΑΤ. Τα στοιχεία αυτά διαχωρίστηκαν σε (1) ατυχήματα, όπου προσμετρούνται οι τραυματίες και σε (2) αριθμό νεκρών. Έγινε επίσης αντιστοίχιση των δήμων στους οποίους καταγράφηκαν τα ατυχήματα με τα νησιά στα οποίοι αυτοί βρίσκονται.

4.3 Βάση Δεδομένων Διπλωματικής Εργασίας

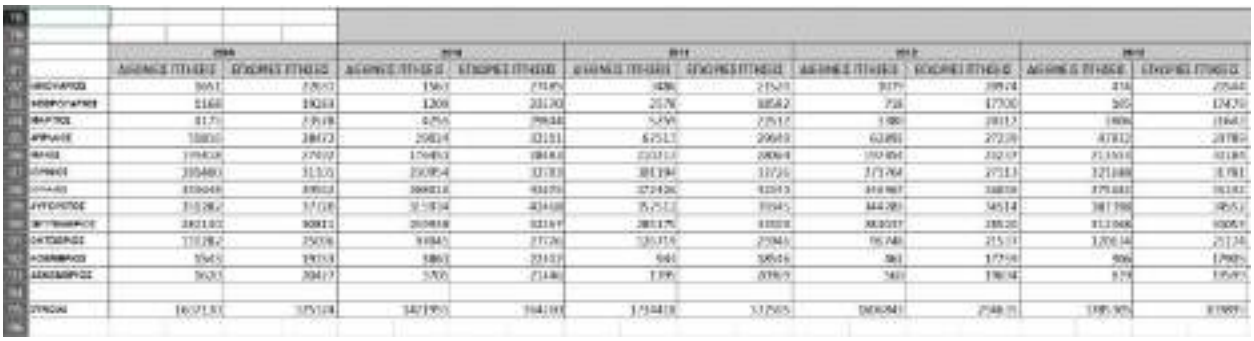
Προκειμένου να γίνει η επεξεργασία των δεδομένων που συλλέχθηκαν, δημιουργήθηκε μια ενιαία βάση δεδομένων σε ένα αρχείο του προγράμματος Excel.

Αρχικά, χρησιμοποιήθηκε ένα φύλλο Excel για κάθε νησί, ώστε να αναλυθούν ξεχωριστά τα χαρακτηριστικά τους. Το κάθε φύλλο ξεκινάει με τις αφίξεις που σημειώθηκαν στα λιμάνια, με κάθε στήλη να διαχωρίζει τους μήνες μεταξύ τους και κάθε στήλη τις χρονιές. Στις στήλες, επίσης, παρατηρούνται οι συνολικές αναχωρήσεις κάθε μήνα, αλλά και ο αριθμός των κομμένων εισιτηρίων για ΙΧ και Φορτηγά.

Παρακάτω φαίνεται το παράδειγμα της Ρόδου:

Εικόνα 4.2: Ενδεικτικά δεδομένα μηνιαίων αφίξεων στον λιμένα της Ρόδου

Στον παρακάτω πίνακα, φαίνονται τα δεδομένα των μηνιαίων αφίξεων στο αεροδρόμιο της Ρόδου:



Εικόνα 4.3: Ενδεικτικά δεδομένα μηνιαίων αφίξεων στο αεροδρόμιο της Ρόδου

Τα παραπάνω στοιχεία αθροίστηκαν σε έναν συγκεντρωτικό πίνακα, που περιγράφει τις συνολικές αφίξεις ανά μήνα και έτος για κάθε εξεταζόμενο νησί, όπως φαίνεται παρακάτω στο παράδειγμα της Ρόδου:

Ρόδος	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
ΙΑΝΟΥΑΡΙΟΣ	38782	41364	37019	30794	29375	30919	33946	37300	40832	42583
ΦΕΒΡΟΥΑΡΙΟΣ	29278	33036	29087	24637	23842	26839	27229	34226	36150	36840
ΜΑΡΤΙΟΣ	40151	48441	37403	33926	33857	30473	36011	45990	45693	55062
ΑΠΡΙΛΙΟΣ	100499	75455	112524	101889	83855	103143	109420	123607	149870	147523
ΜΑΙΟΣ	243595	218552	249957	233274	256232	276717	271121	304240	313486	373232
ΙΟΥΝΙΟΣ	330200	280884	348567	312707	366202	395127	367733	419083	447836	478346
ΙΟΥΛΙΟΣ	420545	381809	441354	405330	432190	481731	469725	549273	561901	582629
ΑΥΓΟΥΣΤΟΣ	426006	394254	426931	405332	445985	476287	491845	552682	562697	566650
ΣΕΠΤΕΜΒΡΙΟΣ	331984	302903	338327	326540	359013	360639	369618	437510	464171	480181
ΟΚΤΩΒΡΙΟΣ	150201	139931	161725	128971	157580	178516	174705	241716	270880	258042
ΝΟΕΜΒΡΙΟΣ	32105	36955	29592	26920	28477	33297	47731	43384	52126	51870
ΔΕΚΕΜΒΡΙΟΣ	32114	36081	32130	29088	28691	33991	39147	38251	41228	41815

Εικόνα 4.4: Ενδεικτικά δεδομένα μηνιαίων αφίξεων στη Ρόδο

Τέλος, όσον αφορά τα ατυχήματα και τον αριθμό των νεκρών, δημιουργήθηκε ένας νέος πίνακας. Προετοιμάζοντας τα δεδομένα για τη γλώσσα προγραμματισμού R, όλα τα χρονικά στοιχεία τοποθετήθηκαν στις γραμμές το ένα κάτω από το άλλο, με άλλες δύο στήλες, τα «ατυχήματα» και οι «νεκροί», να συμπληρώνουν τον πίνακα:

225	Ρόδος			
226			Ατυχήματα	Νεκροί
227		2009 Ιανουάριος	9	2
228		2009 Φεβρουάριος	7	1
229		2009 Μάρτιος	5	
230		2009 Απρίλιος	8	
231		2009 Μάιος	8	1
232		2009 Ιούνιος	22	6
233		2009 Ιούλιος	16	4
234		2009 Αύγουστος	19	2
235		2009 Σεπτέμβριος	12	4
236		2009 Οκτώβριος	10	
237		2009 Νοέμβριος	1	
238		2009 Δεκέμβριος	8	
239		2010 Ιανουάριος	12	2
240		2010 Φεβρουάριος	9	2
241		2010 Μάρτιος	10	
242		2010 Απρίλιος	20	1
243		2010 Μάιος	18	2
244		2010 Ιούνιος	24	3
245		2010 Ιούλιος	30	1
246		2010 Αύγουστος	29	3
247		2010 Σεπτέμβριος	25	5
248		2010 Οκτώβριος	19	4

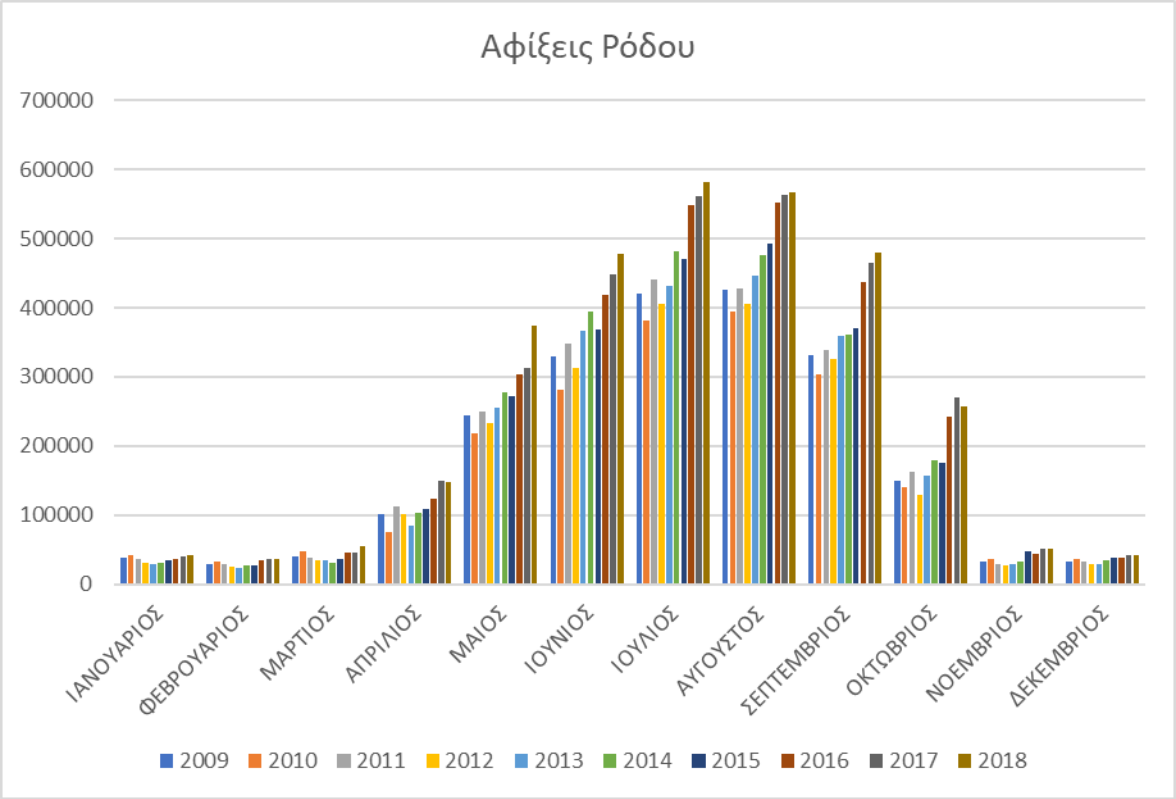
Εικόνα 4.5: Ενδεικτικά δεδομένα ατυχημάτων και θανάτων στη Ρόδο

4.4 Περιγραφική Ανάλυση

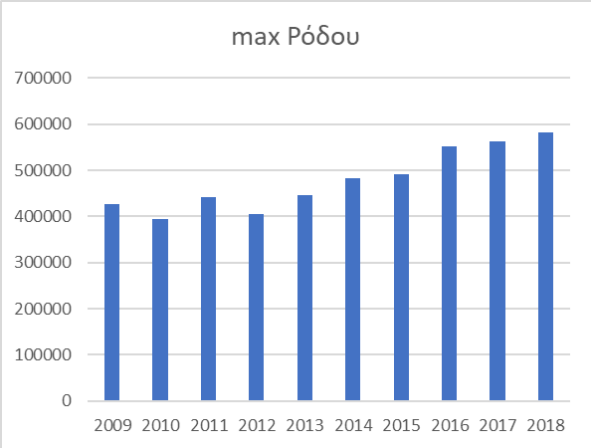
Με τη βοήθεια της βάσης δεδομένων που δημιουργήθηκε στα υπολογιστικά φύλλα του Excel, πραγματοποιήθηκε η περιγραφική ανάλυση των δεδομένων για κάθε νησί που εξετάστηκε στην παρούσα Διπλωματική Εργασία. Ωστόσο, προκειμένου να προκύψουν ασφαλέστερα συμπεράσματα, τα νησιά χωρίστηκαν σε **4 ομάδες** με βάση την τοποθεσία τους. Οι ομάδες αποτελούνται από τα **Δωδεκάνησα**, τις **Κυκλάδες**, το **Ιόνιο Πέλαγος** και το **Βόρειο/Κεντρικό Αιγαίο**. Παρακάτω παρουσιάζονται συγκεντρωτικά γραφήματα για 4 νησιά με μεγάλο αριθμό αφίξεων από καθεμία από τις 4 ομάδες.

4.4.1 Το παράδειγμα της Ρόδου

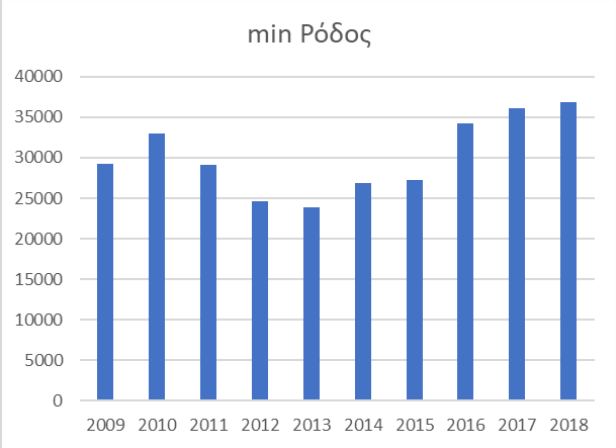
Η Ρόδος αποτελεί το νησί των Δωδεκανήσων που επιλέχθηκε να παρουσιαστεί, καθώς είναι το μεγαλύτερο της συγκεκριμένης ομάδας, αλλά και επειδή δέχεται τον μεγαλύτερο αριθμό αφίξεων κατά τη διάρκεια του έτους. Παρατηρείται, επίσης, μεγάλη διαφορά στον αριθμό των θανάτων και των ατυχημάτων.



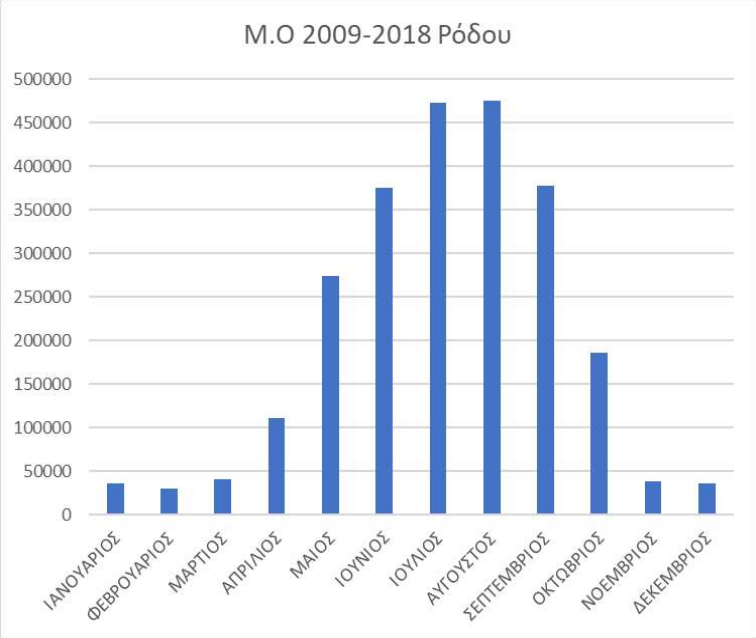
Πίνακας 4.1: Οι μηνιαίες αφίξεις της δεκαετίας για τη Ρόδου



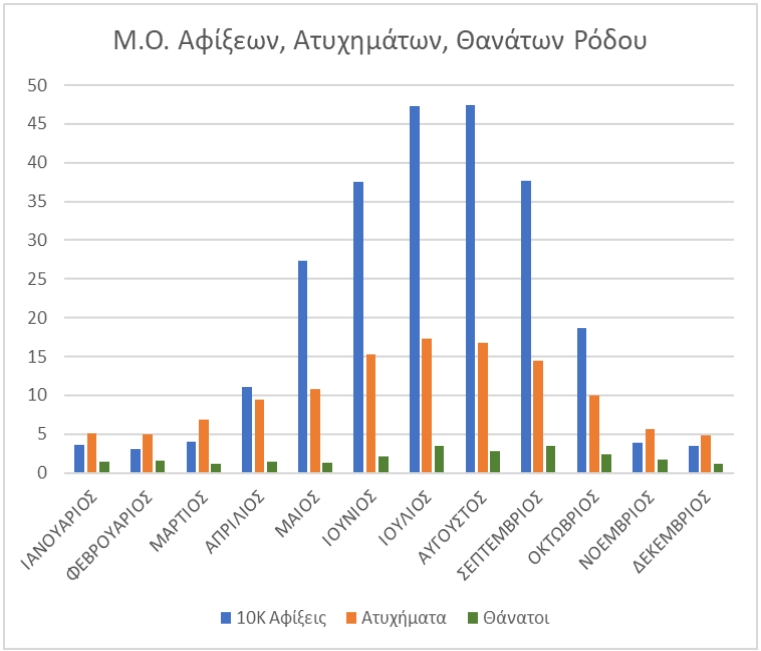
Πίνακας 4.2: Μέγιστες μηνιαίες αφίξεις ανά έτος (Ρόδος)



Πίνακας 4.3: Ελάχιστες μηνιαίες αφίξεις ανά έτος (Ρόδος)



Πίνακας 4.4: Μ.Ο. μηνιαίων αφίξεων δεκαετίας για Ρόδο



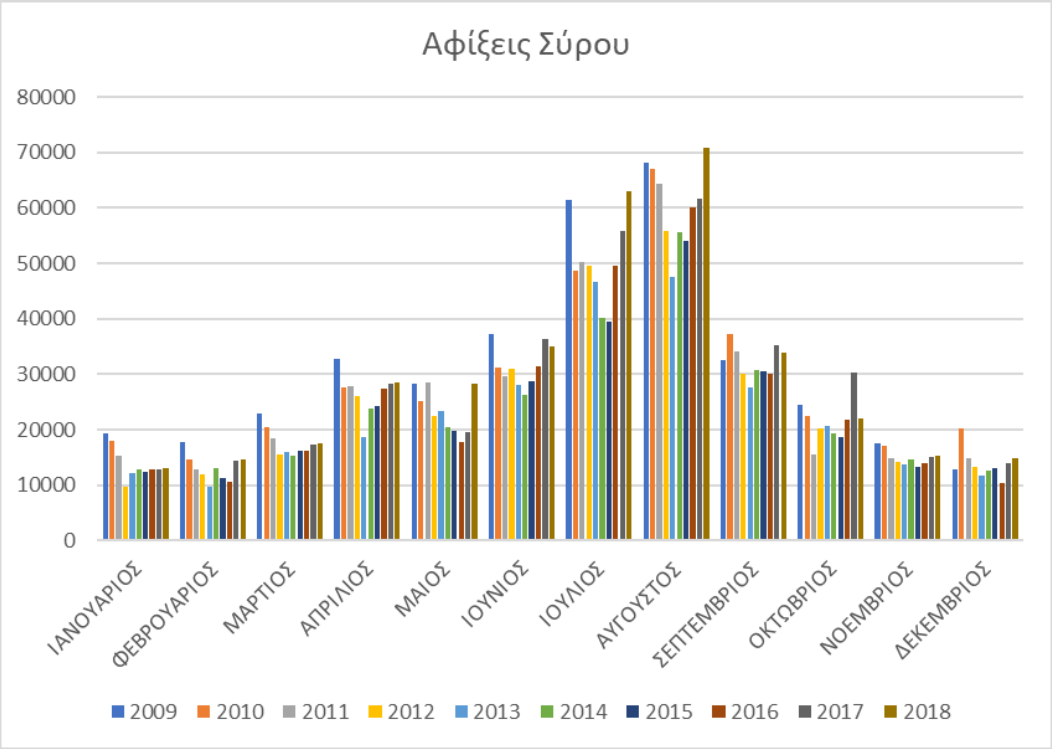
Πίνακας 4.5: Μ.Ο. μηνιαίων αφίξεων, ατυχημάτων, θανάτων δεκαετίας για Ρόδο

Από τα γραφήματα προκύπτει ότι οι αφίξεις στο νησί αυξάνονται την περίοδο Μαΐου – Οκτωβρίου, με τον μέγιστο αριθμό των αφίξεων να παρατηρείται τον Ιούλιο και τον Αύγουστο. Τον Απρίλιο, επίσης, παρατηρούνται περισσότερες αφίξεις σε σχέση με τους χειμερινούς μήνες, αλλά σημαντικά λιγότερες από σε σχέση με την περίοδο Μαΐου – Οκτωβρίου. Με το πέρασμα των ετών, οι αφίξεις των τουριστών αυξάνονται.

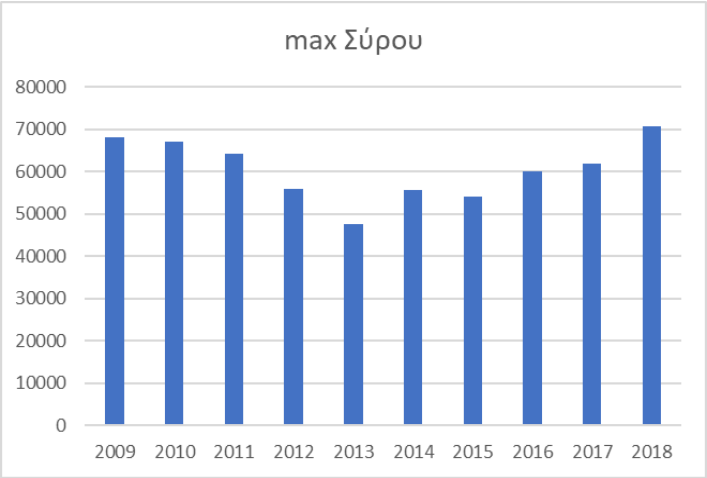
Τα τελευταία τρία έτη (2016-2018) οι αφίξεις είναι σημαντικά αυξημένες σε σχέση με το 2009, ενώ την τριετία 2012-2014 παρατηρούνται τα μικρότερα νούμερα της δεκαετίας, προτού μεγαλώσουν ξανά. Ο αριθμός των ατυχημάτων και των νεκρών φαίνεται να ακολουθεί την τάση των αφίξεων των τουριστών και αυξάνεται σημαντικά τους μήνες Ιούλιο και Αύγουστο.

4.4.2 Το παράδειγμα της Σύρου

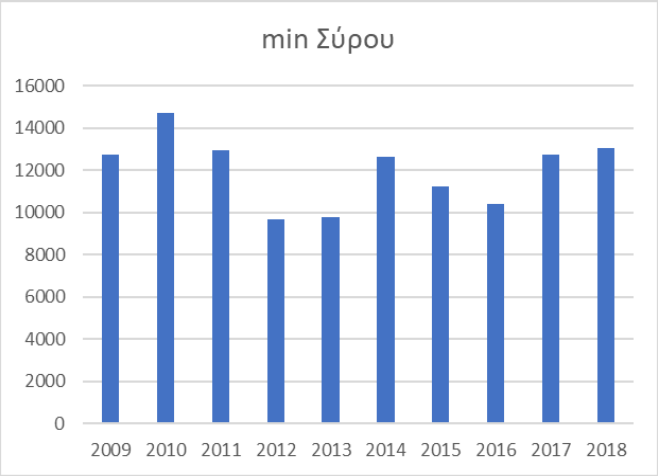
Από την ομάδα των Κυκλάδων, το νησί που επιλέχθηκε να παρουσιαστεί είναι η Σύρος καθώς, παρ' όλο που δεν είναι το μεγαλύτερο της συγκεκριμένης περιοχής, παρουσιάζει πολύ μεγάλο αριθμό αφίξεων για τη δεκαετία που εξετάστηκε, οπότε τα αποτελέσματα συγκεντρώνουν ενδιαφέρον.



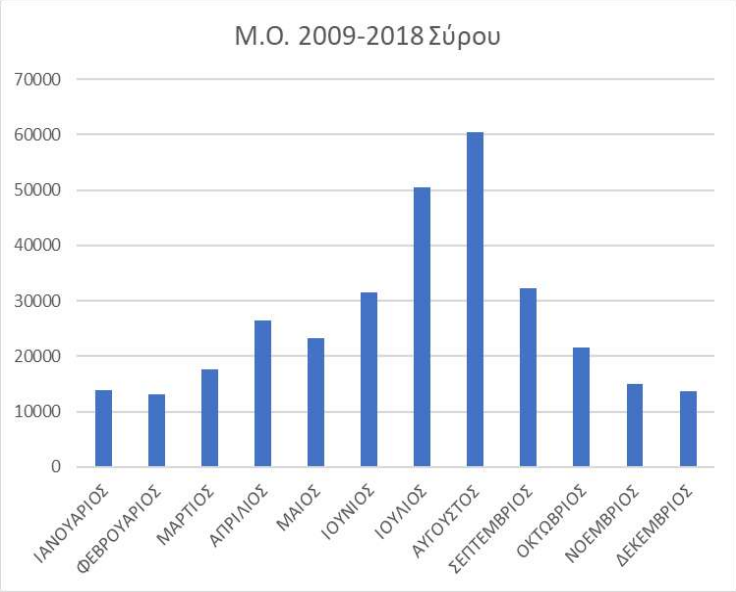
Πίνακας 4.6: Οι μηνιαίες αφίξεις της δεκαετίας για τη Σύρο



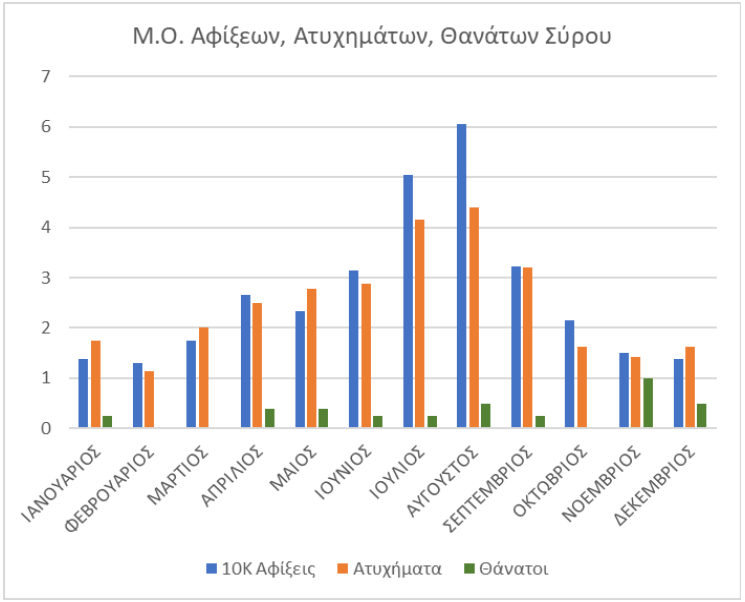
Πίνακας 4.7: Μέγιστες μηνιαίες αφίξεις ανά έτος (Σύρος)



Πίνακας 4.8: Ελάχιστες μηνιαίες αφίξεις ανά έτος (Σύρος)



Πίνακας 4.9: Μ.Ο. μηνιαίων αφίξεων δεκαετίας για Σύρο

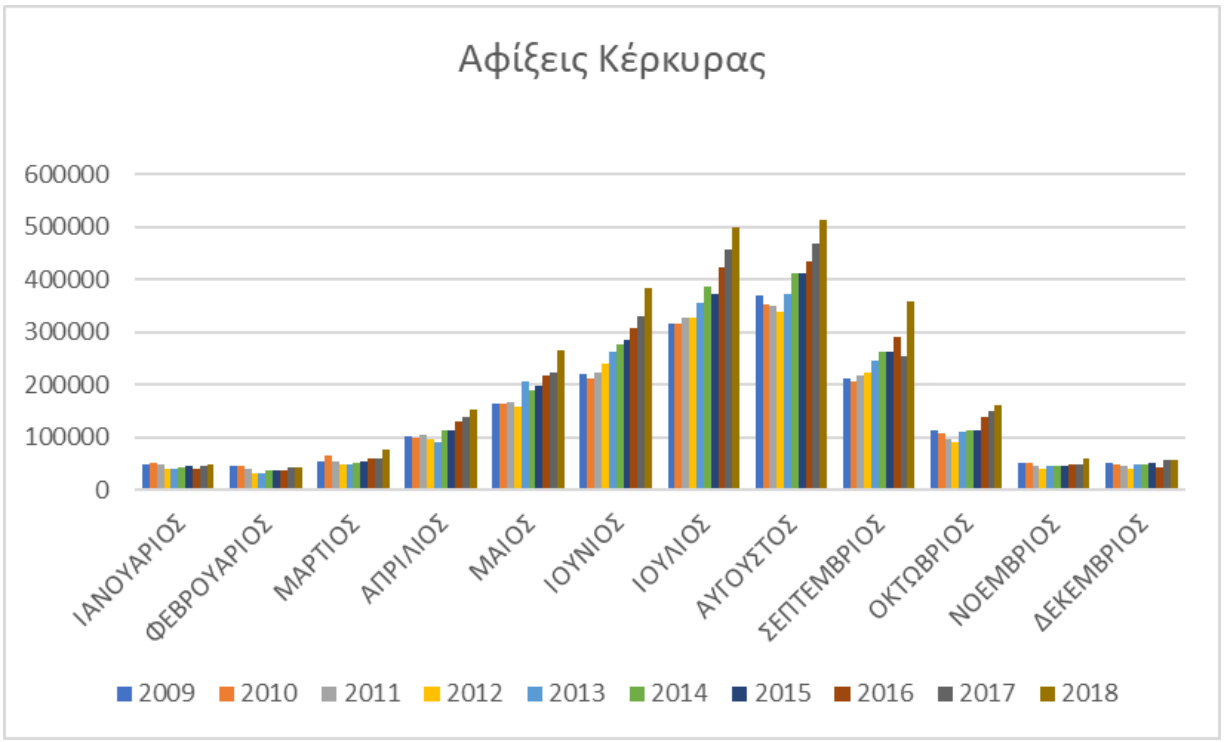


Πίνακας 4.10: Μ.Ο. μηνιαίων αφίξεων, ατυχημάτων, θανάτων δεκαετίας για Σύρο

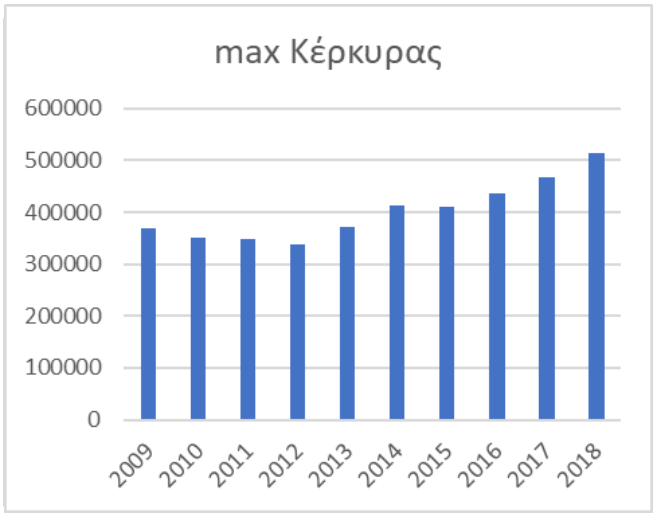
Τα γραφήματα δείχνουν αύξηση των αφίξεων την περίοδο Μαρτίου – Αυγούστου που ακολουθείτε από μία σταδιακή μείωση στους μήνες του Σεπτεμβρίου και Οκτωβρίου. Η επισκεψιμότητα του Ιουλίου και του Αυγούστου είναι εμφανώς πολύ μεγαλύτερη από αυτήν που παρατηρείται τους υπόλοιπους μήνες. Την περίοδο 2009 – 2013 φαίνεται μια σταδιακή μείωση στις αφίξεις, που ακολουθείται από μια αντίστοιχη αύξηση την περίοδο 2014 – 2018 και επιστροφή στα νούμερα του 2009. Ο αριθμός των ατυχημάτων, όπως και των αφίξεων, αυξάνεται τους καλοκαιρινούς μήνες.

4.4.3 Το παράδειγμα της Κέρκυρας

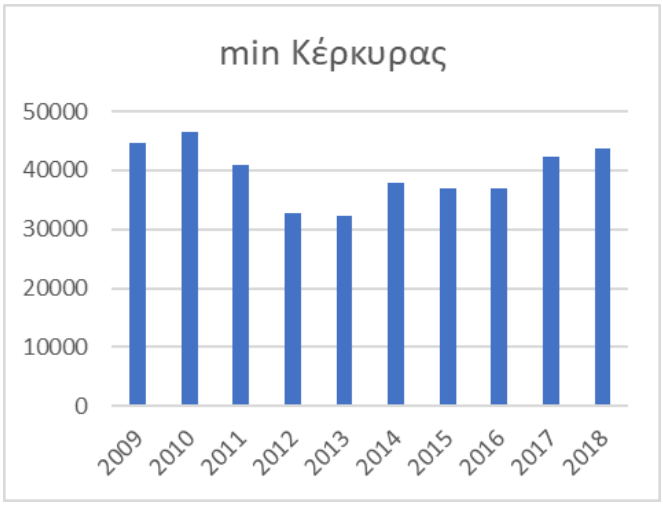
Η Κέρκυρα είναι το νησί του Ιονίου που επιλέχθηκε να παρουσιαστεί από αυτά που εξετάστηκαν, καθώς πέρα από τις περισσότερες αφίξεις, παρουσίασε και έναν μεγάλο αριθμό θανάτων και ατυχημάτων.



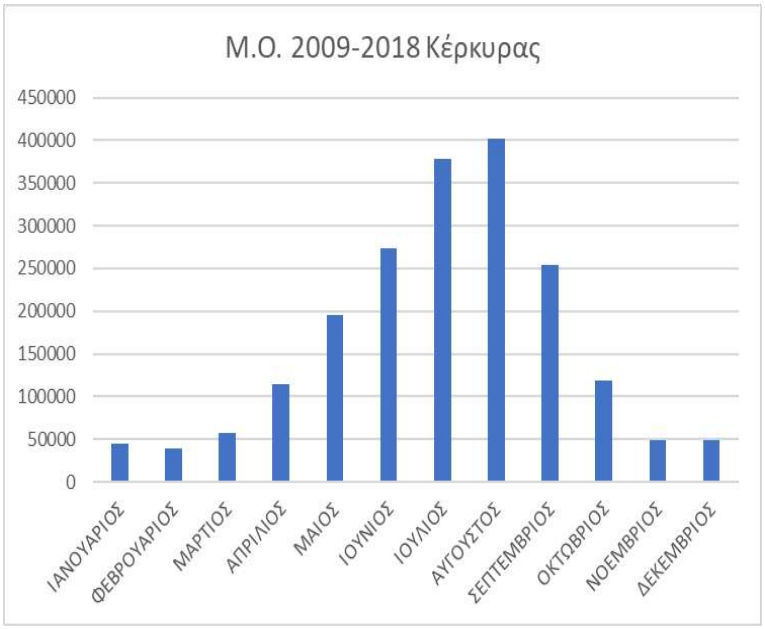
Πίνακας 4.11: Οι μηνιαίες αφίξεις της δεκαετίας για τη Κέρκυρα



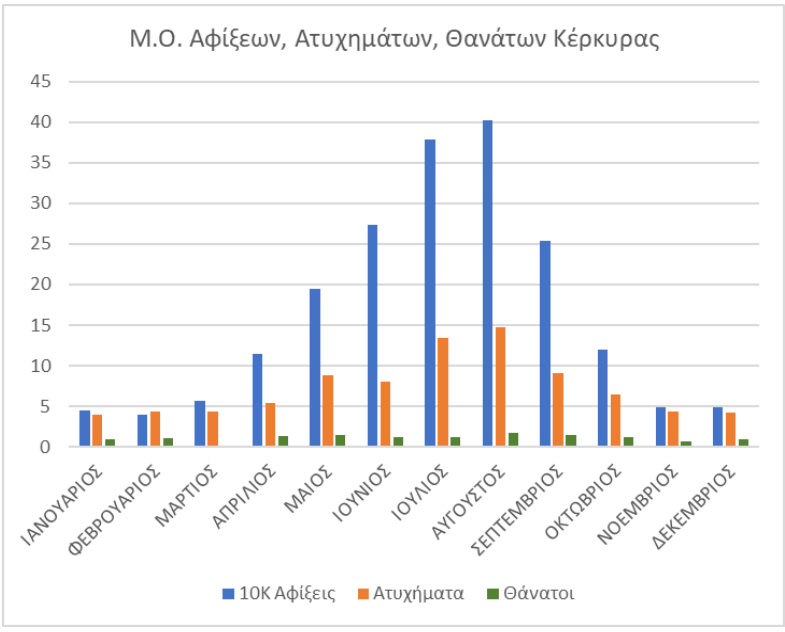
Πίνακας 4.12: Μέγιστες μηνιαίες αφίξεις ανά έτος (Κέρκυρα)



Πίνακας 4.13: Ελάχιστες μηνιαίες αφίξεις ανά έτος (Κέρκυρα)



Πίνακας 4.14: Μ.Ο. μηνιαίων αφίξεων δεκαετίας για Κέρκυρα

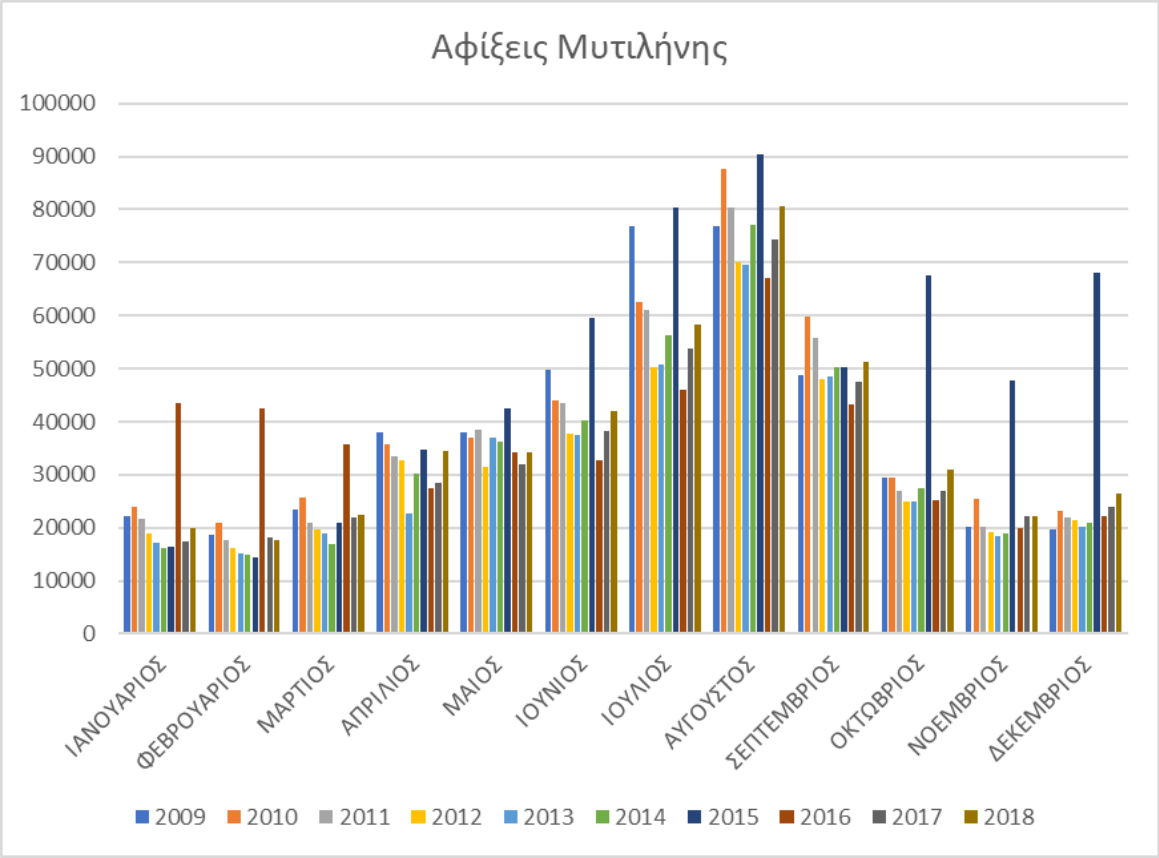


Πίνακας 4.15: Μ.Ο. μηνιαίων αφίξεων, ατυχημάτων, θανάτων δεκαετίας για Κέρκυρα

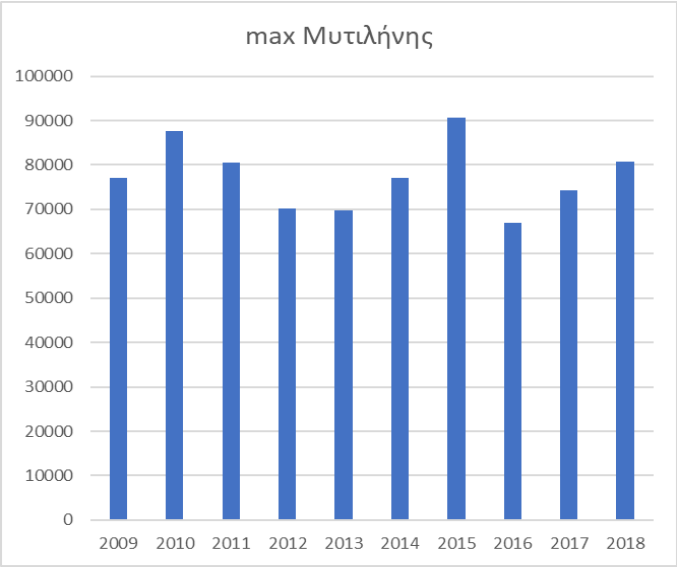
Τα γραφήματα δείχνουν σταθερή αύξηση στις αφίξεις την περίοδο Απριλίου – Αυγούστου και μείωση την περίοδο Σεπτεμβρίου – Νοεμβρίου, ενώ οι χειμερινοί μήνες παραμένουν σε σταθερά επίπεδα. Επίσης, παρατηρείται σταδιακή αύξηση στις ετήσιες αφίξεις σε όλη την εξεταζόμενη περίοδο. Την τάση των αφίξεων ακολουθούν οι αριθμοί των μηνιαίων ατυχημάτων και θανάτων, παρουσιάζοντας αύξηση τους ίδιους μήνες.

4.4.4 Το παράδειγμα της Μυτιλήνης

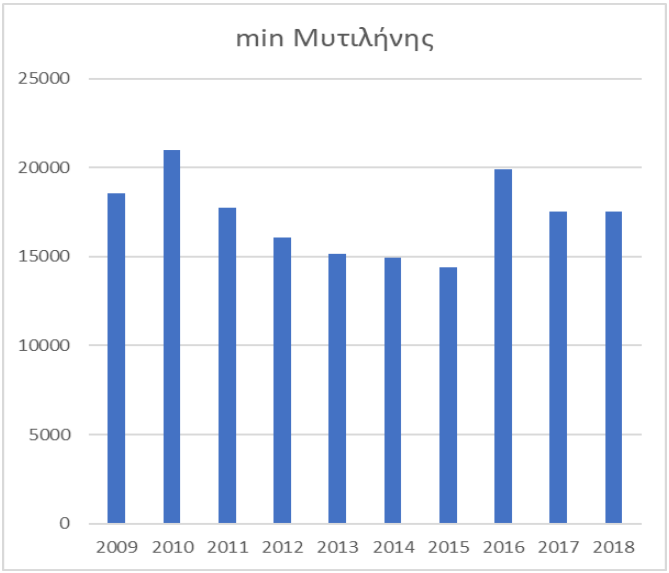
Ακολουθούν τα στατιστικά δεδομένα της Μυτιλήνης, η οποία είναι το νησί που επιλέχθηκε να αντιπροσωπεύσει την ομάδα του Κεντρικού και Βορείου Αιγαίου. Σε σχέση με τα άλλα νησιά που εξετάστηκαν, αυτό κατείχε το μεγαλύτερο αριθμό αφίξεων, θανάτων αλλά και ατυχημάτων.



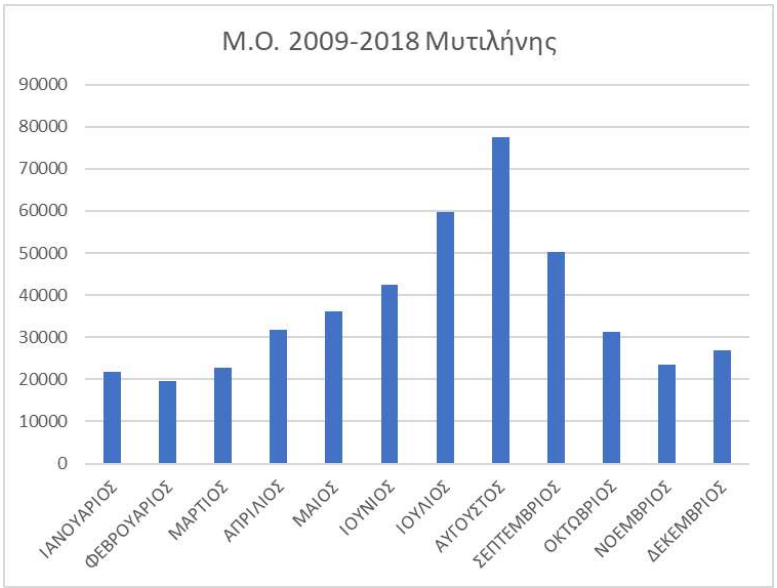
Πίνακας 4.16: Οι μηνιαίες αφίξεις της δεκαετίας για τη Μυτιλήνη



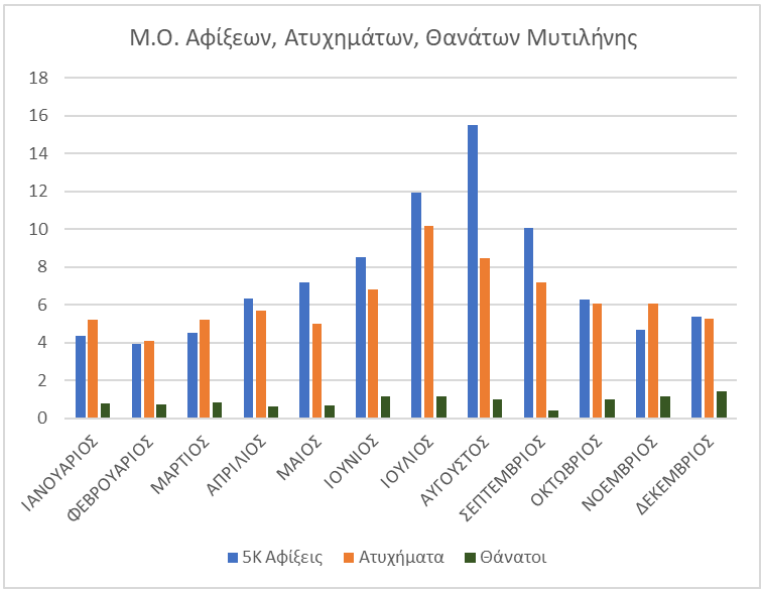
Πίνακας 4.17: Μέγιστες μηνιαίες αφίξεις ανά έτος (Μυτιλήνη)



Πίνακας 4.18: Ελάχιστες μηνιαίες αφίξεις ανά έτος (Μυτιλήνη)



Πίνακας 4.19: Μ.Ο. μηνιαίων αφίξεων δεκαετίας για Μυτιλήνη



Πίνακας 4.20: Μ.Ο. μηνιαίων αφίξεων, ατυχημάτων, θανάτων δεκαετίας για Μυτιλήνη

Τα γραφήματα δείχνουν αύξηση των αφίξεων τους καλοκαιρινούς μήνες, με το τρίμηνο του Ιουλίου – Σεπτεμβρίου να είναι αυτό στο οποίο εμφανίζεται η μέγιστη επισκεψιμότητα του νησιού. Δεν φαίνεται να υπάρχει κάποια ιδιαίτερη τάση όσον αφορά τις ετήσιες αυξήσεις, καθώς τα έτη στα οποία εμφανίστηκαν οι περισσότερες αφίξεις ήταν τα 2009 και 2015, με το 2010 να έχει τον δεύτερο μεγαλύτερο μηνιαίο αριθμό αφίξεων και τον μεγαλύτερο από τους ελάχιστους.

Κεφάλαιο 5: Εφαρμογή Μεθοδολογίας

5.1 Εισαγωγή

Στο παρόν κεφάλαιο παρουσιάζεται η αναλυτική περιγραφή της μεθοδολογίας και των αποτελεσμάτων που προέκυψαν από αυτήν. Αφότου συλλέχθηκαν τα δεδομένα που αναλύθηκαν στο Κεφάλαιο 4, διαμορφώθηκαν κατάλληλα ώστε να εισαχθούν στην γλώσσα προγραμματισμού R.

Στόχος ήταν η εύρεση μοντέλων τα οποία αφενός να καταδεικνύουν την επιρροή του τουρισμού στα οδικά ατυχήματα και αφετέρου, να προβλέπουν τα οδικά ατυχήματα και τους θανάτους, βάσει των ήδη υπαρχόντων στοιχείων.

5.2 Μετατροπή Δεδομένων με το Excel

Αρχικά, εισήχθησαν όλα τα απαραίτητα δεδομένα σε ένα φύλλο του προγράμματος Excel, στην ακόλουθη μορφή:

1η στήλη: Ημερομηνία (**Date**)

2η στήλη: Νησιά (**Islands**)

3η στήλη: Νεκροί σε ατυχήματα (**Fatalities**)

4η στήλη: Ατυχήματα (**Crashes**)

5η στήλη: 1000 Αφίξεις (**Arrivals**)

Ο λόγος που προτιμήθηκε αυτή η μορφή είναι ο τρόπος ανάγνωσης και ανάλυσης των δεδομένων από την R.

Τα νησιά διαχωρίστηκαν ανάλογα την ομάδα στην οποία ανήκουν, γεγονός που οδήγησε στην δημιουργία 4 διαφορετικών φύλλων Excel. Σκοπός του εγχειρήματος αυτού είναι η σύγκριση των χαρακτηριστικών των ομάδων μεταξύ τους. Τα 4 αρχεία ονομάστηκαν: **(1) Dodekanisa.xlsx (2) Kyklades.xlsx (3) Ionio.xlsx (4) Kentriko&BoreioAigaio.xlsx**

Τα χαρακτηριστικά προκύπτουν από την επιλογή του μοντέλου που χρησιμοποιήθηκε.

5.3 Εφαρμογή Generalized Linear Model (GLM)

Αρχικά, γίνεται εγκατάσταση των **βιβλιοθηκών** που είναι απαραίτητες στην επιτέλεση του κώδικα. Αυτά είναι το “readxl”, που είναι υπεύθυνο για την ανάγνωση των αρχείων Excel του υπολογιστή, το “dplyr”, το οποίο διαχειρίζεται και επεξεργάζεται στα δεδομένα και η “MASS”, η οποία χρησιμοποιεί το πακέτο της Αρνητικής Διωνυμικής Κατανομής . Με την εντολή “library” γίνεται η φόρτωση των βιβλιοθηκών.

Έπειτα, γίνεται **διαχωρισμός των δεδομένων**, ώστε να καθιστεί δυνατή η πρόβλεψη από το μοντέλο. Με τις εντολές train_end_date <- as.Date("2016-12-31") και test_start_date <- as.Date("2017-01-01"), το μοντέλο προβλέπει από τον **Ιανουάριο του 2017 εώς το Δεκέμβριο του 2018**, με ότι δεδομένο έχει συλλέξει **μέχρι τον Δεκέμβριο του 2016**. Αντίστοιχα, με τις εντολές train_data <- db[db$Date <= train_end_date,] και test_data <- db[db$Date > train_end_date,], δημιουργούνται σύνολα δεδομένων, τόσο για την “εκπαιδευόμενη” περίοδο, όσο και για τη “δοκιμαστική”. Επιπλέον, ελέγχεται και για τις δύο περιόδους ο αριθμός των γραμμών από τις οποίες αποτελούνται.

Στο επόμενο βήμα, επιλέγεται το κατάλληλο μοντέλο, σύμφωνα με τα δεδομένα των στηλών “Fatalities” και “Crashes”, δηλαδή ένα εκ των **Poisson** και **Negative Binomial**. Με την εντολή mesos <- mean() υπολογίζεται ο μέσος όρος, με την diak <- var() υπολογίζεται η διακύμανση και με την if (diak > mesos) print("Negative Binomial") else print("Poisson") τυπώνεται η κατάλληλη επιλογή μοντέλου. Εφόσον η διακύμανση, είτε στα Fatalities, είτε στα Crashes, είναι μεγαλύτερη από τον μέσο όρο, τότε στον κώδικα θα τυπωθεί η φράση “Negative Binomial”.

Σε αυτό το σημείο ελέγχθηκε το αποτέλεσμα της τελευταίας εντολής για τα Fatalities και τα Crashes όλων των ομάδων νησιών και σε κάθε περίπτωση τυπώνεται η φράση “Negative Binomial”.

Ξεκινώντας από τα Fatalities, το μοντέλο εκπαιδεύεται για αρνητική διωνυμική κατανομή με την εντολή negb <- glm.nb(Fatalities ~ Arrivals, data = train_data), η οποία αντλεί στοιχεία από

την “εκπαιδευόμενη” περίοδο, προκειμένου να βρεθεί ο **βαθμός συσχέτισης** μεταξύ των Fatalities και Arrivals.

Οι προβλέψεις των θανατηφόρων ατυχημάτων για την διετία 2017-2018 επιτυγχάνονται με την εντολή test_data$Predicted_Fatalities <- predict(negb, newdata = test_data, type = "response").

Προκειμένου να κριθεί επιτυχημένο το μοντέλο, χρειάζεται να αξιολογηθεί από τους δείκτες **Μέσου Απόλυτου Σφάλματος (MAE)** και **Τετραγωνικού Μέσου Σφάλματος (RMSE)**.

Η ίδια διαδικασία ακολουθείται, εξίσου, για τα Crashes, δηλαδή το μοντέλο εκπαιδεύεται για την αρνητική διωνυμική κατανομή και συνεχίζει με τις προβλέψεις για τα ατυχήματα για την περίοδο 2017-2018.

5.3.1 Εφαρμογή GLM σε όλα τα νησιά

Το μοντέλο εφαρμόζεται αρχικά για όλα τα νησιά, ώστε να ελεγχθεί ο βαθμός συσχέτισης αφίξεων με θανάτους και ατυχήματα σε μεγαλύτερη κλίμακα. Από την ανάλυση προέκυψε ότι το μοντέλο χρησιμοποιεί γραμμές **3648** για εκπαίδευση και **912** γραμμές για τις προβλέψεις του.

Ξεκινώντας από τα αποτελέσματα των **Fatalities**, το μοντέλο δείχνει ότι ο συντελεστής συσχέτισης για τις αφίξεις είναι 0.00978 ($p < 0.001$), που σημαίνει ότι **η αύξηση των αφίξεων συσχετίζεται με αύξηση των θανάτων κατά 0,98%** . Επίσης, ο δείκτης **McFadden R^2** ισούται με **0.12**, που δείχνει τη ότι το συγκεκριμένο μοντέλο εξηγεί το 12% της διακύμανσης της εξαρτημένης μεταβλητής, και ο συντελεστής **AICc** ισούται με **3011.21**, δείχνοντας την καταλληλότητα του μοντέλου.

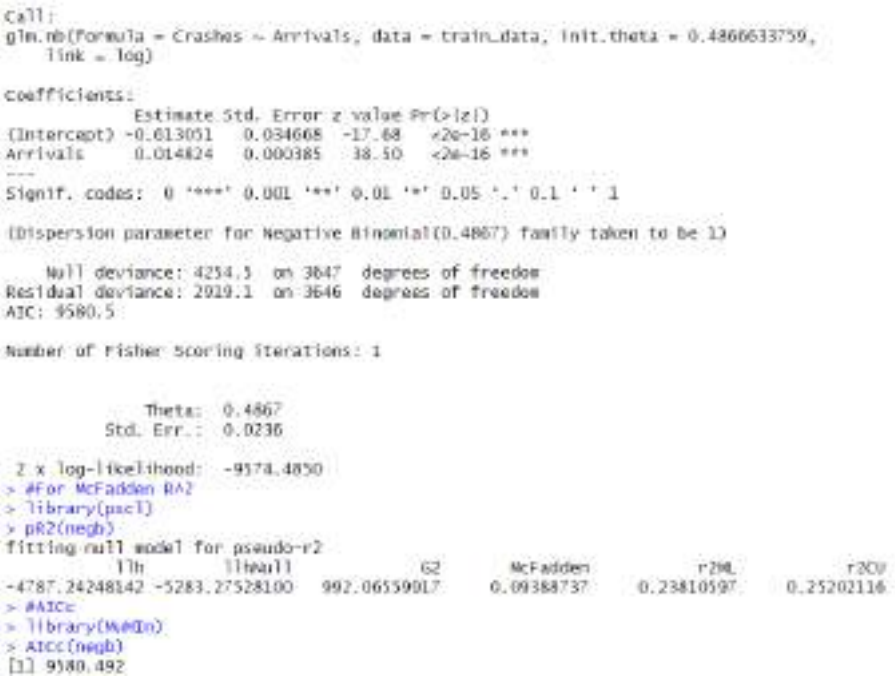
Παρακάτω φαίνονται τα ακριβή και αναλυτικά αποτελέσματα της αρνητικής διωνυμικής κατανομής που προέκυψαν για τους θανάτους.

Εικόνα 5.1: Αποτελέσματα GLM για θανάτους σε όλα τα νησιά

Στη συνέχεια, το μοντέλο προέβλεψε τους θανάτους για την περίοδο 2017-2018 και δημιούργησε έναν πίνακα **πραγματικών θανάτων – προβλεπόμενων θανάτων**. Τα σφάλματα για την πρόβλεψη θανάτων στην περίοδο δοκιμής ήταν:

- Mean Absolute Error (MAE): 0.41
- Root Mean Square Error (RMSE): 1.67

Ακολουθώντας τα ίδια βήματα για τα **Crashes**, ο συντελεστής για τις αφίξεις είναι 0.0148 ($p < 0.001$), που υποδηλώνει, όπως και στην περίπτωση των Fatalities, ότι **περισσότερες αφίξεις σχετίζονται με περισσότερα ατυχήματα κατά 1.49%**. Το μοντέλο έχει **McFadden R^2** που ισούται με **0.10** και το **AICc** είναι **9580.49**, εξηγώντας τη διακύμανση σε σχέση με τους θανάτους, αλλά όχι σε τόσο ικανοποιητικό βαθμό.



Εικόνα 5.2: Αποτελέσματα GLM για ατυχήματα σε όλα τα νησιά

Όπως και στην προηγούμενη περίπτωση, το μοντέλο δημιούργησε έναν πίνακα **πραγματικών ατυχημάτων – προβλεπόμενων ατυχημάτων** και τα σφάλματα προέκυψαν:

- Mean Absolute Error (MAE): 19.66
- Root Mean Square Error (RMSE): 177.93

Ερμηνεία των Αποτελεσμάτων:

- Οι θετικοί συντελεστές για τις Αφίξεις και στα δύο μοντέλα υποδεικνύουν ότι καθώς αυξάνεται ο αριθμός των αφίξεων στα νησιά, αυξάνονται και οι θάνατοι και τα τροχαία ατυχήματα.
- Το AIC και το R^2 του McFadden δείχνουν μια ικανοποιητική προσαρμογή του μοντέλου, αν και οι τιμές R^2 είναι σχετικά χαμηλές, γεγονός που υποδεικνύει ότι οι αφίξεις εξηγούν μόνο ένα μέρος της μεταβλητότητας στους θανάτους.
- Το Μέσο Απόλυτο Σφάλμα (MAE) για τους θανάτους είναι 0.41, και το Τετραγωνικό Μέσο Σφάλμα (RMSE) είναι 1.67, υποδηλώνοντας ότι το μοντέλο προβλέπει ικανοποιητικά.
- Η προσαρμογή του μοντέλου είναι κάπως χειρότερη για τα ατυχήματα, με μεγαλύτερη τιμή AIC και υψηλότερες τιμές MAE (19.66) και RMSE (177.93).

Στη συνέχεια ακολουθήθηκε η ίδια διαδικασία για κάθε ομάδα νησιών ξεχωριστά, ώστε να προκύψουν πιο συγκεκριμένα παραδείγματα για τη συμπεριφορά της κάθε περιοχής.

5.3.2 Εφαρμογή GLM στα Δωδεκάνησα

Το πρώτο μοντέλο αφορά στην ομάδα των νησιών των Δωδεκανήσων. Από την ανάλυση προέκυψε ότι το μοντέλο χρησιμοποιεί **960** γραμμές για εκπαίδευση και **240** γραμμές για τις προβλέψεις του, αλλά και ότι η κατανομή που ακολουθείται για Fatalities και Crashes είναι η **Negative Binomial.**

Ξεκινώντας από τα αποτελέσματα των **Fatalities**, το μοντέλο δείχνει ότι ο συντελεστής συσχέτισης για τις αφίξεις είναι 0.00909 ($p < 0.001$), που σημαίνει ότι **η αύξηση των αφίξεων συσχετίζεται με αύξηση των θανάτων κατά 0.91%**. Επίσης, ο δείκτης **McFadden R^2** ισούται με **0.16**, που δείχνει τη ότι το συγκεκριμένο μοντέλο εξηγεί το 16% της διακύμανσης της εξαρτημένης μεταβλητής, και ο συντελεστής **AICc** ισούται με **954.66**, δείχνοντας την καταλληλότητα του μοντέλου.

Παρακάτω φαίνονται τα ακριβή και αναλυτικά αποτελέσματα της αρνητικής διωνυμικής κατανομής που προέκυψαν για τους θανάτους.

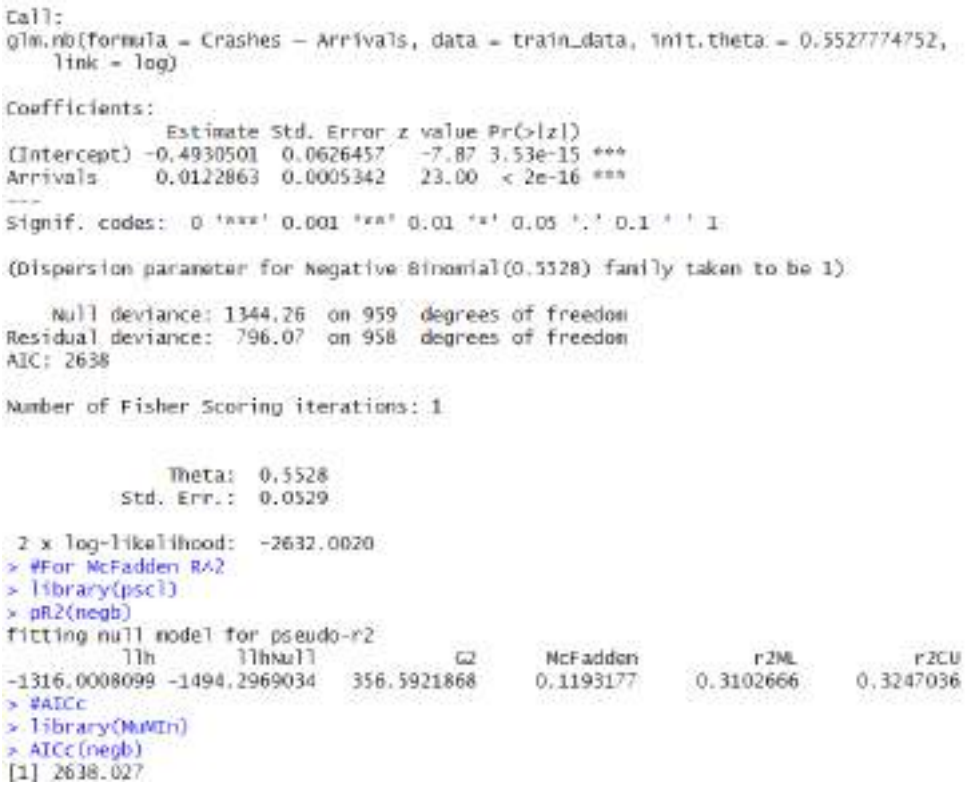
```
Call:
glm.nb(formula = Fatalities ~ Arrivals, data = train_data, init.theta = 0.7883336755,
    link = log)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.1981973  0.1048765  -20.96   <2e-16 ***
Arrivals     0.0090872  0.0005642   16.11   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(0.7883) family taken to be 1)

    Null deviance: 711.06  on 959  degrees of freedom
Residual deviance: 451.69  on 958  degrees of freedom
AIC: 954.64

Number of Fisher Scoring iterations: 1


              Theta:  0.788
          Std. Err.:  0.194

 2 x log-likelihood:  -948.636

> pR2(negb)
fitting null model for pseudo-r2
         llh      llhNull           G2     McFadden         r2ML         r2CU
-474.3177543 -564.5680409  180.5005732    0.1598572    0.1714031    0.2478548
> #AICc
> library(MuMIn)
> AICc(negb)
[1] 954.6606
```

Εικόνα 5.3: Αποτελέσματα GLM για Fatalities στην ομάδα των Δωδεκανήσων

Έπειτα, το μοντέλο προέβλεψε τους θανάτους για την περίοδο 2017-2018 και δημιούργησε έναν πίνακα **πραγματικών θανάτων – προβλεπόμενων θανάτων**. Τα σφάλματα για την πρόβλεψη θανάτων στην περίοδο δοκιμής ήταν:

- Mean Absolute Error (MAE): 0.64
- Root Mean Square Error (RMSE): 2.42

Ακολουθώντας τα ίδια βήματα για τα **Crashes**, ο συντελεστής για τις αφίξεις είναι 0.0123 ($p < 0.001$), που υποδηλώνει, όπως και στην περίπτωση των Fatalities, ότι **περισσότερες αφίξεις σχετίζονται με περισσότερα ατυχήματα κατά 1.24%**. Το μοντέλο έχει **McFadden R^2** που ισούται με **0.12** και το **AICc** είναι **2638.03**, εξηγώντας τη διακύμανση σε σχέση με τις αφίξεις, βέβαια σε χαμηλότερο βαθμό αυτή τη φορά.



```
Call:
glm.nb(formula = Crashes ~ Arrivals, data = train_data, init.theta = 0.5527774752,
    link = log)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.4930501  0.0626457   -7.87 3.53e-15 ***
Arrivals     0.0122863  0.0005342   23.00  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(0.5528) family taken to be 1)

    Null deviance: 1344.26  on 959  degrees of freedom
Residual deviance:  796.07  on 958  degrees of freedom
AIC: 2638

Number of Fisher Scoring iterations: 1


              Theta:  0.5528
          Std. Err.:  0.0529

 2 x log-likelihood:  -2632.0020
> #For McFadden R^2
> library(pscl)
> pR2(negb)
fitting null model for pseudo-r2
         llh       llhNull            G2      McFadden          r2ML          r2CU
-1316.0008099 -1494.2969034   356.5921868     0.1193177     0.3102666     0.3247036
> #AICc
> library(MuMIn)
> AICc(negb)
[1] 2638.027
```

Εικόνα 5.4: Αποτελέσματα GLM για Crashes στην ομάδα των Δωδεκανήσων

Όπως και στην προηγούμενη περίπτωση, το μοντέλο δημιούργησε έναν πίνακα **πραγματικών ατυχημάτων – προβλεπόμενων ατυχημάτων** και τα σφάλματα προέκυψαν:

- Mean Absolute Error (MAE): 15.40
- Root Mean Square Error (RMSE): 87.42

Τα σφάλματα των Crashes προέκυψαν μεγαλύτερα από τα αντίστοιχα των fatalities, γεγονός που εξηγείται από την μεγάλη διαφορά στις τιμές των πραγματικών θανάτων και των πραγματικών ατυχημάτων που το μοντέλο είχε να διαχειριστεί.

Ερμηνεία των Αποτελεσμάτων:

- Οι θετικοί συντελεστές για τις Αφίξεις και στα δύο μοντέλα υποδεικνύουν ότι καθώς αυξάνεται ο αριθμός των αφίξεων στα νησιά, αυξάνονται και οι θάνατοι και τα τροχαία ατυχήματα.
- Το AIC και το R^2 του McFadden δείχνουν μια ικανοποιητική προσαρμογή του μοντέλου, αν και οι τιμές R^2 είναι σχετικά χαμηλές, γεγονός που υποδεικνύει ότι οι αφίξεις εξηγούν μόνο ένα μέρος της μεταβλητότητας στους θανάτους.

- Το Μέσο Απόλυτο Σφάλμα (MAE) για τους θανάτους είναι 0.64, και το Τετραγωνικό Μέσο Σφάλμα (RMSE) είναι 2.42, υποδηλώνοντας ότι το μοντέλο προβλέπει ικανοποιητικά.
- Η προσαρμογή του μοντέλου είναι κάπως χειρότερη για τα ατυχήματα, με μεγαλύτερη τιμή AIC και υψηλότερες τιμές MAE (15.40) και RMSE (87.42).

5.3.3 Εφαρμογή GLM στις Κυκλάδες

Το δεύτερο μοντέλο αφορά στην ομάδα των νησιών των Κυκλάδων. Από την ανάλυση προέκυψε ότι το μοντέλο χρησιμοποιεί **1344** γραμμές για εκπαίδευση και **336** γραμμές για τις προβλέψεις του, αλλά και ότι η κατανομή που ακολουθείται για Fatalities και Crashes είναι η **Negative Binomial.**

Ξεκινώντας από τα αποτελέσματα των **Fatalities**, το μοντέλο δείχνει ότι ο συντελεστής συσχέτισης για τις αφίξεις είναι 0.0112 ($p < 0.001$), που σημαίνει ότι **η αύξηση των αφίξεων συσχετίζεται με αύξηση των θανάτων 1.13%**. Επίσης, ο δείκτης **McFadden R^2** ισούται με **0.11**, που δείχνει τη ότι το συγκεκριμένο μοντέλο εξηγεί το 11% της διακύμανσης της εξαρτημένης μεταβλητής και ο συντελεστής **AICc** ισούται με **628.6**, δείχνοντας την καταλληλότητα του μοντέλου.

Παρακάτω φαίνονται τα ακριβή και αναλυτικά αποτελέσματα της αρνητικής διωνυμικής κατανομής που προέκυψαν για τους θανάτους.

```
Call:
glm.nb(formula = Fatalities ~ Arrivals, data = train_data, init.theta = 33.15413874,
    link = log)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.200551   0.136012  -23.53   <2e-16 ***
Arrivals     0.011216   0.001026   10.94   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(33.1541) family taken to be 1)

    Null deviance: 520.76  on 1343  degrees of freedom
Residual deviance: 439.71  on 1342  degrees of freedom
AIC: 628.6

Number of Fisher Scoring iterations: 1


              Theta:  33
          Std. Err.:  383

 2 x log-likelihood:  -622.599

> pR2(negb)
fitting null model for pseudo-r2
         llh       llhNull            G2      McFadden          r2ML          r2CU
-311.29954254 -349.92122630   77.24336753    0.11037251    0.05585238    0.13759061
> #AICc
> library(MuMIn)
> AICc(negb)
[1] 628.617
```

Εικόνα 5.5: Αποτελέσματα GLM για Fatalities στην ομάδα των Κυκλάδων

Έπειτα, για την περίοδο 2017-2018, το μοντέλο δημιούργησε έναν πίνακα που αποτελείται από τις τιμές των **πραγματικών** και των **προβλεπόμενων θανάτων,** με τα σφάλματα της πρόβλεψης στην δοκιμαστική περίοδο να είναι:

- Μέσο Απόλυτο Σφάλμα (MAE): 0.16
- Ριζική Μέση Τετραγωνική Απόκλιση (RMSE): 0.34

Για τα **Crashes**, το μοντέλο έδειξε επίσης **σημαντική θετική σχέση μεταξύ των αφίξεων και των ατυχημάτων κατά 1.84%**(συντελεστής = 0.0182, $p < 0.001$), με δείκτη **McFadden R^2** που ισούται με **0.11** και AICc ίσο με **2491.1**.

```
Call:
glm.nb(formula = Crashes ~ Arrivals, data = train_data, init.theta = 0.5481233457,
    link = log)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.3017857  0.0673487  -19.33   <2e-16 ***
Arrivals     0.0181781  0.0008857   20.52   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(0.5481) family taken to be 1)

    Null deviance: 1305.34  on 1343  degrees of freedom
Residual deviance:  908.67  on 1342  degrees of freedom
AIC: 2491.1

Number of Fisher Scoring iterations: 1


              Theta:  0.5481
          Std. Err.:  0.0582

 2 x log-likelihood:  -2485.1230
> #For McFadden R^2
> library(pscl)
> pR2(negb)
fitting null model for pseudo-r2
         llh      llhNull           G2     McFadden         r2ML         r2CU
-1242.5617053 -1393.2316820  301.3399533    0.1081442    0.2008537    0.2297506
> #AICc
> library(MuMIn)
> AICc(negb)
[1] 2491.141
```

Εικόνα 5.6: Αποτελέσματα GLM για Crashes στην ομάδα των Κυκλάδων

Τέλος, το μοντέλο δημιούργησε εξίσου έναν πίνακα **πραγματικών ατυχημάτων – προβλεπόμενων ατυχημάτων**, με σφάλματα πρόβλεψης:

- **Μέσο Απόλυτο Σφάλμα (MAE)**: 3.40
- **Ριζική Μέση Τετραγωνική Απόκλιση (RMSE)**: 15.83

Ερμηνεία των Αποτελεσμάτων:

- Και στα δύο μοντέλα, οι συντελεστές για τις αφίξεις είναι θετικοί, υποδηλώνοντας ότι η αύξηση των αφίξεων συνδέεται με την αύξηση των θανάτων και των τροχαίων ατυχημάτων.
- Το AIC και το R^2 του McFadden δείχνουν μια ικανοποιητική προσαρμογή του μοντέλου, αν και οι τιμές R^2 είναι σχετικά χαμηλές, γεγονός που υποδεικνύει ότι οι αφίξεις εξηγούν μόνο ένα μέρος της μεταβλητότητας στους θανάτους.
- Το Μέσο Απόλυτο Σφάλμα (MAE) για τους θανάτους είναι 0.16, και το Τετραγωνικό Μέσο Σφάλμα (RMSE) είναι 0.34, υποδηλώνοντας ότι το μοντέλο προβλέπει ικανοποιητικά.

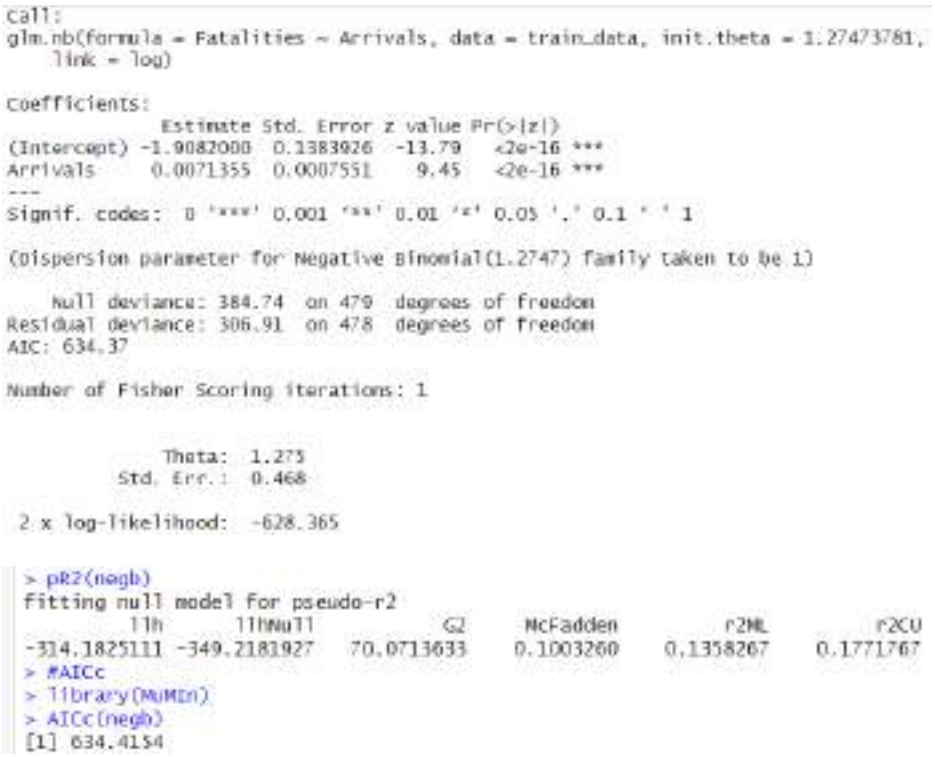
- Η προσαρμογή του μοντέλου είναι κάπως χειρότερη για τα ατυχήματα, με μεγαλύτερη τιμή AIC και υψηλότερες τιμές MAE (3.40) και RMSE (15.83).

5.3.4 Εφαρμογή GLM στο Ιόνιο

Το τρίτο μοντέλο αφορά στην ομάδα των νησιών των Ιόνιων Νήσων. Από την ανάλυση προέκυψε ότι το μοντέλο χρησιμοποιεί **480** γραμμές για εκπαίδευση και **120** γραμμές για τις προβλέψεις του, αλλά και ότι η κατανομή που ακολουθείται για Fatalities και Crashes είναι η **Negative Binomial.**

Ξεκινώντας από τα αποτελέσματα των **Fatalities**, το μοντέλο δείχνει ότι ο συντελεστής συσχέτισης για τις αφίξεις είναι 0.0071 ($p < 0.001$), που σημαίνει ότι **η αύξηση των αφίξεων συσχετίζεται με αύξηση των θανάτων κατά 0.72%**. Επίσης, ο δείκτης **McFadden R^2** ισούται με **0.10**, που δείχνει τη ότι το συγκεκριμένο μοντέλο εξηγεί το 10% της διακύμανσης της εξαρτημένης μεταβλητής και ο συντελεστής **AICc** ισούται με **634.4**, δείχνοντας την καταλληλότητα του μοντέλου.

Παρακάτω φαίνονται τα ακριβή και αναλυτικά αποτελέσματα της αρνητικής διωνυμικής κατανομής που προέκυψαν για τους θανάτους.

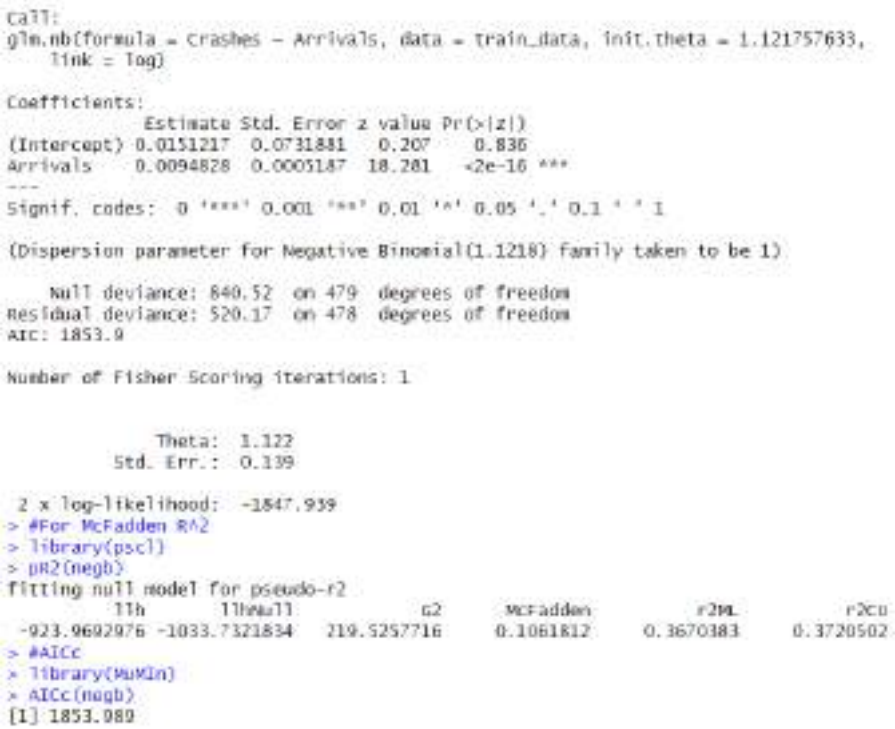


Εικόνα 5.7: Αποτελέσματα GLM για Fatalities στην ομάδα του Ιονίου

Στη συνέχεια, το μοντέλο δημιούργησε τον πίνακα **πραγματικών θανάτων – προβλεπόμενων θανάτων** για την περίοδο 2017-2018, με τα αντίστοιχα σφάλματα να προκύπτουν:

- Μέσο Απόλυτο Σφάλμα (MAE): 0.59
- Ριζική Μέση Τετραγωνική Απόκλιση (RMSE): 1.08

Αντίστοιχα για τα **Crashes**, το μοντέλο δείχνει ότι οι αφίξεις έχουν **θετική και σημαντική επίδραση στον αριθμό των ατυχημάτων κατά 0.95%** ($p < 0.001$), με εκτιμώμενο συντελεστή 0.0095, ο **McFadden R^2** είναι **0.11** και το **AICc** είναι **1853.99**.



Εικόνα 5.8: Αποτελέσματα GLM για Crashes στην ομάδα του Ιονίου

Το μοντέλο δημιούργησε και σε αυτήν την περίπτωση έναν πίνακα **πραγματικών ατυχημάτων – προβλεπόμενων ατυχημάτων**. Τα σφάλματα που προκύπτουν είναι τα ακόλουθα:

- Μέσο Απόλυτο Σφάλμα (MAE): 5.78
- Ριζική Μέση Τετραγωνική Απόκλιση (RMSE): 18.60

Ερμηνεία των Αποτελεσμάτων:

- Η αύξηση των αφίξεων συνδέεται με αύξηση τόσο στους θανάτους όσο και στα τροχαία ατυχήματα, όπως δείχνουν οι θετικοί συντελεστές και στα δύο μοντέλα.
- Το AIC και το R^2 του McFadden δείχνουν μια ικανοποιητική προσαρμογή του μοντέλου, αν και οι τιμές R^2 είναι σχετικά χαμηλές, γεγονός που υποδεικνύει ότι οι αφίξεις εξηγούν μόνο ένα μέρος της μεταβλητότητας στους θανάτους.
- Το Μέσο Απόλυτο Σφάλμα (MAE) για τους θανάτους είναι 0.59, και το Τετραγωνικό Μέσο Σφάλμα (RMSE) είναι 1.08, υποδηλώνοντας ότι το μοντέλο προβλέπει ικανοποιητικά.
- Η προσαρμογή του μοντέλου είναι κάπως χειρότερη για τα ατυχήματα, με μεγαλύτερη τιμή AIC και υψηλότερες τιμές MAE (5.78) και RMSE (18.60).

5.3.5 Εφαρμογή GLM στο Κεντρικό και Βόρειο Αιγαίο

Το τέταρτο μοντέλο αφορά στην ομάδα των νησιών των Ιόνιων Νήσων. Από την ανάλυση προέκυψε ότι το μοντέλο χρησιμοποιεί **864** γραμμές για εκπαίδευση και **216** γραμμές για τις προβλέψεις του, αλλά και ότι η κατανομή που ακολουθείται για Fatalities και Crashes είναι η **Negative Binomial.**

Το μοντέλο για τα **Fatalities** βρίσκει ότι ο **McFadden R^2** ισούται με **0.06**, καθιστώντας το μοντέλο **ακατάλληλο** για τα δεδομένα.

```
Call:
glm.nb(formula = Fatalities ~ Arrivals, data = train_data, init.theta = 0.4758548148,
    link = log)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.69744    0.15640  -17.25  < 2e-16 ***
Arrivals     0.03354    0.00410    8.18 2.85e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(0.4759) family taken to be 1)

    Null deviance: 389.57  on 863  degrees of freedom
Residual deviance: 341.04  on 862  degrees of freedom
AIC: 708.52

Number of Fisher Scoring iterations: 1


              Theta:  0.476
          Std. Err.:  0.136

 2 x log-likelihood:  -702.516

> pR2(negb)
fitting null model for pseudo-r2
         llh       llhNull            G2      McFadden          r2ML          r2CU
-351.25778614 -374.93344757   47.35132287    0.06314630    0.05333005    0.09192169
> #AICc
> library(MuMIn)
> AICc(negb)
[1] 708.5435
```

Εικόνα 5.9: Αποτελέσματα GLM για Fatalities στην ομάδα του Κ/Β Αιγαίου

Το μοντέλο για τα **Crashes** δείχνει ότι οι αφίξεις έχουν μία οριακή αλλά **στατιστικά σημαντική θετική επίδραση στα τροχαία ατυχήματα κατά 5.9%** (εκτίμηση συντελεστή: 0.05742). Ο **McFadden R^2** είναι **0.10** και το **AICc** ισούται με **2227.65**.

```
Call:
glm.nb(formula = Crashes ~ Arrivals, data = train_data, init.theta = 0.5582755902,
    link = log)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.205609   0.088459  -13.63   <2e-16 ***
Arrivals     0.057420   0.002761   20.80   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(0.5583) family taken to be 1)

    Null deviance: 976.19  on 863  degrees of freedom
Residual deviance: 673.98  on 862  degrees of freedom
AIC: 2227.6

Number of Fisher Scoring iterations: 1


              Theta:  0.5583
          Std. Err.:  0.0529

 2 x log-likelihood:  -2221.6210
> #For McFadden R^2
> library(pscl)
> pR2(negb)
fitting null model for pseudo-r2
         llh       llhNull            G2      McFadden          r2ML          r2CU
-1110.81025228 -1234.16526866   246.71003276     0.09995016     0.24839474     0.26353444
> #AICc
> library(MuMIn)
> AICc(negb)
[1] 2227.648
```

Εικόνα 5.10: Αποτελέσματα GLM για Crashes στην ομάδα του Κ/Β Αιγαίου

Τέλος, γίνεται σύγκριση μεταξύ **πραγματικών και προβλεπόμενων θανάτων** και τα σφάλματα που προκύπτουν είναι:

- Μέσο Απόλυτο Σφάλμα (MAE): 4.35
- Ριζική Μέση Τετραγωνική Απόκλιση (RMSE): 17.52

Ερμηνεία των Αποτελεσμάτων:

- Η αύξηση των αφίξεων συνδέεται με αύξηση στα τροχαία ατυχήματα, σύμφωνα με τον θετικό συντελεστή του μοντέλου, ωστόσο λόγω χαμηλού R^2 στους θανάτους, δεν προκύπτει ασφαλές συμπέρασμα για τη συσχέτιση τους με τις αφίξεις και δεν είναι χρήσιμο να πραγματοποιηθούν προβλέψεις.

- Η προσαρμογή του μοντέλου για τα ατυχήματα με τιμές MAE (4.35) και RMSE (17.52) κρίνεται και σε αυτή την περίπτωση ικανοποιητική.

5.4 Εφαρμογή Random Forest

Αρχικά, γίνεται εγκατάσταση των **βιβλιοθηκών** που είναι απαραίτητες στην επιτέλεση του κώδικα. Αυτά είναι το “readxl”, που είναι υπεύθυνο για την ανάγνωση των αρχείων Excel του υπολογιστή, το “dplyr”, το οποίο διαχειρίζεται και επεξεργάζεται στα δεδομένα και η “randomForest”, που είναι υπεύθυνη για την εφαρμογή του μοντέλου. Με την εντολή “library” γίνεται η φόρτωση των βιβλιοθηκών.

Τα δεδομένα χωρίζονται σε **σετ εκπαίδευσης** (2009-2016) και **σετ ελέγχου** (2017-2018). Το σετ εκπαίδευσης χρησιμοποιείται για να κατασκευαστεί το μοντέλο, ενώ το σετ ελέγχου για να αξιολογηθεί η ακρίβεια του.

Στη συνέχεια, ο κώδικας χρησιμοποιεί την εντολή **randomForest** για να δημιουργήσει ένα μοντέλο που προβλέπει τους θανάτους με βάση τις αφίξεις (Fatalities ~ Arrivals) χρησιμοποιώντας 500 δέντρα. Το μοντέλο αναλύει τη σχέση μεταξύ αφίξεων και θανάτων, εξετάζοντας πολλαπλές δειγματοληψίες από τα δεδομένα για να δημιουργήσει διάφορα δέντρα αποφάσεων. Επίσης, υπολογίζει τις **προβλέψεις των θανάτων** για το σετ ελέγχου, συνδυάζει τις προβλέψεις με τις **πραγματικές τιμές** για να μπορεί να συγκρίνει τις διαφορές και υπολογίζει το σφάλμα ρίζας μέσου τετραγώνου **RMSE** και το μέσο απόλυτο σφάλμα **MAE** για να μετρήσει την ακρίβεια του μοντέλου.

Τέλος, ο κώδικας χρησιμοποιεί την εντολή **randomForest** για να προβλέψει τα τροχαία ατυχήματα με βάση τις αφίξεις (Crashes ~ Arrivals). Το μοντέλο προσπαθεί να μάθει τη σχέση μεταξύ του αριθμού αφίξεων και των ατυχημάτων και υπολογίζει τις προβλέψεις στο σετ ελέγχου, το RMSE και το MAE για τα τροχαία ατυχήματα με τον ίδιο τρόπο όπως και για τους θανάτους.

5.4.1 Εφαρμογή Random Forest για όλα τα νησιά

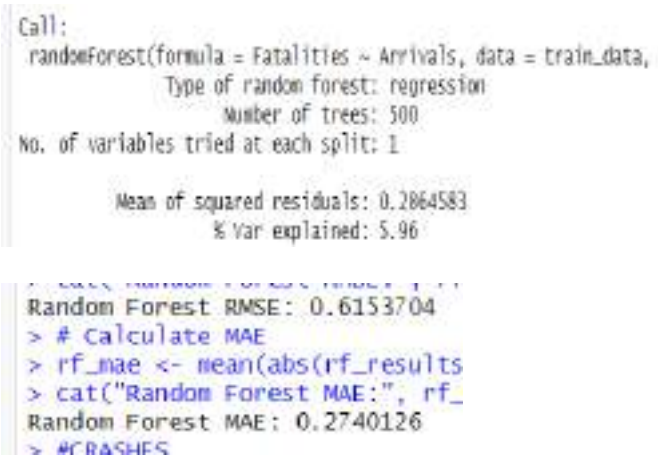
Το μοντέλο εφαρμόζεται αρχικά για όλα τα νησιά, ώστε να ελεγχθεί ο βαθμός συσχέτισης αφίξεων με θανάτους και ατυχήματα σε μεγαλύτερη κλίμακα. Οπότε, το μοντέλο εκπαιδεύτηκε χρησιμοποιώντας δεδομένα από τις αφίξεις για να προβλέψει τα **Fatalities** για την περίοδο 2009-2016 (**3648** γραμμές για εκπαίδευση, **912** για πρόβλεψη των ετών 2017-2018). η μέση

τιμή των τετραγωνικών υπολοίπων (**Mean Squared Residuals**) είναι **0.286,** αριθμός που δείχνει πόσο αποκλίνουν οι προβλέψεις από τις πραγματικές τιμές και το ποσοστό διακύμανσης είναι **5.96%**. Τα σφάλματα έχουν τις ακόλουθες τιμές:

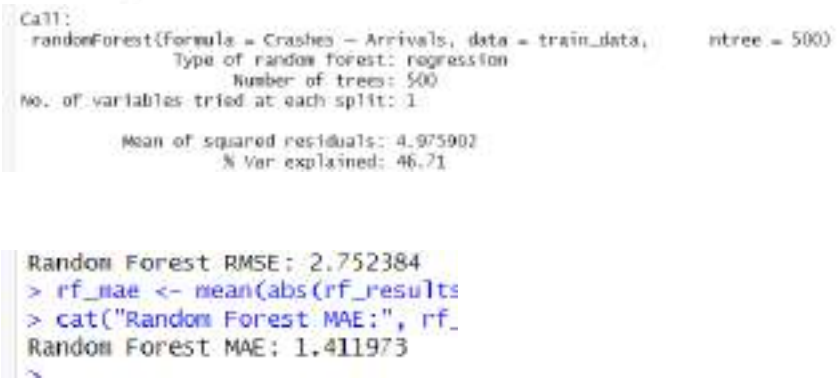
- Μέσο Απόλυτο Σφάλμα (MAE): 0.27
- Ριζική Μέση Τετραγωνική Απόκλιση (RMSE): 0.62

Το μοντέλο για τα **Crashes** χρησιμοποιώντας τα ίδια δεδομένα εκπαίδευσης, καταλήγει σε **Mean Squared Residuals** ίσο με **4.976** και σε ένα ικανοποιητικό ποσοστό διακύμανσης (**46.71%**). Τα σφάλματα ήταν:

- Μέσο Απόλυτο Σφάλμα (MAE): 1.41
- Ριζική Μέση Τετραγωνική Απόκλιση (RMSE): 2.75



Εικόνα 5.11: Αποτελέσματα Random Forest για Fatalities σε όλα τα νησιά



Εικόνα 5.12: Αποτελέσματα Random Forest για Crashes σε όλα τα νησιά

Ερμηνεία των Αποτελεσμάτων:

- Το ποσοστό διακύμανσης που εξηγείται για Fatalities από το μοντέλο είναι χαμηλό (5.96%). Αυτό δείχνει ότι οι αφίξεις εξηγούν μόνο ένα μέρος της μεταβλητότητας των θανάτων και ότι επιπλέον παράμετροι χρειάζεται να ληφθούν υπόψιν προκειμένου να γίνει ακριβής πρόβλεψη στον αριθμό των θανάτων.
- Το ποσοστό διακύμανσης που εξηγείται για Crashes είναι υψηλότερο (46.71%). Οι αφίξεις φαίνεται να έχουν σημαντική επίδραση στην πρόβλεψη των ατυχημάτων. Αυτό υποδηλώνει ότι όσο αυξάνονται οι τουρίστες, αυξάνονται και τα ατυχήματα.
- Τα σφάλματα για τους θανάτους είναι ικανοποιητικά με MAE (0.27) και RMSE (0.62).

- Τα σφάλματα για τα ατυχήματα είναι εξίσου πολύ ικανοποιητικά με MAE (1.41) και RMSE (2.75).

Στα επόμενα υποκεφάλαια ακολουθήθηκε η ίδια διαδικασία, αλλά τα νησιά χωρίστηκαν σε ομάδες, ώστε να προκύψουν διαφορετικά συμπεράσματα για κάθε περιοχή.

5.4.2 Εφαρμογή Random Forest για Δωδεκάνησα

Το πρώτο μοντέλο εκπαιδεύτηκε χρησιμοποιώντας δεδομένα από τις αφίξεις για να προβλέψει τα **Fatalities** για την περίοδο 2009-2016 (**960** γραμμές για εκπαίδευση, **240** για πρόβλεψη των ετών 2017-2018).

Η μέση τιμή των τετραγωνικών υπολοίπων (**Mean Squared Residuals**) ήταν **0.456**, που δείχνει πόσο αποκλίνουν οι προβλέψεις από τις πραγματικές τιμές. Για την αξιολόγηση της ακρίβειας, προέκυψαν τα σφάλματα:

- Μέσο Απόλυτο Σφάλμα (MAE): 0.29
- Ριζική Μέση Τετραγωνική Απόκλιση (RMSE): 0.65

Το δεύτερο μοντέλο εκπαιδεύτηκε με βάση τις αφίξεις για την πρόβλεψη των **Crashes**, χρησιμοποιώντας και πάλι δεδομένα για την περίοδο 2009-2016. Το ποσοστό της διακύμανσης που εξηγείται από το μοντέλο ήταν **63.15%**, υποδεικνύοντας ότι το Random Forest αποδίδει πολύ καλύτερα στις προβλέψεις των τροχαίων ατυχημάτων σε σχέση με τους θανάτους (23.16%).

Η μέση τιμή των τετραγωνικών υπολοίπων για τα τροχαία ατυχήματα ήταν **5.962** και προέκυψαν τα ακόλουθα σφάλματα:

- Μέσο Απόλυτο Σφάλμα (MAE): 1.06
- Ριζική Μέση Τετραγωνική Απόκλιση (RMSE): 2.10

Εικόνα 5.13: Αποτελέσματα Random Forest για Fatalities στην ομάδα των Δωδεκανήσων

Εικόνα 5.14: Αποτελέσματα Random Forest για Crashes στην ομάδα των Δωδεκανήσων

Ερμηνεία των Αποτελεσμάτων:

- Το ποσοστό διακύμανσης που εξηγείται για Fatalities από το μοντέλο είναι χαμηλό (22.87%). Αυτό δείχνει ότι οι αφίξεις εξηγούν μόνο ένα μέρος της μεταβλητότητας των θανάτων και ότι επιπλέον παράμετροι χρειάζεται να ληφθούν υπόψιν προκειμένου να γίνει ακριβής πρόβλεψη στον αριθμό των θανάτων.
- Το ποσοστό διακύμανσης που εξηγείται για Crashes είναι υψηλότερο (63.09%). Οι αφίξεις φαίνεται να έχουν σημαντική επίδραση στην πρόβλεψη των ατυχημάτων. Αυτό υποδηλώνει ότι όσο αυξάνονται οι τουρίστες, αυξάνονται και τα ατυχήματα.
- Τα σφάλματα για τους θανάτους είναι ικανοποιητικά με MAE (0.29) και RMSE (0.65).
- Τα σφάλματα για τα ατυχήματα είναι εξίσου πολύ ικανοποιητικά με MAE (1.06) και RMSE (2.10).

5.4.3 Εφαρμογή Random Forest για Κυκλάδες

Το μοντέλο για τα **Fatalities** χρησιμοποιεί **1344** γραμμές εκπαίδευσης και **336** γραμμές για προβλέψεις δείχνοντας ότι η μέση τιμή των τετραγωνικών υπολοίπων (**Mean Squared Residuals**) είναι **0.0982,** το ποσοστό διακύμανσης είναι αρνητικό (**-26.63%)** και τα σφάλματα έχουν τις ακόλουθες τιμές:

- Μέσο Απόλυτο Σφάλμα (MAE): 0.16
- Ριζική Μέση Τετραγωνική Απόκλιση (RMSE): 0.39

Το μοντέλο για τα **Crashes** χρησιμοποιώντας τα ίδια δεδομένα εκπαίδευσης, καταλήγει σε **Mean Squared Residuals** ίσο με **2.1623,** χαμηλό ποσοστό διακύμανσης (**16.53%**) και στα παρακάτω σφάλματα:

- Μέσο Απόλυτο Σφάλμα (MAE): 0.96
- Ριζική Μέση Τετραγωνική Απόκλιση (RMSE): 1.91

Εικόνα 5.15: Αποτελέσματα Random Forest για Fatalities στην ομάδα των Κυκλάδων

Εικόνα 5.16: Αποτελέσματα Random Forest για Crashes στην ομάδα των Κυκλάδων

Ερμηνεία των Αποτελεσμάτων:

- Το ποσοστό διακύμανσης που εξηγείται για Fatalities είναι αρνητικό, που σημαίνει ότι οι αφίξεις δεν προσφέρουν χρήσιμες πληροφορίες για την πρόβλεψη των θανάτων.
- Το ποσοστό διακύμανσης που εξηγείται για Crashes είναι χαμηλό, αλλά θετικό, δηλαδή οι αφίξεις παρέχουν περιορισμένες πληροφορίες για την πρόβλεψη των ατυχημάτων.
- Τα σφάλματα για τους θανάτους είναι ικανοποιητικά με MAE (0.16) και RMSE (0.39).
- Τα σφάλματα για τα ατυχήματα είναι εξίσου πολύ ικανοποιητικά με MAE (0.96) και RMSE (1.91).

5.4.4 Εφαρμογή Random Forest για Ιόνιο

Το μοντέλο για τα **Fatalities** χρησιμοποιεί **480** γραμμές εκπαίδευσης και **120** γραμμές για προβλέψεις, καταλήγοντας σε ένα **Mean Squared Residuals** ίσο με **0.472**, ένα μικρό ποσοστό διακύμανσης (**2.84%**) και στα ακόλουθα σφάλματα:

- Μέσο Απόλυτο Σφάλμα (MAE): 0.40
- Ριζική Μέση Τετραγωνική Απόκλιση (RMSE): 0.72

Το μοντέλο για τα **Crashes** χρησιμοποιώντας τα ίδια δεδομένα εκπαίδευσης, καταλήγει σε **Mean Squared Residuals** ίσο με **7.972** , ικανοποιητικό ποσοστό διακύμανσης (**53.14%**) και στα παρακάτω σφάλματα:

- Μέσο Απόλυτο Σφάλμα (MAE): 1.92
- Ριζική Μέση Τετραγωνική Απόκλιση (RMSE): 3.42

```
> # Fit Random Forest model on training data
> rf_model <- randomForest(Fatalities ~ Arrivals, data = train_data, ntree = 500)
> # View the model summary
> print(rf_model)

Call:
 randomForest(formula = Fatalities ~ Arrivals, data = train_data,      ntree = 500)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 1

          Mean of squared residuals: 0.470787
                    % Var explained: 2.84
> # Make predictions on the test data
> rf_predictions <- predict(rf_model, newdata = test_data)
> # Combine predictions with actual values for comparison
> rf_results <- data.frame(
+   Date = test_data$Date,
+   Actual_Fatalities = test_data$Fatalities,
+   Predicted_Fatalities = rf_predictions
+ )
> # View the first few rows of the results
> print(head(rf_results))
        Date Actual_Fatalities Predicted_Fatalities
1 2017-01-01                0          3.386667e-02
2 2017-02-01                0          1.379867e-01
3 2017-03-01                1          3.660000e-02
4 2017-04-01                1         -1.756371e-16
5 2017-05-01                0          1.960000e-01
6 2017-06-01                0          5.040000e-02
> # Calculate RMSE
> rf_rmse <- sqrt(mean((rf_results$Actual_Fatalities - rf_results$Predicted_Fatalities)^2))
```

Εικόνα 5.17: Αποτελέσματα Random Forest για Fatalities στην ομάδα του Ιονίου

```
> #CRASHES
> rf_model <- randomForest(Crashes ~ Arrivals, data = train_data, ntree = 500)
> print(rf_model)

Call:
 randomForest(formula = Crashes ~ Arrivals, data = train_data,      ntree = 500)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 1

          Mean of squared residuals: 8.014813
                    % Var explained: 53.14
> rf_predictions <- predict(rf_model, newdata = test_data)
> rf_results <- data.frame(
+   Date = test_data$Date,
+   Actual_Crashes = test_data$Crashes,
+   Predicted_Crashes = rf_predictions
+ )
> print(head(rf_results))
        Date Actual_Crashes Predicted_Crashes
1 2017-01-01              1          3.1891000
2 2017-02-01              4          7.1658000
3 2017-03-01              2          0.8653867
4 2017-04-01              2          2.1090333
5 2017-05-01              3          3.4936000
6 2017-06-01              5          2.8760333
> rf_rmse <- sqrt(mean((rf_results$Actual_Crashes - rf_results$Predicted_Crashes)^2))
> cat("Random Forest RMSE:", rf_rmse, "\n")
```

Εικόνα 5.18: Αποτελέσματα Random Forest για Crashes στην ομάδα του Ιονίου

Ερμηνεία των Αποτελεσμάτων:

- Οι αφίξεις έχουν μικρή σημασία στην πρόβλεψη των θανάτων, καθώς ο συντελεστής συσχέτισης, αν και θετικός, είναι πολύ μικρός (2.84%).
- Οι αφίξεις έχουν σημαντικό ρόλο στην πρόβλεψη των ατυχημάτων (συντελεστής συσχέτισης 53.47%), κάτι που υποδηλώνει ότι καθώς αυξάνεται ο τουρισμός, αυξάνεται και η πιθανότητα ατυχημάτων.
- Τα σφάλματα για τους θανάτους είναι ικανοποιητικά με MAE (0.40) και RMSE (0.72).
- Τα σφάλματα για τα ατυχήματα είναι εξίσου πολύ ικανοποιητικά με MAE (1.93) και RMSE (3.44).

5.4.5 Εφαρμογή Random Forest για Κεντρικό και Βόρειο Αιγαίο

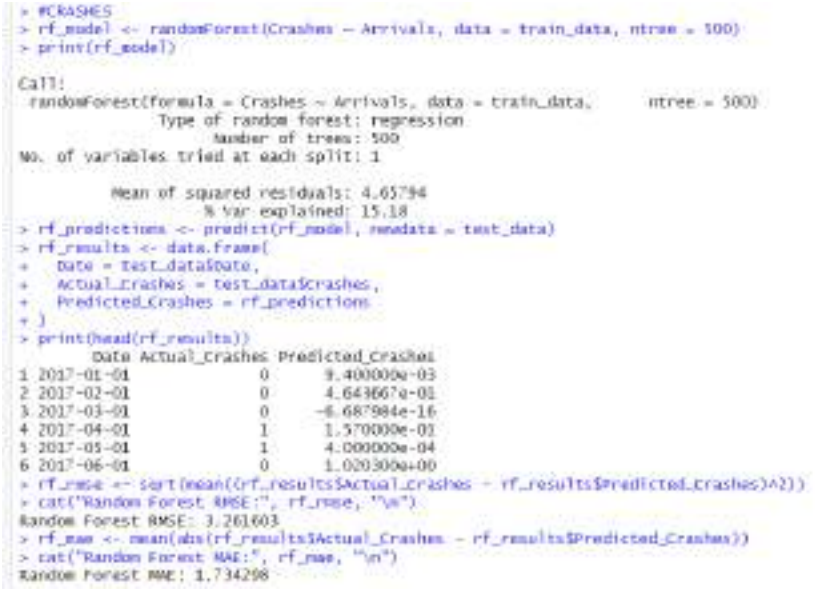
Το μοντέλο για τα **Fatalities** χρησιμοποιεί **864** γραμμές εκπαίδευσης και **216** γραμμές για προβλέψεις και βγάζει ένα **Mean Squared Residuals** ίσο με **0.245**, αρνητικό ποσοστό διακύμανσης και τα ακόλουθα σφάλματα:

- Μέσο Απόλυτο Σφάλμα (MAE): 0.24
- Ριζική Μέση Τετραγωνική Απόκλιση (RMSE): 0.50

Με τα ίδια δεδομένα εκπαίδευσης, το μοντέλο για τα **Crashes** δίνει ένα **Mean Squared Residuals** που ισούται με **4.664**, ποσοστό διακύμανσης της τάξης του 15% και τα παρακάτω σφάλματα:

- Μέσο Απόλυτο Σφάλμα (MAE): 1.74
- Ριζική Μέση Τετραγωνική Απόκλιση (RMSE): 3.26

Εικόνα 5.19: Αποτελέσματα Random Forest για Fatalities στην ομάδα του Κ/Β Αιγαίου



Εικόνα 5.20: Αποτελέσματα Random Forest για Crashes στην ομάδα του Κ/Β Αιγαίου

Ερμηνεία των Αποτελεσμάτων:

- Οι αφίξεις δεν φαίνεται να παίζουν σημαντικό ρόλο στην πρόβλεψη των θανάτων με βάση το παρόν μοντέλο.
- Οι αφίξεις δείχνουν έναν μέτριο ρόλο στην πρόβλεψη των ατυχημάτων, αλλά ένα μεγάλο μέρος της διακύμανσης παραμένει ανεξήγητο.
- Τα σφάλματα για τους θανάτους είναι ικανοποιητικά με MAE (0.24) και RMSE (0.50).
- Τα σφάλματα για τα ατυχήματα είναι εξίσου πολύ ικανοποιητικά με MAE (1.74) και RMSE (3.24).

5.5 Αποτελέσματα Μοντέλων

Τα μοντέλα για όλες τις ομάδες νησιών έδειξαν ότι υπάρχει θετική συσχέτιση μεταξύ του αριθμού των αφίξεων και του αριθμού των θανάτων και των ατυχημάτων.

Ειδικότερα, ο μεγαλύτερος συντελεστής συσχέτισης των μοντέλων για τους θανάτους παρουσιάζεται στις Κυκλάδες, με τα Δωδεκάνησα να έρχονται δεύτερα και το Ιόνιο τρίτο, ενώ το πενιχρό McFadden R^2 του μοντέλου GLM για την ομάδα του Κεντρικού και Βόρειου Αιγαίου οδηγεί στο συμπέρασμα ότι δεν είναι στατιστικά σημαντικό, άρα και μη αποδεκτό.

Όσον αφορά τα μοντέλα με εξαρτημένη μεταβλητή τα ατυχήματα, ο μεγαλύτερος συντελεστής συσχέτισης παρουσιάζεται για το μοντέλο της ομάδας των νησιών του Κεντρικού και Βορείου

Αιγαίου και ακολουθούν με φθίνουσα σειρά οι Κυκλάδες και τα Δωδεκάνησα, με το Ιόνιο να έρχεται τελευταίο.

Με τη χρήση των GLM για τη συσχέτιση τω αφίξεων με τους θανάτους σε οδικά ατυχήματα, τα μικρότερα σφάλματα πρόβλεψης προκύπτουν για την ομάδα των Κυκλάδων, με το Ιόνιο να ακολουθεί και τα Δωδεκάνησα να βρίσκονται στο τέλος. Στα αντίστοιχα μοντέλα των ατυχημάτων, τα μικρότερα σφάλματα πρόβλεψης παρατηρούνται για τις Κυκλάδες, με το Κεντρικό και Βόρειο Αιγαίο, το Ιόνιο και τα Δωδεκάνησα να έπονται.

Χρησιμοποιήθηκε επίσης η μέθοδος Random Forest, προκειμένου να συσχετιστούν οι αφίξεις των τουριστών με τα οδικά ατυχήματα και τους θανάτους σε αυτά στις ίδιες ομάδες νησιών. Σε σχέση με τα μοντέλα για τον αριθμό των θανάτων σε οδικά ατυχήματα, προέκυψε μόνο ένα ικανοποιητικό μοντέλο, για την ομάδα των Δωδεκανήσων με αποδεκτό σφάλμα πρόβλεψης.

Για τα δε μοντέλα συσχέτισης αφίξεων ατυχημάτων, προέκυψαν αποδεκτά μοντέλα για όλες τις ομάδες νησιών. Επισημαίνεται παρ’ όλα αυτά, ότι για τις Κυκλάδες και το Κεντρικό και Βόρειο Αιγαίο, το ποσοστό διακύμανσης που εξηγείται είναι σχετικά χαμηλό. Τα μικρότερα σφάλματα πρόβλεψης υπολογίστηκαν για τις Κυκλάδες και τα Δωδεκάνησα, με το Κεντρικό και Βόρειο Αιγαίο και το Ιόνιο να ακολουθούν.

Γενικότερα, συγκρίνοντας τις δύο μεθόδους ως προς τα σφάλματα προβλέψεων, πιο χαμηλοί δείκτες MAE και RMSE παρατηρούνται για τα μοντέλα αφίξεων-ατυχημάτων με τη μέθοδο Random Forest.

Κεφάλαιο 6: Συμπεράσματα

6.1 Σύνοψη Αποτελεσμάτων

Η παρούσα Διπλωματική Εργασία στοχεύει να διερευνήσει τα οδικά ατυχήματα στα ελληνικά νησιά και πιο συγκεκριμένα **την επιρροή του τουρισμού στα ατυχήματα με τραυματίες, αλλά και στον αριθμό των νεκρών σε οδικά ατυχήματα**. Για τον σκοπό αυτόν δημιουργήθηκαν μοντέλα συσχέτισης ατυχημάτων – αφίξεων τουριστών και μοντέλα συσχέτισης θανάτων-αφίξεων τουριστών.

Για την επίτευξη των στόχων της Διπλωματικής Εργασίας, **συλλέχθηκαν δεδομένα οδικών ατυχημάτων και αφίξεων τουριστών** σε λιμάνια και αεροδρόμια 38 ελληνικών νησιών σε μηνιαία βάση. Τα απαραίτητα για την μελέτη δεδομένα αντλήθηκαν από την Ελληνική Στατιστική Υπηρεσία (ΕΛΣΤΑΤ) για τις αφίξεις σε λιμάνια και από το Ινστιτούτο των Συνδέσμων Ελληνικών Τουριστικών Επιχειρήσεων (ΙΝΣΕΤΕ) για τις αφίξεις τουριστών σε αεροδρόμια. Επιπλέον, μηνιαία στοιχεία ατυχημάτων με τραυματίες και νεκρούς σε οδικά ατυχήματα αντλήθηκαν από τη βάση των αναλυτικών δεδομένων τροχαίων ατυχημάτων του ΕΜΠ, όπως έχουν καταγραφεί στην ΕΛΣΤΑΤ. Όλα τα δεδομένα αφορούν στην περίοδο 2009-2018.

Έπειτα, **αναπτύχθηκαν δύο διαφορετικά μοντέλα για τη συσχέτιση των οδικών ατυχημάτων με τις αφίξεις των τουριστών: α) Generalized Linear Model (GLM) και β) Random Forest**. Το κάθε μοντέλο προέβλεψε τον αριθμό των οδικών ατυχημάτων και θανάτων σε αυτά με βάση τις τουριστικές αφίξεις για τα τελευταία δύο έτη (2017 και 2018), διαβάζοντας τις χρονοσειρές των οκτώ πρώτων ετών της βάσης δεδομένων. Με την χρήση των δεικτών MAE (Mean Absolute Error) και RMSE (Root Mean Squared), τέθηκε δυνατή η σύγκριση των δύο μεθοδολογιών, προκειμένου να ελεγχθεί η ακρίβεια των προβλέψεων. Σημειώνεται ότι τα εξεταζόμενα νησιά χωρίστηκαν σε τέσσερις γεωγραφικές ομάδες (Κυκλάδες, Δωδεκάνησα, Ιόνιο, Κεντρικό και Βόρειο Αιγαίο), για τις οποίες αναπτύχθηκαν ξεχωριστά μοντέλα.

Τα αποτελέσματα που προέκυψαν παρουσιάζονται συγκεντρωτικά στον Πίνακα 6.1. Και οι δυο τύποι μοντέλων έδειξαν ότι για κάθε ομάδα νησιών, οι αφίξεις συσχετίζονται θετικά με τα ατυχήματα και τους θανάτους που σημειώνονται στις υπό μελέτη περιοχές. Αυτό το γεγονός υποδεικνύει ότι η αύξηση των αφίξεων και του τουρισμού, συνεπάγεται αύξηση στα οδικά ατυχήματα και τους θανάτους κατά τις τουριστικές περιόδους.

Ειδικότερα, τα μοντέλα GLM έδειξαν ότι στην περιοχή του Κεντρικού και Βορείου Αιγαίου, η **επιρροή των τουριστικών αφίξεων στον αριθμό των ατυχημάτων είναι μεγαλύτερη** σε σχέση

με τα νησιά του Ιονίου, των Δωδεκανήσων και των Κυκλάδων, το οποίο ενδεχομένως οφείλεται πέρα από τις επικρατούσες συνθήκες (οδική υποδομή, ετοιμότητα πρώτων βοηθειών, κτλ.), και στο γεγονός ότι σε αυτά τα νησιά οι επισκέψεις ξένων τουριστών είναι λιγότερες.

Συγκρίνοντας τα **μοντέλα GLM αφίξεων-θανάτων**, προέκυψε ότι δεν υπάρχει συσχέτιση για την ομάδα νησιών του Κεντρικού και Βορείου Αιγαίου, ενώ στις άλλες τρεις περιοχές που εξετάστηκαν, υπήρχε μία μικρή συσχέτιση αφίξεων – θανάτων, και πιο συγκεκριμένα στις περιοχές των Δωδεκανήσων και των Κυκλάδων. Σημειώνεται ότι ο αριθμός των νεκρών σε οδικά ατυχήματα που καταγράφεται στα νησιά σε μηνιαία βάση είναι μικρός, γεγονός που επηρεάζει την σημαντικότητα των στατιστικών μοντέλων.

Αντίστοιχα μοντέλα αφίξεων-ατυχημάτων και αφίξεων-θανάτων αναπτύχθηκαν με τη μέθοδο **Random Forest**. Σε σχέση με τα μοντέλα για τον αριθμό των θανάτων σε οδικά ατυχήματα, προέκυψε μόνο ένα ικανοποιητικό μοντέλο, για την ομάδα των Δωδεκανήσων με αποδεκτό σφάλμα πρόβλεψης. Για τα δε μοντέλα συσχέτισης αφίξεων ατυχημάτων, προέκυψαν αποδεκτά μοντέλα για όλες τις ομάδες νησιών. Επισημαίνεται παρ’ όλα αυτά, ότι για τις Κυκλάδες και το Κεντρικό και Βόρειο Αιγαίο, το ποσοστό διακύμανσης που εξηγείται είναι σχετικά χαμηλό. Τα μικρότερα σφάλματα πρόβλεψης υπολογίστηκαν για τις Κυκλάδες και τα Δωδεκάνησα, με το Κεντρικό και Βόρειο Αιγαίο και το Ιόνιο να ακολουθούν.

Γενικότερα, συγκρίνοντας τις δύο μεθόδους ως προς τα σφάλματα προβλέψεων, πιο χαμηλοί δείκτες MAE και RMSE παρατηρούνται για τα μοντέλα αφίξεων-ατυχημάτων με τη μέθοδο **Random Forest, η οποία φαίνεται να προβλέπει πιο ικανοποιητικά**.

	Generalized Linear Model (GLM)					
	Fatalities			Crashes		
Ομάδα Νησιών	**Συντελεστής**	**McFadden R^2**	**AICc**	**Συντελεστής**	**McFadden R^2**	**AICc**
Όλα τα νησιά	0.00977 < p = 0.001	0.12	3011.21	0.0148 < p = 0.001	0.10	9580.49
Δωδεκάνησα	0.00909 < p = 0.001	0.16	954.66	0.0123 < p = 0.001	0.12	2638.03
Κυκλάδες	0.01122 < p = 0.001	0.11	628.61	0.0182 < p = 0.001	0.11	2491.14
Ιόνιο	0.00714 < p = 0.001	0.10	634.42	0.0095 < p = 0.001	0.11	1853.99
Κεντρικό/Βόρειο Αιγαίο	0.03354 < p = 0.001	0.06	708.54	0.0574 < p = 0.001	0.10	2227.65

Ομάδα Νησιών	**Fatalities MAE**	**Fatalities RMSE**	**Crashes MAE**	**Crashes RMSE**
Όλα τα νησιά	0.41	1.67	19.66	177.93
Δωδεκάνησα	0.64	2.42	15.40	87.42
Κυκλάδες	0.16	0.34	3.40	15.83
Ιόνιο	0.59	1.07	5.78	18.60
Κεντρικό/Βόρειο Αιγαίο	0.27	0.51	4.35	17.52

	Random Forest			
	Fatalities		Crashes	
Ομάδα Νησιών	**Mean of Squared Residual**	**% of Var explained**	**Mean of Squared Residual**	**% of Var explained**
Όλα τα νησιά	0.286	5.96	4.976	46.71
Δωδεκάνησα	0.454	23.16	5.977	63.15
Κυκλάδες	0.099	-26.53	2.185	16.53
Ιόνιο	0.471	2.84	8.015	53.14
Κεντρικό/Βόρειο Αιγαίο	0.245	-21.47	4.658	15.18

Ομάδα Νησιών	**Fatalities MAE**	**Fatalities RMSE**	**Crashes MAE**	**Crashes RMSE**
Όλα τα νησιά	0.27	0.62	1.41	2.75
Δωδεκάνησα	0.29	0.65	1.06	2.10
Κυκλάδες	0.16	0.40	0.96	1.91
Ιόνιο	0.40	0.72	1.92	3.42
Κεντρικό/Βόρειο Αιγαίο	0.24	0.50	1.73	3.26

Πίνακας 6.1: Συγκεντρωτικός Πίνακας Αποτελεσμάτων

6.2 Συνολικά Συμπεράσματα

Σε αυτήν την υποενότητα παρουσιάζονται τα βασικά συμπεράσματα που προέκυψαν από την ανάλυση των αποτελεσμάτων της στατιστικής ανάλυσης.

1. Ο **τουρισμός, και γενικότερα οι αφίξεις σε κάποιο νησί, συσχετίζονται θετικά με τον αριθμό των ατυχημάτων και των νεκρών** σε οδικά ατυχήματα που καταγράφονται σε εκείνη την χρονική περίοδο.

2. Παρ΄ όλο που οι αφίξεις συσχετίστηκαν με τα ατυχήματα σε όλες τις ομάδες νησιών που εξετάστηκαν, τα αποτελέσματα διέφεραν. Είναι πιθανό σε **νησιά που εμφάνιζαν μικρό αριθμό ατυχημάτων και μεγάλο αριθμό αφίξεων**, να επικρατούν καλύτερες οδικές συνθήκες, είτε διότι το οδικό δίκτυο είναι καταλληλότερο είτε διότι η συμπεριφορά των οδηγών (ντόπιων και τουριστών) είναι καλύτερη.

3. Στην περιοχή του **Κεντρικού και Βορείου Αιγαίου**, η επιρροή των τουριστικών αφίξεων στον αριθμό των ατυχημάτων είναι μεγαλύτερη σε σχέση με τα νησιά του Ιονίου, των Δωδεκανήσων και των Κυκλάδων. Αυτό ενδεχομένως οφείλεται τόσο στις επικρατούσες συνθήκες (πιθανώς χειρότερη οδική υποδομή, ακατάλληλη συμπεριφορά οδηγών, ελλιπής αστυνόμευση, κτλ.), τόσο και στο γεγονός ότι σε αυτά τα νησιά οι επισκέψεις ξένων τουριστών είναι λιγότερες. Είναι γνωστό και από τη βιβλιογραφία ότι οι ξένοι τουρίστες τείνουν να προσαρμόζονται δυσκολότερα σε ένα άγνωστο σε αυτούς οδικό περιβάλλον, επομένως, πιθανώς στα συγκεκριμένα νησιά, λόγω χαμηλότερου ξένου τουρισμού, να μην έχουν ληφθεί τα κατάλληλα μέτρα για τη σωστή προσαρμογή τους.

4. Οι **αφίξεις επηρεάζουν σε μικρότερο βαθμό τους θανάτους** που καταγράφονται εκείνη την περίοδο, με τα μοντέλα να δείχνουν ότι σε κάποιες περιπτώσεις δεν υπάρχει συσχέτιση αυτών των δύο μεταβλητών. Αυτό πιθανώς οφείλεται στο γεγονός ότι ο αριθμός των νεκρών σε οδικά ατυχήματα που καταγράφονται σε μηνιαία βάση στα νησιά είναι σημαντικά μικρός, το οποίο δεν επιτρέπει την ανάπτυξη στατιστικά σημαντικών μοντέλων. Παρ’ όλα αυτά, η μικρή συσχέτιση αφίξεων και θανάτων πιθανώς υποδεικνύει την ύπαρξη και άλλων παραγόντων που συμβάλλουν στη σοβαρότητα των ατυχημάτων, όπως επικίνδυνες συμπεριφορές (υψηλές ταχύτητες, οδήγηση υπό την επήρεια αλκοόλ, κτλ.), οδικές υποδομές, παροχή πρώτων βοηθειών και περίθαλψης μετά το ατύχημα, κτλ., οι οποίοι θα μπορούσαν να διερευνηθούν.

5. Τέλος, το γεγονός ότι τα ατυχήματα αυξάνονται με τις αφίξεις, άλλα όχι οι θάνατοι, οδηγεί στο συμπέρασμα ότι, πιθανώς στην πλειοψηφία των περιπτώσεων, στα **ατυχήματα αυτά**

εμπλέκονται οδηγοί που δυσκολεύονται να αφομοιώσουν τις κυκλοφοριακές συνθήκες της περιοχής, ωστόσο ο τρόπος οδήγησης τους δεν είναι τόσο επικίνδυνος ώστε να οδηγήσει σε πολύ σοβαρό ατύχημα.

6.3 Προτάσεις για περαιτέρω έρευνα

Η παρούσα Εργασία θα μπορούσε να ενισχυθεί μελλοντικά, προσθέτοντας και **επιπλέον μεταβλητές** πέρα από τις αφίξεις, ώστε να μελετηθεί σε μεγαλύτερο βάθος το ποσοστό που αυτές επηρεάζουν τα ατυχήματα και τους θανάτους στα ελληνικά νησιά.

Επίσης, παρατηρήθηκε ότι **το σύνολο το θανάτων** δεν έδωσε σαφή συμπεράσματα ως προς την συσχέτισή τους με τις αφίξεις στα νησιά. Παρουσιάζει ενδιαφέρον, ωστόσο, η δυνατότητα να συσχετιστούν οι θάνατοι σε συνδυασμό με τους βαριά τραυματίες που σημειώνονται σε μηνιαία βάση στις περιοχές, ώστε να μελετηθεί εάν τα μοντέλα εμφανίζουν μεγαλύτερη συσχέτιση μεταξύ αυτών των δύο μεταβλητών και των αφίξεων.

Τέλος, η μελέτη θα μπορούσε να ενισχυθεί, εάν εξίσου συμπεριληφθούν τα δεδομένα αφίξεων των υπόλοιπων νομών της Ελλάδας, ώστε να συγκριθούν οι περιοχές της **ηπειρωτικής με εκείνες της νησιωτικής Ελλάδας** ως προς τη συσχέτιση των αφίξεων με τα ατυχήματα και τους θανάτους.

6.4 Προτάσεις προς την Πολιτεία

Η Πολιτεία θα μπορούσε να συμβάλει στην προτροπή των ατυχημάτων ή, ακόμα, και των θανάτων, μέσω της λήψης μέτρων για την **κατάλληλη προσαρμογή των οδικών συνθηκών στις τουριστικές περιοχές**.

Αυτά περιλαμβάνουν την τοποθέτηση κατάλληλης και ευδιάκριτης από τους χρήστες της οδού, σήμανσης, στις πιο κρίσιμες περιοχές του κάθε τόπου. Επιπλέον, ένα ασφαλές οδικό περιβάλλον θα ελαχιστοποιούσε σε μεγάλο βαθμό τα ατυχήματα, οπότε η ποιότητα των οδοστρωμάτων και η ύπαρξη φωτισμού στις οδούς επιβάλλονται για κάθε τουριστικό προορισμό. Τέλος, η ενίσχυση της **αστυνόμευσης** σε αυτούς τους προορισμούς θα διασφάλιζε την οδική ασφάλεια σε επισκέπτες και ντόπιους.

Κεφάλαιο 7: Βιβλιογραφικές Αναφορές

1. Ελληνική Στατιστική Αρχή, https://www.statistics.gr/

2. INSETE – Ινστιτούτο Συνδέσμου Ελληνικών Τουριστικών Επιχειρήσεων, https://insete.gr/

3. National Technical University of Athens, Road Safety Observatory, https://www.nrso.ntua.gr/

4. Bellos V., Ziakopoulos A., Yannis G., "Investigation of the effect of tourism on road crashes". Journal of Transportation Safety & Security, Vol.12, Issue 6, January 2019, pp. 782-799.

5. Nikolaou D., Folla K., Bellos E., Yannis G., "Tourism and Road Accidents in Greece", Proceedings of the 9th International Congress on Transportation Research, organised by The Hellenic Institute of Transportation Engineers (HITE) and the Hellenic Institute of Transport (HIT/CERTH) (Athens, Greece, 24-25 October 2019)

6. Ziakopoulos A., Michelaraki E., Nikolaou D., Folla K., Yannis G. (2023), "Association Rule Mining for Island and Mainland Road Crash Injuries in Greece", Transport Research Arena (TRA) Conference

7. Linda Walker & Stephen J. Page (2004) "The Contribution of Tourists and Visitors to Road Traffic Accidents: A Preliminary Analysis of Trends and Issues for Central Scotland", Current Issues in Tourism, 7:3, 217-241

8. Page, S. J., & Meyer, D. (1996). "Tourist accidents", Annals of Tourism Research, 23(3), 666–690

Παράρτημα: Κώδικας GLM και Random Forest στην R

Εισαχθέντα δεδομένα στην R:

	A	B	C	D	E
1	**Date**	**Island**	**Fatalities**	**Crashes**	**Arrivals**
2	2009-01-0	Limnos	0	1	5416
3	2009-02-0	Limnos	0	1	4611
4	2009-03-0	Limnos	0	0	5500
5	2009-04-0	Limnos	0	0	8303
6	2009-05-0	Limnos	0	0	7814
7	2009-06-0	Limnos	0	1	12922
8	2009-07-0	Limnos	0	1	23288
9	2009-08-0	Limnos	0	0	24284
10	2009-09-0	Limnos	0	0	10260
11	2009-10-0	Limnos	0	0	7267
12	2009-11-0	Limnos	0	0	4915
13	2009-12-0	Limnos	0	0	5078

Κώδικας GLM:

```
#START #####
rm(list=ls())
Sys.sleep(0.1)
options(scipen = 7)

# Install necessary packages
#install.packages("readxl")
#install.packages("dplyr")
# Load the required libraries
library(readxl)
library(dplyr)

# Read the Excel file
db <- read_excel("Dodekanisa.xlsx")

# View the first few rows of the data
head(db)
# Convert the 'Date' column to Date format
db$Date <- as.Date(db$Date)
# Check the structure of the data
str(db)
```

```
# Fit a GLM model (Poisson or Negative Binomial) using the training data
# Poisson or NB regression #####
#Fatalities
mesos <- mean(train_data$Fatalities)
diak <- var(train_data$Fatalities)
if (diak > mesos) print("Negative Binomial") else print("Poisson")
#Crashes
mesos <- mean(train_data$Crashes)
diak <- var(train_data$Crashes)
if (diak > mesos) print("Negative Binomial") else print("Poisson")
```

```
library(MASS)
negb <- glm.nb(Fatalities ~ Arrivals, data = train_data)
summary(negb)
#For McFadden R^2
library(pscl)
pR2(negb)
#AICc
library(MuMIn)
AICc(negb)
```

```r
# Define the training and test period
train_end_date <- as.Date("2016-12-31")  # 8 years of data
test_start_date <- as.Date("2017-01-01")  # Next 2 years of data

# Split into training and test sets
train_data <- db[db$Date <= train_end_date, ]
test_data <- db[db$Date > train_end_date, ]

# Check the number of rows in each dataset
cat("Training Data Rows:", nrow(train_data), "\n")
cat("Test Data Rows:", nrow(test_data), "\n")
```

```r
# Make predictions on the test data
test_data$Predicted_Fatalities <- predict(negb, newdata = test_data, type = "response")

# View the first few predictions
head(test_data[, c("Date", "Fatalities", "Predicted_Fatalities")])

# Evaluate model performance on the test data

# Calculate Mean Absolute Error (MAE)
mae_value <- mean(abs(test_data$Fatalities - test_data$Predicted_Fatalities))
cat("Mean Absolute Error (MAE):", mae_value, "\n")

# Calculate Root Mean Square Error (RMSE)
rmse_value <- sqrt(mean((test_data$Fatalities - test_data$Predicted_Fatalities)^2))
cat("Root Mean Square Error (RMSE):", rmse_value, "\n")
```

Κώδικας GLM με λύση για τα Δωδεκάνησα:

```
> # Define the training and test period
> train_end_date <- as.Date("2016-12-31")  # 8 years of data
> test_start_date <- as.Date("2017-01-01")  # Next 2 years of data
> # Split into training and test sets
> train_data <- db[db$Date <= train_end_date, ]
> test_data <- db[db$Date > train_end_date, ]
> # Check the number of rows in each dataset
> cat("Training Data Rows:", nrow(train_data), "\n")
Training Data Rows: 960
> cat("Test Data Rows:", nrow(test_data), "\n")
Test Data Rows: 240
> # Fit a GLM model (Poisson or Negative Binomial) using the training data
> # Poisson or NB regression #####
> #Fatalities
> mesos <- mean(train_data$Fatalities)
> diak <- var(train_data$Fatalities)
> if (diak > mesos) print("Negative Binomial") else print("Poisson")
[1] "Negative Binomial"
> #Crashes
> mesos <- mean(train_data$Crashes)
> diak <- var(train_data$Crashes)
> if (diak > mesos) print("Negative Binomial") else print("Poisson")
[1] "Negative Binomial"
```

```
#CRASHES
negb <- glm.nb(Crashes ~ Arrivals, data = train_data)
summary(negb)
#For McFadden R^2
library(pscl)
pR2(negb)
#AICc
library(MuMIn)
AICc(negb)

# Make predictions on the test data
test_data$Predicted_Crashes <- predict(negb, newdata = test_data, type = "response")

# View the first few predictions
head(test_data[, c("Date", "Crashes", "Predicted_Crashes")])

# Evaluate model performance on the test data

# Calculate Mean Absolute Error (MAE)
mae_value <- mean(abs(test_data$Crashes - test_data$Predicted_Crashes))
cat("Mean Absolute Error (MAE):", mae_value, "\n")

# Calculate Root Mean Square Error (RMSE)
rmse_value <- sqrt(mean((test_data$Crashes - test_data$Predicted_Crashes)^2))
cat("Root Mean Square Error (RMSE):", rmse_value, "\n")
```

```
# A tibble: 6 × 3
  Date       Crashes Predicted_Crashes
  <date>       <dbl>             <dbl>
1 2017-01-01       5              1.01
2 2017-02-01       4              0.952
3 2017-03-01       4              1.07
4 2017-04-01      10              3.85
5 2017-05-01       5             28.7
6 2017-06-01      17            150.
> # Calculate Mean Absolute Error (MAE)
> mae_value <- mean(abs(test_data$Crashes - test_data$Predicted_Crashes))
> cat("Mean Absolute Error (MAE):", mae_value, "\n")
Mean Absolute Error (MAE): 15.40126
> # Calculate Root Mean Square Error (RMSE)
> rmse_value <- sqrt(mean((test_data$Crashes - test_data$Predicted_Crashes)^2))
> cat("Root Mean Square Error (RMSE):", rmse_value, "\n")
Root Mean Square Error (RMSE): 87.42191
```

```
> # Make predictions on the test data
> test_data$Predicted_Fatalities <- predict(negb, newdata = test_data, type = "response")
> # View the first few predictions
> head(test_data[, c("Date", "Fatalities", "Predicted_Fatalities")])
# A tibble: 6 × 3
  Date       Fatalities Predicted_Fatalities
  <date>          <dbl>                <dbl>
1 2017-01-01          1                0.161
2 2017-02-01          2                0.154
3 2017-03-01          2                0.168
4 2017-04-01          2                0.433
5 2017-05-01          0                1.92
6 2017-06-01          1                6.50
> # Calculate Mean Absolute Error (MAE)
> mae_value <- mean(abs(test_data$Fatalities - test_data$Predicted_Fatalities))
> cat("Mean Absolute Error (MAE):", mae_value, "\n")
Mean Absolute Error (MAE): 0.6449612
> # Calculate Root Mean Square Error (RMSE)
> rmse_value <- sqrt(mean((test_data$Fatalities - test_data$Predicted_Fatalities)^2))
> cat("Root Mean Square Error (RMSE):", rmse_value, "\n")
Root Mean Square Error (RMSE): 2.424423
```

Κώδικας Random Forest:

```r
# START #####
rm(list=ls()) #REMOVE FOR CLEAN SLATE !!!
Sys.sleep(0.1) #Breather
options(scipen = 7)

# Install necessary packages
#install.packages("readxl")
#install.packages("dplyr")
# Install randomForest package
#install.packages("randomForest")

# Load the required libraries
library(readxl)
library(dplyr)
# Load the randomForest library
library(randomForest)

# Read the Excel file
db <- read_excel("bodekarisa.xlsx")

# View the first few rows of the data
head(db)
# Convert the "Date" column to Date format
db$Date <- as.Date(db$Date)

# Check the structure of the data
str(db)
# Define the training and test periods
train_end_date <- as.Date("2016-12-31")  # 8 years of data
test_start_date <- as.Date("2017-01-01") # Next 2 years of data

# Split into training and test sets
train_data <- db[db$Date <= train_end_date, ]
test_data <- db[db$Date > train_end_date, ]

# Check the number of rows in each dataset
cat("Training Data Rows:", nrow(train_data), "\n")
cat("Test Data Rows:", nrow(test_data), "\n")
```

```r
# Fit Random Forest model on training data
rf_model <- randomForest(Fatalities ~ Arrivals, data = train_data, ntree = 500)

# View the model summary
print(rf_model)

# Make predictions on the test data
rf_predictions <- predict(rf_model, newdata = test_data)

# Combine predictions with actual values for comparison
rf_results <- data.frame(
  Date = test_data$Date,
  Actual_Fatalities = test_data$Fatalities,
  Predicted_Fatalities = rf_predictions
)

# View the first few rows of the results
print(head(rf_results))

# Calculate RMSE
rf_rmse <- sqrt(mean((rf_results$Actual_Fatalities - rf_results$Predicted_Fatalities)^2))
cat("Random Forest RMSE:", rf_rmse, "\n")

# Calculate MAE
rf_mae <- mean(abs(rf_results$Actual_Fatalities - rf_results$Predicted_Fatalities))
cat("Random Forest MAE:", rf_mae, "\n")
```

```r
#CRASHES
rf_model <- randomForest(Crashes ~ Arrivals, data = train_data, ntree = 300)
print(rf_model)

rf_predictions <- predict(rf_model, newdata = test_data)

rf_results <- data.frame(
  Date = test_data$Date,
  Actual_Crashes = test_data$Crashes,
  Predicted_Crashes = rf_predictions
)
print(head(rf_results))

rf_rmse <- sqrt(mean((rf_results$Actual_Crashes - rf_results$Predicted_Crashes)^2))
cat("Random Forest RMSE:", rf_rmse, "\n")

rf_mae <- mean(abs(rf_results$Actual_Crashes - rf_results$Predicted_Crashes))
cat("Random Forest MAE:", rf_mae, "\n")
```