



**ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ
ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ
ΤΜΗΜΑ ΠΟΛΙΤΙΚΩΝ ΜΗΧΑΝΙΚΩΝ**

Διπλωματική Εργασία

ΑΝΑΛΥΣΗ ΚΑΙ ΠΡΟΒΛΕΨΗ ΟΔΙΚΩΝ ΑΤΥΧΗΜΑΤΩΝ ΜΕΣΩ ΧΡΟΝΟΣΕΙΡΩΝ

ΠΑΥΛΟΣ ΓΚΟΥΓΚΟΥΛΗΣ



Υπεβλήθη για την εκπλήρωση μέρους των απαιτήσεων για την απόκτηση του
Διπλώματος Πολιτικού Μηχανικού

ΒΟΛΟΣ 2025

© 2025 Πάυλος Γκουγκουλής

Η έγκριση της διπλωματικής εργασίας από το Τμήμα Πολιτικών Μηχανικών της Πολυτεχνικής Σχολής του Πανεπιστημίου Θεσσαλίας δεν υποδηλώνει αποδοχή των απόψεων του συγγραφέα (Ν. 5343/32 αρ. 202 παρ. 2).

Εγκρίθηκε από τα Μέλη της Τριμελούς Εξεταστικής Επιτροπής:

Πρώτος Εξεταστής (Επιβλέπων)	Δρ. Θεοφιλάτος Αθανάσιος Επίκουρος Καθηγητής, Τμήμα Πολιτικών Μηχανικών, Πανεπιστήμιο Θεσσαλίας
Δεύτερος Εξεταστής	Δρ. Φράγκου Αθανάσιος Μέλος Εργαστηριακού Διδακτικού Προσωπικού (ΕΔΙΠ), Τμήμα Πολιτικών Μηχανικών, Πανεπιστήμιο Θεσσαλίας
Τρίτος Εξεταστής	Δρ. Σπηλιωτόπουλος Μάριος Μέλος Εργαστηριακού Διδακτικού Προσωπικού (ΕΔΙΠ), Τμήμα Πολιτικών Μηχανικών, Πανεπιστήμιο Θεσσαλίας

Ευχαριστίες

Πρώτα απ' όλα, θέλω να ευχαριστήσω τον επιβλέποντα της διπλωματικής εργασίας μου, Επίκουρο Καθηγητή κ. Θεοφιλάτο Αθανάσιο, για την πολύτιμη βοήθεια και καθοδήγησή του κατά τη διάρκεια της δουλειάς μου. Επίσης, είμαι ευγνώμων στα υπόλοιπα μέλη της εξεταστικής επιτροπής, Καθηγητές κκ. Φράγκου Αθανάσιος και Σπηλιωτόπουλος Μάριος για την ανάγνωση της εργασίας μου και για τις πολύτιμες υποδείξεις τους. Οφείλω επίσης ευχαριστίες στον Καθηγητή κ. Γιαννή Γιώργο του Τομέα Μεταφορών στο Τμήμα Πολιτικών Μηχανικών του Εθνικού Μετσόβιου Πολυτεχνείου, για την παροχή των αναλυτικών δεδομένων Τροχαίων Ατυχημάτων περιόδου 1996 - 2022. Ευχαριστώ επίσης από καρδιάς, την κοπέλα μου Κουστίκα Ευθυμία για τις πολύτιμες συμβουλές της, την υπομονή, υποστήριξη και την συμπαράστασή της όλα αυτά τα χρόνια για πράγματα που χάραξαν καθοριστικά την πορεία μου. Μεγάλες ευχαριστίες πηγαίνουν και στους φίλους μου ονομαστικά: Γιώργος, Βαγγελιώ, Μαρία, Αριστέα, Βαγγέλης, Άννα, Άρης, Νίκη, Φίλιππος για όσα μου έχουν προσφέρει από την στιγμή που τους γνώρισα. Ένα μεγάλο ευχαριστώ οφείλω και στους συμφοιτητές και συνεργάτες μου, Κυριαζή Κάλλη, Δημήτριο Γιαλούρη και Δημήτριο Βαϊνά, για τη στήριξη, την παρέα και τη συνεργασία μας όλα αυτά τα χρόνια. Τέλος, και πάνω απ' όλα, είμαι ευγνώμων στους γονείς μου, Αναστασία Καλαμπούκα και Ιωάννη Γκουγκουλή για την ολόψυχη αγάπη και υποστήριξή τους. Αφιερώνω αυτή την εργασία σε όλους τους παραπάνω.

Πάυλος Γκουγκουλής

ΑΝΑΛΥΣΗ ΚΑΙ ΠΡΟΒΛΕΨΗ ΟΔΙΚΩΝ ΑΤΥΧΗΜΑΤΩΝ ΜΕΣΩ ΧΡΟΝΟΣΕΙΡΩΝ

Παύλος Γκουγκουλής

Πανεπιστήμιο Θεσσαλίας, Τμήμα Πολιτικών Μηχανικών, 2025

Επιβλέπων Καθηγητής: Θεοφύλατος Αθανάσιος, Επίκουρος Καθηγητής

Περίληψη

Το φαινόμενο των τροχαίων ατυχημάτων παραμένει μια από τις σοβαρότερες απειλές για τη δημόσια υγεία στην Ελλάδα και διεθνώς. Η παρούσα εργασία επικεντρώνεται στην ανάλυση και πρόβλεψη των τροχαίων ατυχημάτων και θανάτων, αξιοποιώντας χρονοσειρές εβδομαδιαίων και μηνιαίων δεδομένων από το 1996 έως το 2022. Μέσω εμπλουτισμού των δεδομένων με κοινωνικοοικονομικούς δείκτες, εποχικά χαρακτηριστικά και πολιτικές μεταβλητές, εφαρμόζονται και συγκρίνονται στατιστικά μοντέλα ARIMA / SARIMA και νευρωνικά δίκτυα τύπου LSTM. Η αξιολόγηση των μοντέλων πραγματοποιείται με περιγραφικές και ποσοτικές μεθόδους, αναδεικνύοντας την υπεροχή του LSTM στη βραχυπρόθεσμη πρόβλεψη. Η εργασία αναδεικνύει τη δυναμική των σύγχρονων μεθόδων πρόβλεψης στη μοντελοποίηση σύνθετων κοινωνικών φαινομένων. Τα αποτελέσματα παρέχουν χρήσιμο υπόβαθρο για αναλυτικά εργαλεία υποστήριξης λήψης αποφάσεων σε ζητήματα οδικής ασφάλειας και πρόληψης ατυχημάτων.

Λέξεις Κλειδιά: *Τροχαία Ατυχήματα, Πρόβλεψη, Χρονοσειρές, ARIMA, SARIMA, LSTM, Κοινωνικοοικονομικά Δεδομένα*

ANALYSIS AND PREDICTION OF ROAD ACCIDENTS USING TIME SERIES

Pavlos Gkougkoulis

University of Thessaly, Department of Civil Engineering, 2025

Supervisor: Theofilatos Athanasios, Assistant Professor

Abstract

Road traffic accidents remain one of the most significant threats to public health in Greece and worldwide. This study focuses on the analysis and forecasting of traffic accidents and fatalities using time series data on a weekly and monthly basis from 1996 to 2022. By enriching the dataset with socioeconomic indicators, seasonal patterns, and policy-related variables, statistical models ARIMA / SARIMA and LSTM neural networks are applied and compared. Model evaluation is conducted using both descriptive and quantitative methods, highlighting the superiority of the LSTM approach in short-term forecasting. The study demonstrates the potential of modern predictive methods in modeling complex social phenomena. The results provide a valuable foundation for analytical tools that support data-driven decision-making in traffic safety policy and accident prevention.

Keywords: *Traffic Accidents, Forecasting, Time Series, ARIMA, SARIMA, LSTM, Socioeconomic Data*

Πίνακας Περιεχομένων

Κεφάλαιο 1	Εισαγωγή	1
1.1	Διατύπωση του Προβλήματος	2
1.2	Στόχοι της Διπλωματικής Εργασίας	2
1.3	Πεδίο Εφαρμογής και Περιορισμοί	3
Κεφάλαιο 2	Βιβλιογραφική Ανασκόπηση	4
2.1	Τάσεις Τροχαίων Ατυχημάτων: Παγκόσμια και Ελληνική Πραγματικότητα	4
2.2	Προσδιοριστικοί Παράγοντες Τροχαίων Ατυχημάτων	5
2.3	Τιμές Καυσίμων, Οικονομικοί Δείκτες και Οδική Ασφάλεια	7
2.4	Μοντελοποίηση Χρονοσειρών Τροχαίων Ατυχημάτων	9
2.4.1	Κλασικά Μοντέλα: ARIMA, SARIMA	9
2.4.2	Μετρικές Αξιολόγησης στην Πρόβλεψη Χρονοσειρών (ARIMA / SARIMA)	11
2.5	Μηχανική Μάθηση για Προβλέψεις Χρονοσειρών	15
2.5.1	LSTM Και RNNs	15
2.5.2	Μετρικές Αξιολόγησης στην Πρόβλεψη Χρονοσειρών (LSTM)	21
Κεφάλαιο 3	Συλλογή και Διαχείριση Δεδομένων	25
3.1	Ανάκτηση Δεδομένων από την ΒΔ SANTRA	25
3.1.1	Σχόλια και Περιγραφή των Δεδομένων από την ΒΔ SANTRA	26
3.2	Συλλογή Δεδομένων από Διάφορες Πηγές για Κοινωνικοοικονομικούς Παράγοντες	29
3.2.1	Σχόλια και Περιγραφή των Κοινωνικοοικονομικών Δεδομένων από Διάφορες Πηγές	30
3.3	Αρχικοί Μετασχηματισμοί στα Δεδομένα Πριν την Εισαγωγή τους στη ΒΔ	34
3.4	Δημιουργία Βάσης Δεδομένων	36
3.4.1	Σχέσεις Μεταξύ Μεταβλητών και Διάκριση των Data Types	36
3.4.2	Διάγραμμα Οντοτήτων-Συσχετίσεων (ER) της Βάσης Δεδομένων	37
3.4.3	Εισαγωγή Δεδομένων στη ΒΔ RoadAccidents	38
3.4.4	Αξιολόγηση της Συμπλήρωσης Ελλιπών Τιμών και Αιτιολόγηση της Απόρριψής της	40
3.4.5	Queries Εξαγωγής Δεδομένων και Τελικά Datasets	41

Κεφάλαιο 4	Προεπεξεργασία Δεδομένων και Feature Engineering	44
4.1	Εισαγωγή Δεδομένων και ορισμός του Working Directory	44
4.2	Βασικός Καθαρισμός και Μετατροπές Τύπων Μεταβλητών	45
4.3	Δημιουργία Χρονικών Δεικτών και Ενοποίηση Ημερομηνιών	45
4.4	Διαδικασία και Μεθοδολογία του Feature Engineering	46
4.4.1	Season, Quarter, isSummer, isWinter	46
4.4.2	Holidays and Special Events	47
4.4.3	Week / Month / Year Lags	47
4.4.4	Week/Month/Year Rolling Averages	48
4.4.5	Επιπλέον Δείκτες και Binary Flags για Νομοθεσίες	48
4.4.6	Cumulative Features	50
4.4.7	Αφαίρεση Μεταβλητών με Missing Values και Incomplete Data Entire	50
Κεφάλαιο 5	Exploratory Data Analysis (EDA)	52
5.1	Περιγραφική Στατιστική	52
5.1.1	Εβδομαδιαία Δεδομένα	54
5.1.2	Μηνιαία Δεδομένα	54
5.1.3	Ετήσια Δεδομένα	55
5.2	Επαγωγική Στατιστική	55
5.2.1	Εβδομαδιαία Δεδομένα	56
5.2.2	Μηνιαία Δεδομένα	58
5.2.3	Ετήσια Δεδομένα	59
5.3	Επιπλέον Στατιστικές Αναλύσεις και Plots	60
5.3.1	Time Series Plots	60
5.3.2	Plots Επιπλέον Περιγραφικών Στατιστικών	68
5.3.3	Plots Ατυχημάτων σε Συσχέτιση με τις Εποχές και τις Αργίες	73
5.3.4	Correlation Matrices	76
Κεφάλαιο 6	Στατιστική Μοντελοποίηση Χρονοσειρών	81
6.1	Μεθοδολογία και Workflow για την Στατιστική Μοντελοποίηση	82
6.2	Διαδικασία Επιλογής και Εξέτασης Μοντέλου Πρόβλεψης	83
6.2.1	Δημιουργία Αντικειμένου Time Series στην R	85
6.2.2	Διαγνωστικά Plots και Tests και Seasonal / Normal Differencing	86

6.2.3	Manual Fitting και Fitting με Auto.Arima()	89
6.2.4	Residuals Checks	91
6.2.5	Grid Searches	91
6.2.6	Log Transformations	92
6.3	Αποτελέσματα από την Μεθοδολογία Επιλογής Μοντέλων ARIMA / SARIMA	93
6.4	Τελική Εκτίμηση των Στατιστικών Μοντέλων και Προβλέψεις	101
6.4.1	ARIMA Μοντέλο για Πρόβλεψη Εβδομαδιαίων Τροχαίων Ατυχημάτων	101
6.4.2	SARIMA Μοντέλο για Πρόβλεψη Εβδομαδιαίων Τροχαίων Ατυχημάτων	104
6.4.3	ARIMA Μοντέλο για Πρόβλεψη Μηνιαίων Τροχαίων Ατυχημάτων	107
6.4.4	SARIMA Μοντέλο για Πρόβλεψη Μηνιαίων Τροχαίων Ατυχημάτων	109
6.4.5	Grid Search Αποτέλεσμα για τα Μηνιαία Δεδομένα Τροχαίων Ατυχημάτων	112
6.4.6	Επιπλέον Πίνακες και Μετρικές για τα Μοντέλα Τροχαίων Ατυχημάτων	115
6.4.7	ARIMA Μοντέλο για Πρόβλεψη Εβδομαδιαίων Θανάτων από Τροχαία Ατυχήματα	119
6.4.8	SARIMA Μοντέλο για Πρόβλεψη Εβδομαδιαίων Θανάτων από Τροχαία Ατυχήματα	121
6.4.9	ARIMA Μοντέλο για Πρόβλεψη Μηνιαίων Θανάτων από Τροχαία Ατυχήματα	124
6.4.10	SARIMA Μοντέλο για Πρόβλεψη Μηνιαίων Θανάτων από Τροχαία Ατυχήματα	126
6.4.11	Grid Search Αποτελέσματα για Εβδομαδιαία και Μηνιαία Δεδομένα Θανάτων από Τροχαία Ατυχήματα	128
6.4.12	Επιπλέον Πίνακες και Μετρικές για τα Μοντέλα Θανάτων από Τροχαία Ατυχήματα	133
Κεφάλαιο 7	Υλοποίηση NN LSTM	139
7.1	Μεθοδολογία Υλοποίησης NN LSTM	140
7.1.1	Dataset Preparation και Hyperparameter Grid Search	141
7.1.2	Training και Predictions του LSTM Βάση των Βέλτιστων Hyperparameters	143
7.1.3	Αναδρομική Πρόβλεψη και Στάσιμα Σύνολα Δεδομένων	145
7.1.4	One-Step vs Multi-Step Forecasting	146
7.2	Αποτελέσματα Εφαρμογής NN LSTM	147
7.2.1	One-Step Πρόβλεψη Εβδομαδιαίων Τροχαίων Ατυχημάτων	148
7.2.2	One-Step Πρόβλεψη Εβδομαδιαίων Θανάτων από Τροχαία Ατυχήματα	152
7.2.3	One-Step Πρόβλεψη Μηνιαίων Τροχαίων Ατυχημάτων	156
7.2.4	One-Step Πρόβλεψη Μηνιαίων Θανάτων από Τροχαία Ατυχήματα	160
7.3	Υλοποίηση του Recursive Forecasting	164
7.3.1	Προβλέψεις Εβδομαδιαίων Τροχαίων Ατυχημάτων	165
7.3.2	Προβλέψεις Εβδομαδιαίων Θανάτων από Τροχαία Ατυχήματα	166
7.3.3	Προβλέψεις Μηνιαίων Τροχαίων Ατυχημάτων	167

7.3.4	Προβλέψεις Μηνιαίων Θανάτων από Τροχαία Ατυχήματα	168
Κεφάλαιο 8 Σύγκριση Αποτελεσμάτων		169
8.1	ARIMA vs SARIMA Μοντέλα	170
8.2	ARIMA / SARIMA Μοντέλα vs LSTM NN	171
8.3	Ακρίβεια Προβλέψεων στα Εβδομαδιαία vs Μηνιαία Dataset	172
Κεφάλαιο 9 Συμπεράσματα και Προτάσεις για Μελλοντική Έρευνα		173
9.1	Περιορισμοί της Μελέτης	174
9.2	Προτάσεις για Μελλοντική Έρευνα	175
Βιβλιογραφία		177
ΠΑΡΑΡΤΗΜΑ: Ενδεικτικά Σημεία Κώδικα		179
9.3	SQL Queries Εισαγωγής Δεδομένων στην ΒΔ	179
9.3.1	Εισαγωγή Εβδομαδιαίων Δεδομένων στην ΒΔ	179
9.3.2	Aggregate Μηνιαίων Δεδομένων στην ΒΔ	179
9.3.3	Aggregate Ετήσιων Δεδομένων στην ΒΔ	180
9.4	SQL Queries Εξαγωγής Δεδομένων στην ΒΔ	180
9.4.1	Εξαγωγή Εβδομαδιαίου Dataset Τροχαίων Ατυχημάτων	180
9.4.2	Εξαγωγή Εβδομαδιαίου Dataset Τροχαίων Ατυχημάτων με Κοινωνικοοικονομικούς Παράγοντες	180
9.5	Ενδεικτικά Σημεία Κώδικα στην R	181
9.5.1	Ενδεικτικός Κώδικας Εισαγωγής Dummy Variables για Νομοθεσίες	181
9.5.2	Συνάρτηση Υπολογισμού Descriptive Statistics	181
9.5.3	Ενδεικτικό Block Κώδικα για Manual ARIMA / SARIMA Model Selection	182
9.5.4	Ενδεικτικό Block Κώδικα για Εφαρμογή της Auto.Arima() Συνάρτησης	183
9.5.5	Ενδεικτικό Block Κώδικα για τα Residual Checks	184
9.5.6	Ενδεικτικό Block Κώδικα για Grid Search	184
9.6	Ενδεικτικά Σημεία Κώδικα στην Python	185
9.6.1	Ενδεικτικό Hyperparameter Grid και Αρχιτεκτονική LSTM NN	185
9.6.2	Github Repository	187

Κατάλογος Πινάκων

Πίνακας 3.2.1-1: Κοινωνικοοικονομικά Δεδομένα και Πηγές (α)	31
Πίνακας 3.2.1-2: Κοινωνικοοικονομικά Δεδομένα και Πηγές (β)	32
Πίνακας 6.2.6-1: Αποτελέσματα Καλύτερων Μοντέλων Ανάλογα με τον Τρόπο Fitting για το Εβδομαδιαίο Dataset Ατυχημάτων και Θανάτων	94
Πίνακας 6.2.6-2: Αποτελέσματα Καλύτερων Μοντέλων Ανάλογα με τον Τρόπο Fitting για το Μηνιαίο Dataset Ατυχημάτων και Θανάτων	95
Πίνακας 6.2.6-3: Πίνακας Μετρικών Τελικών Μοντέλων ARIMA / SARIMA Τροχαίων Ατυχημάτων από το Εβδομαδιαίο Dataset	96
Πίνακας 6.2.6-4: Πίνακας Μετρικών Τελικών Μοντέλων ARIMA / SARIMA Θανάτων από Τροχαία Ατυχήματα από το Εβδομαδιαίο Dataset	97
Πίνακας 6.2.6-5: Πίνακας Μετρικών Τελικών Μοντέλων ARIMA / SARIMA Τροχαίων Ατυχημάτων από το Μηνιαίο Dataset	99
Πίνακας 6.2.6-6 Πίνακας Μετρικών Τελικών Μοντέλων ARIMA / SARIMA Θανάτων από Τροχαία Ατυχήματα από το Μηνιαίο Dataset	100
Πίνακας 6.4.6-1: Residual Checks στα Μοντέλα ARIMA / SARIMA Τροχαίων Ατυχημάτων	115
Πίνακας 6.4.6-2: Ενδεικτικές Τιμές Προβλέψεων Εβδομαδιαίου Μοντέλου ARIMA(4,1,1) with Drift	115
Πίνακας 6.4.6-3: Ενδεικτικές Τιμές Προβλέψεων Εβδομαδιαίου Μοντέλου SARIMA(2,1,1)(0,1,1)[48]	116
Πίνακας 6.4.6-4: Ενδεικτικές Τιμές Προβλέψεων Μηνιαίου Μοντέλου ARIMA(3,1,2)	117
Πίνακας 6.4.6-5: Ενδεικτικές Τιμές Προβλέψεων Μηνιαίου Μοντέλου SARIMA(2,1,1)(0,1,1)[12]	118
Πίνακας 6.4.6-6: Ενδεικτικές Τιμές Προβλέψεων Μηνιαίου Μοντέλου SARIMA(1,0,2)(1,0,3)[12] (GS)	118
Πίνακας 6.4.12-1: Residual Checks στα Μοντέλα ARIMA / SARIMA Θανάτων από Τροχαία Ατυχημάτων	133
Πίνακας 6.4.12-2: Ενδεικτικές Τιμές Προβλέψεων Εβδομαδιαίου Μοντέλου ARIMA(4,1,1) with Drift	133
Πίνακας 6.4.12-3: Ενδεικτικές Τιμές Προβλέψεων Εβδομαδιαίου Μοντέλου SARIMA(1,1,2)(1,1,0)[48]	134
Πίνακας 6.4.12-4: Ενδεικτικές Τιμές Προβλέψεων Μηνιαίου Μοντέλου ARIMA(3,1,2) with Drift	135

Πίνακας 6.4.12-5: Ενδεικτικές Τιμές Προβλέψεων Μηνιαίου Μοντέλου SARIMA(2,1,1)(1,1,1)[12]	136
Πίνακας 6.4.12-6: Ενδεικτικές Τιμές Προβλέψεων Εβδομαδιαίου Μοντέλου SARIMA(3,0,3)(0,1,3)[48]	137
Πίνακας 6.4.12-7: Ενδεικτικές Τιμές Προβλέψεων Μηνιαίου Μοντέλου SARIMA(1,0,2)(1,1,1)[12]	138
Πίνακας 7.2.1-1: Καλύτερες Hyperparameters από το Grid Search για το LSTM με το Εβδομαδιαίο Dataset Τροχαίων Ατυχημάτων	148
Πίνακας 7.2.1-2: Μετρικές LSTM NN για το Εβδομαδιαίο Dataset Τροχαίων Ατυχημάτων	148
Πίνακας 7.2.1-3: Πρόβλεψη Τροχαίων Ατυχημάτων με το LSTM NN για την 1 ^η εβδομάδα του 2023	148
Πίνακας 7.2.2-1: Καλύτερες Hyperparameters από το Grid Search για το LSTM με το Εβδομαδιαίο Dataset Θανάτων από Τροχαία Ατυχήματα	152
Πίνακας 7.2.2-2: Μετρικές LSTM NN για το Εβδομαδιαίο Dataset Θανάτων από Τροχαία Ατυχήματα	152
Πίνακας 7.2.2-3: Πρόβλεψη Θανάτων από Τροχαία Ατυχήματα με το LSTM NN για την 1 ^η εβδομάδα του 2023	152
Πίνακας 7.2.3-1: Καλύτερες Hyperparameters από το Grid Search για το LSTM με το Μηνιαίο Dataset Τροχαίων Ατυχημάτων	156
Πίνακας 7.2.3-2: Μετρικές LSTM NN για το Μηνιαίο Dataset Τροχαίων Ατυχημάτων	156
Πίνακας 7.2.3-3: Πρόβλεψη Τροχαίων Ατυχημάτων με το LSTM NN για τον 1 ^ο μήνα του 2023	156
Πίνακας 7.2.4-1: Καλύτερες Hyperparameters από το Grid Search για το LSTM με το Μηνιαίο Dataset Θανάτων από Τροχαία Ατυχήματα	160
Πίνακας 7.2.4-2: Μετρικές LSTM NN για το Μηνιαίο Dataset Θανάτων από Τροχαία Ατυχήματα	160
Πίνακας 7.2.4-3: Πρόβλεψη Θανάτων από Τροχαία Ατυχήματα με το LSTM NN για τον 1 ^ο μήνα του 2023	160
Πίνακας 7.3.1-1: Ενδεικτικές Recursive Predictions (2023+) για το Dataset των Εβδομαδιαίων Τροχαίων Ατυχημάτων	165
Πίνακας 7.3.2-1: Ενδεικτικές Recursive Predictions (2023+) για το Dataset των Εβδομαδιαίων Θανάτων από Τροχαία Ατυχήματα	166
Πίνακας 7.3.3-1: Ενδεικτικές Recursive Predictions (2023+) για το Dataset των Μηνιαίων Τροχαίων Ατυχημάτων	167

Πίνακας 7.3.4-1: Ενδεικτικές Recursive Predictions (2023+) για το Dataset των Μηνιαίων Θανάτων από Τροχαία Ατυχήματα.....	168
--	-----

Κατάλογος Σχημάτων

Σχήμα 2.4.1-1: Basic Road Safety Figures, Greece 2014 – 2024 (Road Safety Observatory, 2025) (Slight increase in road fatalities in 2024, Greece, 2025)	5
Σχήμα 2.4.1-1: Road Traffic Fatality Rate per 100 000 Population by WHO Region and Country Income Level, 2021 (World Health Organization, 2023)	7
Σχήμα 2.4.1-1: Vision Zero Διάγραμμα (Folla, 2023)	8
Σχήμα 2.5.1-1: Αρχιτεκτονικές των FeedForward Neural Network (FNN) (a) και Recurrent Neural Network (RNN) (b)	16
Σχήμα 2.5.1-2: Ολική Συνάρτηση του RNN	17
Σχήμα 2.5.1-3: Backpropagation Through Time (BPTT) in RNNs	17
Σχήμα 2.5.1-4: Αρχιτεκτονική ενός LSTM Cell	20
Σχήμα 2.5.1-5: Αρχιτεκτονική ενός LSTM Layer	21
Σχήμα 3.1.1-1: Δεδομένα Ατυχημάτων Ιανουάριος – Μάρτιος από τα Έτη 1996 – 2003 (Μορφή Count Panel Data)	27
Σχήμα 3.1.1-2: Δεδομένα Ατυχημάτων Εβδομάδων 1-26 και Μηνών 1-12 του Έτους 2015 (Μορφή Pivot Table)	28
Σχήμα 3.1.1-3: Φεβ. 1996, 1 ^η «εβδομάδα» είναι η 01/02 – 03/02	29
Σχήμα 3.2.1-1: Δεδομένα Τροχαίων Ατυχημάτων και Θανάτων πριν την Μεταμόρφωση	35
Σχήμα 3.2.1-2: Δεδομένα Τροχαίων Ατυχημάτων και Θανάτων μετά την Μεταμόρφωση	35
Σχήμα 3.4.2-1: ER Schema της BD RoadAccidents.	37
Σχήμα 3.4.5-1: Περιβάλλον pgAdmin4 και Ενδεικτικό Query για Εξαγωγή των Datasets	43
Σχήμα 5.2.1-1: Turkey HSD Αποτελέσματα για Σύγκριση Εβδομαδιαίων Τροχαίων Ατυχημάτων σε Σχέση με την Εποχή	57
Σχήμα 5.2.2-1: Turkey HSD Αποτελέσματα για Σύγκριση Τροχαίων Μηνιαίων Ατυχημάτων σε Σχέση με την Εποχή	59
Σχήμα 5.3.1-1: Χρονοσειρά Τροχαίων Ατυχημάτων στην Ελλάδα 1996 – 2022	61
Σχήμα 5.3.1-2: Χρονοσειρά Θανάτων από Τροχαία Ατυχήματα στην Ελλάδα 1996 – 2022	62
Σχήμα 5.3.1-3: Rolling Sums της Χρονοσειράς Εβδομαδιαίων Τροχαίων Ατυχημάτων	63
Σχήμα 5.3.1-4: Rolling Sums της Χρονοσειράς Εβδομαδιαίων Θανάτων από Τροχαία Ατυχήματα	64
Σχήμα 5.3.1-5: Rolling Averages της Χρονοσειράς Εβδομαδιαίων Τροχαίων Ατυχημάτων	64

Σχήμα 5.3.1-6: Rolling Averages της Χρονοσειράς Εβδομαδιαίων Θανάτων από Τροχαία Ατυχήματα	65
Σχήμα 5.3.1-7: Time Series Plot για Θανάτους από Τροχαία Ατυχήματα ανά κάτοικο 1996 – 2022	66
Σχήμα 5.3.1-8: Time Series Plot για Τροχαία Ατυχήματα ανά Υπάρχοντα Οχήματα 1996 – 2022 ...	67
Σχήμα 5.3.1-9: Time Series Plot για Ποσοστό Θνησιμότητας των Τροχαίων Ατυχημάτων 1996 - 2022	67
Σχήμα 5.3.2-1: Boxplot Κατανομής Τιμών Τροχαίων Ατυχημάτων ανά Μήνα	69
Σχήμα 5.3.2-2: Boxplot Κατανομής Τιμών Θανάτων από Τροχαία Ατυχήματα ανά Μήνα	69
Σχήμα 5.3.2-3: Histogram Κατανομής Τιμών Τροχαίων Ατυχημάτων	71
Σχήμα 5.3.2-4: Histogram Κατανομής Τιμών Θανάτων από Τροχαία Ατυχήματα	71
Σχήμα 5.3.2-5: Scatter Plot Τροχαίων Ατυχημάτων και Θανάτων	72
Σχήμα 5.3.3-1: Μέσος Όρος Τροχαίων Ατυχημάτων Ανα Εβδομάδα για Κάθε Εποχή	73
Σχήμα 5.3.3-2: Μέσος Όρος Θανάτων από Τροχαία Ατυχήματα Ανα Εβδομάδα για Κάθε Εποχή ..	74
Σχήμα 5.3.3-3: Τροχαία Ατυχήματα Όταν Υπάρχουν Αργίες / Διακοπές	75
Σχήμα 5.3.3-4: Θάνατοι από Τροχαία Ατυχήματα Όταν Υπάρχουν Αργίες / Διακοπές	75
Σχήμα 5.3.4-1: Pearson Correlation Matrix για τα Δεδομένα Τροχαίων Ατυχημάτων	77
Σχήμα 5.3.4-2: Spearman Correlation Matrix για τα Δεδομένα Τροχαίων Ατυχημάτων	78
Σχήμα 5.3.4-3: Kendall Correlation Matrix για τα Δεδομένα Τροχαίων Ατυχημάτων	79
Σχήμα 5.3.4-4: Covariance Matrix για τα Δεδομένα Τροχαίων Ατυχημάτων	80
Σχήμα 5.3.4-1: Workflow Στατιστικής Μοντελοποίησης	82
Σχήμα 5.3.4-1: Μεθοδολογία του Model Selection and Testing	84
Σχήμα 6.2.2-1: Ενδεικτικό ACF Plot της Χρονοσειράς των Regular Εβδομαδιαίων Τροχαίων Ατυχημάτων	87
Σχήμα 6.2.2-2: Ενδεικτικό PACF Plot της Χρονοσειράς των Regular Εβδομαδιαίων Τροχαίων Ατυχημάτων	87
Σχήμα 6.2.3-1: Ενδεικτικό Block Κώδικα για Manual Fitting και Fitting με Auto.Arima()	90
Σχήμα 6.2.5-1: Ενδεικτικό Block Κώδικα που πραγματοποιεί Grid Search για Εύρεση Μοντέλου που να Περνάει το Residual Check	92
Σχήμα 6.2.6-1: Αποτελέσματα Grid Search για Fitting Μοντέλων ARIMA / SARIMA με Πρόβλεψη τους Θανάτους από Τροχαία Ατυχήματα (Regular Fatalities Time Series)	98
Σχήμα 6.2.6-2: Αποτελέσματα Grid Search για Fitting Μοντέλων ARIMA / SARIMA με Πρόβλεψη τους Θανάτους από Τροχαία Ατυχήματα (Log Fatalities Time Series)	98

Σχήμα 6.4.1-1: Actual vs Predicted Τιμες Εβδομαδιαίων Τροχαίων Ατυχημάτων του Μοντέλου ARIMA(4,1,1) with Drift.....	102
Σχήμα 6.4.1-2: Προβλέψεις του Μοντέλου ARIMA(4,1,1) για Εβδομαδιαία Τροχαία Ατυχήματα έως το 2026.....	103
Σχήμα 6.4.1-3: Διαγνωστικά των Residuals του Μοντέλου ARIMA(4,1,1) σε Εβδομαδιαία Τροχαία Ατυχήματα	103
Σχήμα 6.4.2-1: Actual vs Predicted Τιμες Εβδομαδιαίων Τροχαίων Ατυχημάτων του Μοντέλου SARIMA(2,1,1)(0,1,1)[48]	105
Σχήμα 6.4.2-2: Προβλέψεις του Μοντέλου SARIMA(2,1,1)(0,1,1)[48] για Εβδομαδιαία Τροχαία Ατυχήματα έως το 2026.....	106
Σχήμα 6.4.2-3: Διαγνωστικά των Residuals του Μοντέλου SARIMA(2,1,1)(0,1,1)[48] σε Εβδομαδιαία Τροχαία Ατυχήματα	106
Σχήμα 6.4.3-1: Actual vs Predicted Τιμες Μηνιαίων Τροχαίων Ατυχημάτων του Μοντέλου ARIMA(3,1,2).....	108
Σχήμα 6.4.3-2: Προβλέψεις του Μοντέλου ARIMA(3,1,2) για Μηνιαία Τροχαία Ατυχήματα έως το 2026.....	108
Σχήμα 6.4.3-3: Διαγνωστικά των Residuals του Μοντέλου ARIMA(3,1,2) σε Μηνιαία Τροχαία Ατυχήματα	109
Σχήμα 6.4.4-1: Actual vs Predicted Τιμες Μηνιαίων Τροχαίων Ατυχημάτων του Μοντέλου SARIMA(2,1,1)(0,1,1)[12]	110
Σχήμα 6.4.4-2: Προβλέψεις του Μοντέλου SARIMA(2,1,1)(0,1,1)[12] για Μηνιαία Τροχαία Ατυχήματα έως το 2026.....	111
Σχήμα 6.4.4-3: Διαγνωστικά των Residuals του Μοντέλου SARIMA(2,1,1)(0,1,1)[12] σε Μηνιαία Τροχαία Ατυχήματα	112
Σχήμα 6.4.5-1: Actual vs Predicted Τιμες Μηνιαίων Τροχαίων Ατυχημάτων του Μοντέλου SARIMA(1,0,2)(1,0,3)[12]	113
Σχήμα 6.4.5-2: Προβλέψεις του Μοντέλου SARIMA(1,0,2)(1,0,3)[12] για Μηνιαία Τροχαία Ατυχήματα έως το 2026.....	114
Σχήμα 6.4.5-3: Διαγνωστικά των Residuals του Μοντέλου SARIMA(1,0,2)(1,0,3)[12] σε Μηνιαία Τροχαία Ατυχήματα	114
Σχήμα 6.4.7-1: Actual vs Predicted Τιμες Εβδομαδιαίων Θανάτων από Τροχαία Ατυχήματα του Μοντέλου ARIMA(4,1,1) with Drift.....	120

Σχήμα 6.4.7-2 Προβλέψεις του Μοντέλου ARIMA(4,1,1) with Drift για Εβδομαδιαίους Θανάτους από Τροχαία Ατυχήματα έως το 2026	120
Σχήμα 6.4.7-3: Διαγνωστικά των Residuals του Μοντέλου ARIMA(4,1,1) with Drift σε Εβδομαδιαίους Θανάτους από Τροχαία Ατυχήματα	121
Σχήμα 6.4.8-1: Actual vs Predicted Τιμες Εβδομαδιαίων Θανάτων από Τροχαία Ατυχήματα του Μοντέλου SARIMA(1,1,2)(1,1,0)[48]	122
Σχήμα 6.4.8-2: Προβλέψεις του Μοντέλου SARIMA(1,1,2)(1,1,0)[48] για Εβδομαδιαίους Θανάτους από Τροχαία Ατυχήματα έως το 2026	123
Σχήμα 6.4.8-3: Διαγνωστικά των Residuals του Μοντέλου SARIMA(1,1,2)(1,1,0)[48] σε Εβδομαδιαίους Θανάτους από Τροχαία Ατυχήματα	123
Σχήμα 6.4.9-1: Actual vs Predicted Τιμες Μηνιαίων Θανάτων από Τροχαία Ατυχήματα του Μοντέλου ARIMA(3,1,2) with Drift	125
Σχήμα 6.4.9-2: Προβλέψεις του Μοντέλου ARIMA(3,1,2) with Drift για Μηνιαίους Θανάτους από Τροχαία Ατυχήματα έως το 2026	125
Σχήμα 6.4.9-3: Διαγνωστικά των Residuals του Μοντέλου ARIMA(3,1,2) with Drift σε Μηνιαίους Θανάτους από Τροχαία Ατυχήματα	126
Σχήμα 6.4.10-1: Actual vs Predicted Τιμες Μηνιαίων Θανάτων από Τροχαία Ατυχήματα του Μοντέλου SARIMA(2,1,1)(1,1,1)[12]	127
Σχήμα 6.4.10-2: Προβλέψεις του Μοντέλου SARIMA(2,1,1)(1,1,1)[12] για Μηνιαίους Θανάτους από Τροχαία Ατυχήματα έως το 2026	127
Σχήμα 6.4.10-3: Διαγνωστικά των Residuals του Μοντέλου SARIMA(2,1,1)(1,1,1)[12] σε Μηνιαίους Θανάτους από Τροχαία Ατυχήματα	128
Σχήμα 6.4.11-1: Actual vs Predicted Τιμες Εβδομαδιαίων Θανάτων από Τροχαία Ατυχήματα του Μοντέλου SARIMA(3,0,3)(0,1,3)[48]	129
Σχήμα 6.4.11-2: Προβλέψεις του Μοντέλου SARIMA(3,0,3)(0,1,3)[48] για Εβδομαδιαίους Θανάτους από Τροχαία Ατυχήματα έως το 2026	129
Σχήμα 6.4.11-3: Διαγνωστικά των Residuals του Μοντέλου SARIMA(3,0,3)(0,1,3)[48] σε Εβδομαδιαίους Θανάτους από Τροχαία Ατυχήματα	130
Σχήμα 6.4.11-4: Actual vs Predicted Τιμες Μηνιαίων Θανάτων από Τροχαία Ατυχήματα του Μοντέλου SARIMA(1,0,2)(1,1,1)[12]	131
Σχήμα 6.4.11-5: Προβλέψεις του Μοντέλου SARIMA(1,0,2)(1,1,1)[12] για Μηνιαίους Θανάτους από Τροχαία Ατυχήματα έως το 2026	131

Σχήμα 6.4.11-6: Διαγνωστικά των Residuals του Μοντέλου SARIMA(1,0,2)(1,1,1)[12] σε Μηνιαίους Θανάτους από Τροχαία Ατυχήματα	132
Σχήμα 7.1.1-1: Hyperparameter Grid για το Grid Search Βέλτιστου Μοντέλου	143
Σχήμα 7.1.2-1: Αρχιτεκτονική NN LSTM.....	144
Σχήμα 7.1.3-1: Recursive Predictions Flowchart	146
Σχήμα 7.1.4-1: Διαφορά Μεταξύ One-Step vs Multi-Step Forecasting.....	147
Σχήμα 7.2.1-1: Διάγραμμα Actual vs Predicted του LSTM NN για το Εβδομαδιαίο Dataset Τροχαίων Ατυχημάτων	149
Σχήμα 7.2.1-2: Ιστόγραμμα των Residuals του LSTM NN για το Εβδομαδιαίο Dataset Τροχαίων Ατυχημάτων	149
Σχήμα 7.2.1-3: Scatter Plot των Predicted vs Actual του LSTM NN για το Εβδομαδιαίο Dataset Τροχαίων Ατυχημάτων.....	150
Σχήμα 7.2.2-1: Διάγραμμα Actual vs Predicted του LSTM NN για το Εβδομαδιαίο Dataset Θανάτων από Τροχαία Ατυχήματα.....	153
Σχήμα 7.2.2-2: Ιστόγραμμα των Residuals του LSTM NN για το Εβδομαδιαίο Dataset Θανάτων από Τροχαία Ατυχήματα	153
Σχήμα 7.2.2-3: Scatter Plot των Predicted vs Actual του LSTM NN για το Εβδομαδιαίο Dataset Θανάτων από Τροχαία Ατυχήματα.....	154
Σχήμα 7.2.3-1: Διάγραμμα Actual vs Predicted του LSTM NN για το Μηνιαίο Dataset Τροχαίων Ατυχημάτων	157
Σχήμα 7.2.3-2: Ιστόγραμμα των Residuals του LSTM NN για το Μηνιαίο Dataset Τροχαίων Ατυχημάτων	157
Σχήμα 7.2.3-3: Scatter Plot των Predicted vs Actual του LSTM NN για το Μηνιαίο Dataset Τροχαίων Ατυχημάτων	158
Σχήμα 7.2.4-1: Διάγραμμα Actual vs Predicted του LSTM NN για το Μηνιαίο Dataset Θανάτων από Τροχαία Ατυχήματα	161
Σχήμα 7.2.4-2: Ιστόγραμμα των Residuals του LSTM NN για το Μηνιαίο Dataset Θανάτων από Τροχαία Ατυχήματα	161
Σχήμα 7.2.4-3: Scatter Plot των Predicted vs Actual του LSTM NN για το Μηνιαίο Dataset Θανάτων από Τροχαία Ατυχήματα.....	162
Σχήμα 7.3.1-1: Recursive Predictions Εβδομαδιαίων Τροχαίων Ατυχημάτων έως το 2026	165

Σχήμα 7.3.2-1: Recursive Predictions Εβδομαδιαίων Θανάτων από Τροχαία Ατυχήματα έως το 2026	166
Σχήμα 7.3.3-1: Recursive Predictions Μηνιαίων Τροχαίων Ατυχημάτων έως το 2026	167
Σχήμα 7.3.4-1: Recursive Predictions Μηνιαίων Θανάτων από Τροχαία Ατυχήματα έως το 2026	168

Κεφάλαιο 1 Εισαγωγή

Η οδική ασφάλεια έχει αποτελέσει σταθερά ένα από τα βασικά πεδία προβληματισμού σε πολιτικό, κοινωνικό και επιστημονικό επίπεδο, ιδιαίτερα από τα μέσα του 20ού αιώνα και έπειτα. Η ραγδαία αύξηση της κυκλοφορίας οχημάτων, σε συνδυασμό με την επιταχυνόμενη αστικοποίηση και την ανάπτυξη των οδικών δικτύων, οδήγησε σε πρωτοφανή άνοδο των τροχαίων ατυχημάτων. Οι συνέπειες υπήρξαν πολυδιάστατες: απώλεια ανθρώπινων ζωών, σοβαροί τραυματισμοί, επιβάρυνση των συστημάτων υγείας, αλλά και σημαντικές οικονομικές απώλειες λόγω της μειωμένης παραγωγικότητας και του κόστους αποκατάστασης (Dalal, Lin, Gifford, & Svanström, 2013). Από τις δεκαετίες του 1960 και 1970, η ανάγκη κατανόησης των αιτίων των ατυχημάτων, οδήγησε στην καθιέρωση οργανωμένων συστημάτων συλλογής δεδομένων και στην ανάπτυξη πολιτικών πρόληψης σε διεθνές επίπεδο (Peden, et al., 2004). Ο στόχος ήταν όχι μόνο η αποτύπωση της υφιστάμενης κατάστασης, αλλά και η αναγνώριση προτύπων και παραγόντων κινδύνου, που θα επέτρεπαν τον σχεδιασμό πιο στοχευμένων παρεμβάσεων.

Στην ελληνική πραγματικότητα, η διάδοση της ιδιοκτησίας οχημάτων και η εντατικοποίηση της κυκλοφορίας με αυτά, γνώρισαν εξαιρετική αύξηση από τις αρχές της δεκαετίας του 1980 και μετά. Ωστόσο, αυτή η εξέλιξη δεν συνοδεύτηκε από ανάλογη πρόοδο στις υποδομές, στη ρύθμιση της κυκλοφορίας ή στην οδηγική εκπαίδευση. Η απουσία οργανωμένων προγραμμάτων πρόληψης, η περιορισμένη επιτήρηση των κανόνων οδικής κυκλοφορίας και οι καθυστερήσεις στην υιοθέτηση τεχνολογιών ασφάλειας συνέβαλαν στην αύξηση της συχνότητας των ατυχημάτων, με ιδιαίτερα αρνητικές συνέπειες για τις νεαρές ηλικίες. Αν και τα τελευταία χρόνια καταγράφεται αισθητή πρόοδος, μέσω έργων αναβάθμισης υποδομών, αυστηροποίησης της νομοθεσίας και χρήσης μέσων τεχνολογικής παρακολούθησης, το ζήτημα των τροχαίων ατυχημάτων εξακολουθεί να αποτελεί μείζονα πρόκληση. Η ανάγκη για διαρκή βελτίωση και προληπτικές παρεμβάσεις παραμένει επίκαιρη, απαιτώντας πλέον πιο σύνθετες και τεκμηριωμένες προσεγγίσεις (ΣΕΣ Επιτροπή Οδικών Μεταφορών, 2025).

1.1 Διατύπωση του Προβλήματος

Παρά την πρόοδο που έχει σημειωθεί τις τελευταίες δεκαετίες στον τομέα της οδικής ασφάλειας, τα τροχαία ατυχήματα εξακολουθούν να αποτελούν μία από τις βασικές αιτίες πρόωρου θανάτου και σοβαρού τραυματισμού, τόσο στην Ελλάδα όσο και διεθνώς. Χιλιάδες ζωές χάνονται ή αλλάζουν δραματικά κάθε χρόνο εξαιτίας ατυχημάτων που, σε μεγάλο βαθμό, μπορούν να προληφθούν (Peden, et al., 2004). Οι συνέπειες δεν είναι μόνο προσωπικές και οικογενειακές, αλλά και κοινωνικές και οικονομικές. Το πρόβλημα δεν είναι στατικό και γι αυτό, μεταβάλλεται με τον χρόνο, επηρεάζεται από εποχικούς, κοινωνικούς και θεσμικούς παράγοντες και απαιτεί συνεχή επαναξιολόγηση. Το γεγονός ότι το πρόβλημα επιμένει, παρά την εφαρμογή διαφόρων μέτρων και τεχνολογικών λύσεων, καταδεικνύει ότι η αντιμετώπισή του δεν έχει ακόμη επιτευχθεί σε ικανοποιητικό βαθμό.

Η ανάγκη για μείωση των ατυχημάτων δεν αφορά μόνο την πρόληψη μέσω καμπανιών ευαισθητοποίησης ή αυστηρότερης νομοθεσίας, αλλά και την κατανόηση του πότε, πού και γιατί συμβαίνουν. Για να σχεδιαστούν πιο στοχευμένες και αποδοτικές παρεμβάσεις, είναι κρίσιμο να υπάρχει πρόβλεψη της πορείας του φαινομένου στο χρόνο. Η κατανόηση των τάσεων και των επαναλαμβανόμενων μοτίβων μπορεί να προσφέρει ουσιαστική υποστήριξη στη λήψη αποφάσεων και την έγκαιρη εφαρμογή πολιτικών. Σε αυτό το πλαίσιο, η επιστημονικά τεκμηριωμένη πρόβλεψη των ατυχημάτων αποτελεί αναγκαίο εργαλείο για την ενίσχυση της οδικής ασφάλειας και τη μείωση των θυμάτων στο μέλλον.

1.2 Στόχοι της Διπλωματικής Εργασίας

Βασικός στόχος της παρούσας διπλωματικής εργασίας είναι η ανάπτυξη, εφαρμογή και αξιολόγηση μοντέλων ανάλυσης χρονοσειρών για την πρόβλεψη του αριθμού των οδικών ατυχημάτων και θανάτων στην Ελλάδα. Η εργασία επιδιώκει να αποτυπώσει τη δυναμική εξέλιξη των ατυχημάτων σε βάθος χρόνου, να ανιχνεύσει εποχικά και διαρθρωτικά μοτίβα και να εκτιμήσει μελλοντικές τάσεις.

Επιμέρους στόχοι της διπλωματικής εργασίας περιλαμβάνουν τη διερεύνηση της καταλληλότητας διαφορετικών μεθόδων χρονοσειριακής ανάλυσης, όπως τα παραδοσιακά στατιστικά μοντέλα και τα σύγχρονα μοντέλα μηχανικής μάθησης, τη μελέτη της επίδρασης

εξωτερικών παραγόντων στα πρότυπα των ατυχημάτων και την αξιολόγηση της ακρίβειας και της πρακτικής χρησιμότητας των προβλέψεων. Ιδιαίτερη έμφαση δίνεται επίσης στη σύγκριση των μοντέλων ως προς την ικανότητά τους να προβλέψουν όχι μόνο τον αριθμό ατυχημάτων αλλά και τη χρονική κατανομή τους, αναδεικνύοντας τις περιόδους αυξημένης επικινδυνότητας.

1.3 Πεδίο Εφαρμογής και Περιορισμοί

Η παρούσα εργασία επικεντρώνεται στην ανάλυση και πρόβλεψη των οδικών ατυχημάτων και θανάτων στην Ελλάδα, αξιοποιώντας εβδομαδιαία δεδομένα για την περίοδο 1996–2022. Η εργασία επικεντρώνεται τόσο στη βραχυπρόθεσμη όσο και στη μεσοπρόθεσμη εκτίμηση της πορείας των ατυχημάτων, με στόχο τη συμβολή στον σχεδιασμό πιο στοχευμένων και αποτελεσματικών πολιτικών οδικής ασφάλειας. Για την υλοποίηση αυτής της προσπάθειας αξιοποιούνται σύγχρονες μέθοδοι χρονοσειριακής ανάλυσης και πρόβλεψης.

Παρά τα πλεονεκτήματα που προσφέρει η χρονοσειριακή ανάλυση, η μελέτη υπόκειται σε ορισμένους περιορισμούς που πρέπει να ληφθούν υπόψη. Ο αριθμός ατυχημάτων και θανάτων αποτελεί το βασικό αντικείμενο ανάλυσης, χωρίς διάκριση ως προς τη σοβαρότητα, τον τύπο ή τη γεωγραφική κατανομή των συμβάντων, περιορίζοντας έτσι τη δυνατότητα εντοπισμού ειδικότερων προτύπων. Επιπλέον, η πρόβλεψη στηρίζεται αποκλειστικά σε ιστορικά δεδομένα και σε εξωτερικές πληροφορίες περιορισμένης έκτασης, γεγονός που ενδέχεται να μην αποτυπώνει επαρκώς απρόβλεπτες κοινωνικές, οικονομικές ή θεσμικές μεταβολές που επηρεάζουν δυναμικά το φαινόμενο. Τέλος, η προσέγγιση βασίζεται στην υπόθεση ότι οι διαδικασίες που διέπουν τη δημιουργία των δεδομένων παραμένουν σχετικά σταθερές στο χρόνο, παραδοχή που σε περιόδους έντονων μεταβολών μπορεί να οδηγήσει σε μειωμένη προγνωστική ακρίβεια.

Κεφάλαιο 2 Βιβλιογραφική Ανασκόπηση

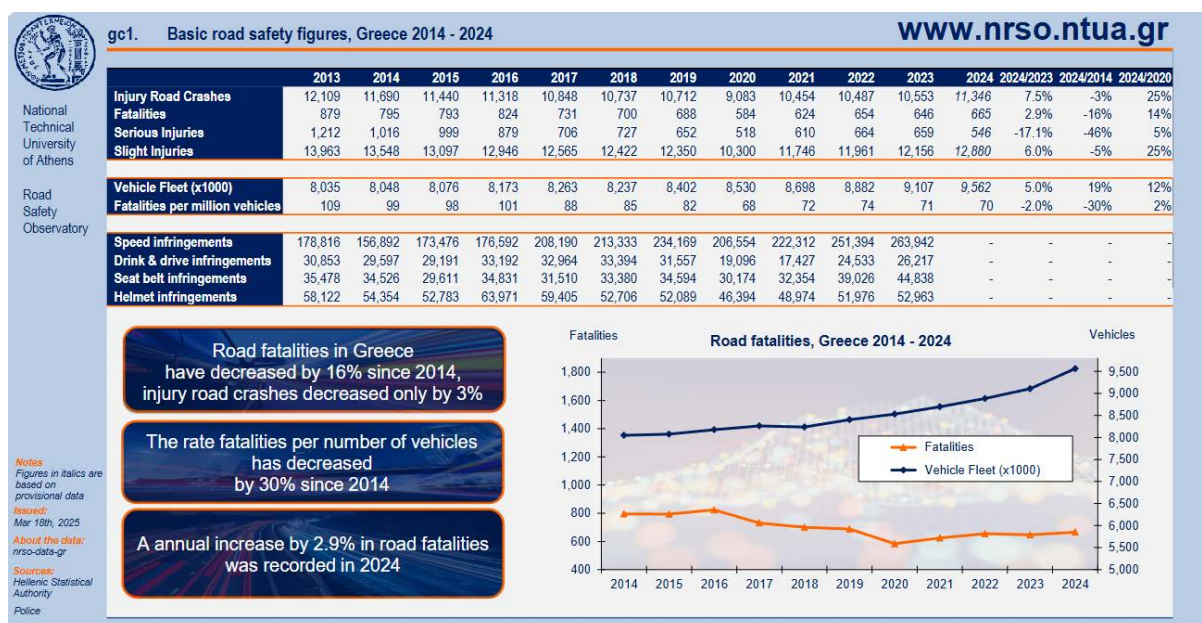
Η ανασκόπηση της υπάρχουσας βιβλιογραφίας αποτελεί κρίσιμο βήμα για την επιστημονική τεκμηρίωση κάθε ερευνητικής προσπάθειας. Στο πλαίσιο της παρούσας εργασίας, η βιβλιογραφική ανασκόπηση επικεντρώνεται αφενός στην παρουσίαση των βασικών χαρακτηριστικών και τάσεων των τροχαίων ατυχημάτων σε παγκόσμιο και εθνικό επίπεδο, και αφετέρου στις μεθοδολογίες που έχουν χρησιμοποιηθεί διεθνώς για την πρόβλεψη και ανάλυσή τους.

2.1 Τάσεις Τροχαίων Ατυχημάτων: Παγκόσμια και Ελληνική Πραγματικότητα

Σε παγκόσμιο επίπεδο, όπως και προαναφέρθηκε, τα τροχαία ατυχήματα συνεχίζουν να αποτελούν μία από τις σημαντικότερες προκλήσεις δημόσιας υγείας και βιώσιμης ανάπτυξης. Σύμφωνα με τα τελευταία στοιχεία του Παγκόσμιου Οργανισμού Υγείας (World Health Organization, 2023), το 2021 καταγράφηκαν περίπου 1,19 εκατομμύρια θάνατοι από τροχαία ατυχήματα, που αντιστοιχούν σε 15 θανάτους ανά 100.000 κατοίκους. Ιδιαίτερα ανησυχητικό είναι το γεγονός ότι οι σοβαροί τραυματισμοί από τροχαία παραμένουν η 1η αιτία θανάτου παγκοσμίως για τις ηλικίες 5–29 ετών, ενώ καταλαμβάνουν τη 12η θέση στις συνολικές αιτίες θανάτου όλων των ηλικιών. Οι κοινωνικές και οικονομικές επιπτώσεις είναι εξίσου σημαντικές, καθώς τα τροχαία ατυχήματα επιφέρουν απώλειες που υπολογίζονται στο 1% έως και 3% του ΑΕΠ σε πολλές χώρες, ενώ σε ορισμένες περιπτώσεις αγγίζουν ακόμη και το 6%. Επιπλέον, επηρεάζουν κυρίως τον πιο παραγωγικό πληθυσμό, δηλαδή, το 69% των θανάτων αφορά άτομα ηλικίας 18–59 ετών, ενώ οι άνδρες πλήττονται δυσανάλογα, με αναλογία θνησιμότητας γυναικών προς άνδρες 1:3.

Στην ελληνική πραγματικότητα, αν και η πορεία της οδικής ασφάλειας έχει παρουσιάσει σημαντική πρόοδο από το 2014, με μείωση των θανάτων από τροχαία κατά 16%

και μείωση του δείκτη θανάτων ανά εκατομμύριο οχημάτων κατά 30%, τα πρόσφατα στοιχεία προκαλούν ανησυχία. Το 2024 καταγράφηκε αύξηση 2,9% στους θανάτους από τροχαία σε σύγκριση με το 2023, σύμφωνα με τα προσωρινά στοιχεία της ΕΛΣΤΑΤ, ενώ η Ελλάδα κατατάχθηκε 25η από τις 27 χώρες της Ε.Ε. όσον αφορά την οδική ασφάλεια, καλύτερη μόνο από τη Βουλγαρία και τη Ρουμανία (Road Safety Observatory, 2025) (Slight increase in road fatalities in 2024, Greece, 2025). Τα αποτελέσματα αυτά αποδεικνύουν ότι η χώρα απέχει σημαντικά από την επίτευξη των στόχων που έχουν τεθεί για το 2025 (-25%) και το 2030 (-50%), αναδεικνύοντας την επείγουσα ανάγκη για εντατικοποίηση των προσπάθειών, τόσο σε επίπεδο θεσμικών παρεμβάσεων όσο και κοινωνικής ευαισθητοποίησης. Οι αριθμοί καταδεικνύουν ότι η πρόοδος δεν είναι γραμμική και ότι απαιτείται συστηματική και διατομεακή προσέγγιση για τη βελτίωση της οδικής ασφάλειας στην Ελλάδα.



Σχήμα 2.4.1-1: Basic Road Safety Figures, Greece 2014 – 2024 (Road Safety Observatory, 2025) (Slight increase in road fatalities in 2024, Greece, 2025)

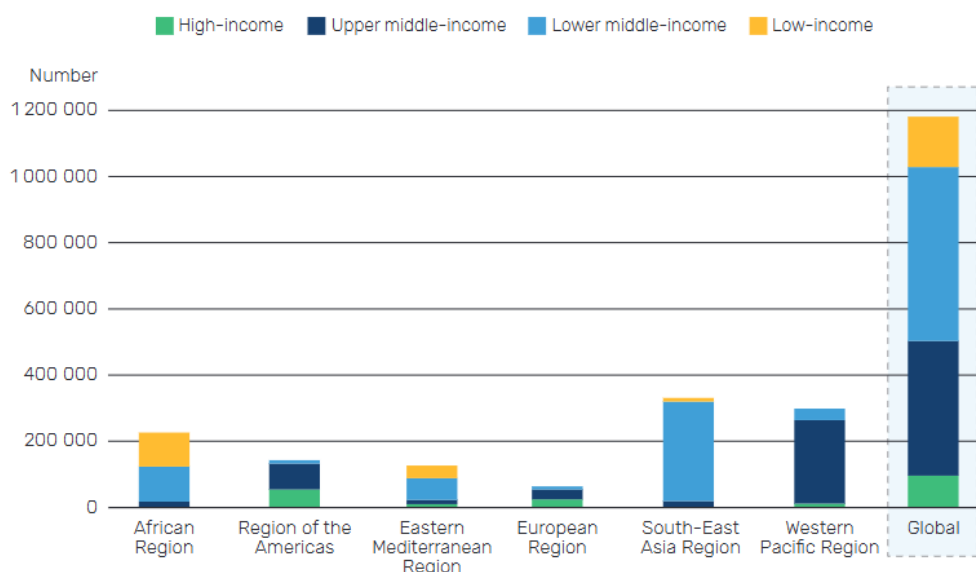
2.2 Προσδιοριστικοί Παράγοντες Τροχαίων Ατυχημάτων

Η πρόκληση τροχαίων ατυχημάτων αποτελεί αποτέλεσμα πολύπλοκης αλληλεπίδρασης μεταξύ παραγόντων που σχετίζονται με τον άνθρωπο, το όχημα, το περιβάλλον και το κοινωνικοοικονομικό πλαίσιο. Ο ανθρώπινος παράγοντας όμως παραμένει καθοριστικός. Η

υπερβολική ταχύτητα, η κατανάλωση αλκοόλ ή άλλων ουσιών, η μη χρήση ζώνης ασφαλείας ή κράνους, καθώς και η απόσπαση της προσοχής, συγκαταλέγονται στις κύριες αιτίες θανάτων από τροχαία. Σύμφωνα με τον Παγκόσμιο Οργανισμό Υγείας, η οδήγηση υπό την επήρεια αλκοόλ αυξάνει σημαντικά τον κίνδυνο θανατηφόρου ατυχήματος, ενώ η μη χρήση ζώνης ασφαλείας παραμένει σημαντικός παράγοντας κινδύνου, ιδίως σε χώρες με χαμηλή επιβολή των σχετικών νόμων (World Health Organization, 2023).

Τα χαρακτηριστικά του οχήματος και η συντήρησή του αποτελούν κρίσιμο παράγοντα για την ασφάλεια των επιβατών και την πιθανότητα πρόκλησης ατυχήματος. Η ύπαρξη βασικών παθητικών μέσων προστασίας, όπως αερόσακοι, ζώνες ασφαλείας, ενισχυμένοι και δομικοί θώρακες, παίζει καθοριστικό ρόλο στη μείωση της σοβαρότητας των τραυματισμών σε περίπτωση σύγκρουσης. Παράλληλα, η μηχανική κατάσταση των συστημάτων του οχήματος, όπως τα φρένα και τα ελαστικά, μπορεί να επηρεάσει άμεσα την ικανότητα αποφυγής ενός ατυχήματος. Ο Παγκόσμιος Οργανισμός Υγείας επισημαίνει ότι η τακτική τεχνική επιθεώρηση των οχημάτων είναι ένα από τα βασικά μέτρα πρόληψης των ατυχημάτων, ωστόσο σε πολλές χώρες, και ιδιαίτερα σε περιοχές με παλαιό και μη ελεγχόμενο στόλο, η συμμόρφωση με τα πρότυπα συντήρησης είναι περιορισμένη (World Health Organization, 2023).

Πέραν των παραπάνω, οι περιβαλλοντικοί και κοινωνικοοικονομικοί παράγοντες επιδρούν ισχυρά τη συχνότητα και σοβαρότητα των τροχαίων. Κακές καιρικές συνθήκες, ανεπαρκής φωτισμός, ελλιπής σήμανση και επικίνδυνη χάραξη οδών αποτελούν συνήθεις εξωτερικούς παράγοντες κινδύνου. Σε επίπεδο κοινωνικοοικονομικών δεδομένων, ο WHO αναφέρει ότι οι φτωχότερες κοινωνικές ομάδες είναι δυσανάλογα πιο εκτεθειμένες σε τροχαία, τόσο ως πεζοί όσο και ως οδηγοί, ενώ αντιμετωπίζουν ταυτόχρονα εμπόδια στην πρόσβαση σε σύγχρονα και ασφαλή μέσα μετακίνησης.



Σχήμα 2.4.1-1: Road Traffic Fatality Rate per 100 000 Population by WHO Region and Country Income Level, 2021 (World Health Organization, 2023)

2.3 Τιμές Καυσίμων, Οικονομικοί Δείκτες και Οδική Ασφάλεια

Η βαθιά οικονομική ύφεση στην Ελλάδα μετά το 2009 επέφερε σημαντικές αλλαγές στην καθημερινότητα και στον τρόπο μετακίνησης, με άμεσες συνέπειες στην οδική ασφάλεια. Η μείωση του ΑΕΠ κατά περισσότερο από 25% και η άνοδος της ανεργίας σε επίπεδα άνω του 25%, σε συνδυασμό με την αύξηση της τιμής της βενζίνης κατά περίπου 50% λόγω φορολογικών μέτρων, οδήγησαν σε σαφή περιορισμό των μετακινήσεων και της χρήσης Ι.Χ., ενώ παρατηρήθηκε αύξηση της χρήσης μέσων μαζικής μεταφοράς. Η μείωση των διανυθέντων χιλιομέτρων, κυρίως στις αστικές περιοχές, επίσης συνέβαλε στη σημαντική πτώση των τροχαίων δυστυχημάτων. Οι θάνατοι από τροχαία μειώθηκαν κατά 45% από το 2009 και κατά 52% σε σύγκριση με το 2004. Ωστόσο, οι αστικές περιοχές εμφάνισαν μικρότερη μείωση (31%) σε σχέση με τις μη αστικές (40,7%), ενώ συγκεκριμένες κατηγορίες χρηστών, όπως οι πεζοί και οι ηλικίες 45–64 ετών, παρουσίασαν αισθητά μικρότερη βελτίωση στην ασφάλειά τους, επιβεβαιώνοντας ότι το αστικό περιβάλλον παραμένει υψηλού κινδύνου (Μρογας, Kopelias, Mitropoulos, & Kepaptsoglou, 2016) (Yannis, Papadimitriou, & Folla, 2014).

Εκτός των παραπάνω, η οικονομική δραστηριότητα μιας χώρας συνδέεται άμεσα με τη χρήση των μεταφορικών μέσων και την ένταση της κυκλοφορίας, γεγονός που επηρεάζει τα

επίπεδα τροχαίας επικινδυνότητας. Σε περιόδους οικονομικής ανάπτυξης, η αύξηση του εισοδήματος ανά κάτοικο, η ενίσχυση της ιδιοκτησίας Ι.Χ. και η αύξηση της κινητικότητας μπορούν να οδηγήσουν σε αύξηση του αριθμού των ατυχημάτων. Αντιθέτως, σε περιόδους ύφεσης, καταγράφεται συνήθως μείωση των μετακινήσεων, ειδικά των μη αναγκαίων, και επομένως και των ατυχημάτων. Ενδεικτικά, κατά την περίοδο 2012–2021, η Ελλάδα σημείωσε μείωση άνω του 35% στους θανάτους από τροχαία ατυχήματα, ξεπερνώντας τον μέσο όρο μείωσης της Ευρωπαϊκής Ένωσης (25%) (European Commission, 2023).

Αντιλαμβανόμενη τη σημασία των συστημικών και οικονομικών παραγόντων στην οδική ασφάλεια, η Ελλάδα έχει υιοθετήσει το Εθνικό Στρατηγικό Σχέδιο Οδικής Ασφάλειας 2021–2030, το οποίο είναι πλήρως ευθυγραμμισμένο με τις αρχές του ευρωπαϊκού πλαισίου Vision Zero και της προσέγγισης Safe System. Το σχέδιο αυτό στοχεύει σε μείωση 50% των θανάτων και σοβαρών τραυματισμών μέχρι το 2030, με απώτερο στόχο τη μηδενική θνησιμότητα έως το 2050. Οι βασικοί άξονες δράσης περιλαμβάνουν μέτρα όπως η αναβάθμιση του οδικού δικτύου (ειδικά στις επαρχιακές περιοχές), ενίσχυση της επιβολής των κανόνων κυκλοφορίας (π.χ. συστηματική χρήση ηλεκτρονικών ελέγχων), βελτίωση της οδηγικής εκπαίδευσης, και παροχή κινήτρων για ανανέωση του στόλου με ασφαλέστερα οχήματα. Ιδιαίτερη έμφαση δίνεται στην προστασία των ευάλωτων χρηστών του οδικού δικτύου, όπως οι πεζοί, οι ποδηλάτες και οι δικυκλιστές, καθώς και στην κοινωνική ευαισθητοποίηση μέσω εκστρατειών ενημέρωσης (Folla, 2023).

Strategic Plan Key Principles

- Based on the principles of Vision Zero and Safe System Approach, a **new holistic approach to road transport system's safety** in Greece for the decade 2021-2030 has been adopted with the ultimate goal of achieving the ambitious vision zero fatalities by 2050.



Σχήμα 2.4.1-1: Vision Zero Διάγραμμα (Folla, 2023)

2.4 Μοντελοποίηση Χρονοσειρών Τροχαίων Ατυχημάτων

Η μοντελοποίηση χρονοσειρών αποτελεί ένα ισχυρό εργαλείο για την ανάλυση και πρόβλεψη του φαινομένου των τροχαίων ατυχημάτων στον χρόνο. Μέσω της στατιστικής μελέτης ιστορικών δεδομένων σε διακριτά χρονικά διαστήματα, μπορούν να εντοπιστούν εποχικότητα, τάσεις, κυκλικές μεταβολές και διαταραχές στη συχνότητα των ατυχημάτων. Η κατανόηση αυτών των μοτίβων επιτρέπει την πρόβλεψη μελλοντικών επιπέδων επικινδυνότητας και την υποστήριξη της λήψης τεκμηριωμένων αποφάσεων στον σχεδιασμό πολιτικών πρόληψης. Η εφαρμογή μοντέλων όπως το ARIMA, τα εποχικά SARIMA, και άλλα Count Data Μοντέλα όπως Poisson, NB, INAR, καθώς και οι πιο πρόσφατες προσεγγίσεις με νευρωνικά δίκτυα τύπου LSTM, έχουν αξιοποιηθεί διεθνώς, για την αξιόπιστη πρόβλεψη ατυχημάτων και την αξιολόγηση μέτρων οδικής ασφάλειας.

2.4.1 ΚΛΑΣΙΚΑ ΜΟΝΤΕΛΑ: ARIMA, SARIMA

Στο πλαίσιο της μελέτης τροχαίων ατυχημάτων και θανάτων, η χρονοσειριακή μοντελοποίηση επιτρέπει τη σύνδεση του παρελθόντος με το μέλλον, προσδιορίζοντας τη δομή της εσωτερικής εξάρτησης των δεδομένων. Μεταξύ των πλέον καθιερωμένων μεθόδων μοντελοποίησης είναι τα μοντέλα ARIMA (AutoRegressive Integrated Moving Average) και η εποχική τους επέκταση, SARIMA (Seasonal ARIMA), τα οποία εισήχθησαν συστηματικά στη στατιστική μέσω της μεθοδολογίας Box–Jenkins (Getahun, 2021).

Το μοντέλο ARIMA(p,d,q) προϋποθέτει την ύπαρξη στασιμότητας στη χρονοσειρά ή την καθιστά στάσιμη μέσω διαφορισμών. Το υποσύνολο "p" (AR) αναφέρεται στη χρονική εξάρτηση από προηγούμενες τιμές της χρονοσειράς (autoregressive component), το "d" εκφράζει τον αριθμό των διαφορισμών που απαιτούνται για να επιτευχθεί στασιμότητα, ενώ το "q" (MA) αντιπροσωπεύει την εξάρτηση από προηγούμενα σφάλματα πρόβλεψης (moving average component). Η μέθοδος απαιτεί έναν κύκλο τεσσάρων φάσεων: (α) προσδιορισμός μοντέλου μέσω ACF/PACF, (β) εκτίμηση παραμέτρων, (γ) διαγνωστικό έλεγχο residuals και (δ) πρόβλεψη. Κάθε στάδιο αξιολογείται με σαφώς καθορισμένα στατιστικά κριτήρια, όπως το AIC (Akaike Information Criterion), το BIC (Bayesian Information Criterion) και το τεστ στασιμότητας Augmented Dickey-Fuller (ADF) (Deretic, Stanimirovic, Awadh, Vujanovic, & Djukic, 2022).

Σε πολλές περιπτώσεις, οι χρονοσειρές εμφανίζουν εποχικές διακυμάνσεις, δηλαδή επαναλαμβανόμενα μοτίβα σε τακτά χρονικά διαστήματα (π.χ. αυξημένα ατυχήματα τους θερινούς μήνες ή τα Σαββατοκύριακα). Σε τέτοιες περιπτώσεις, η εφαρμογή ενός απλού ARIMA μοντέλου είναι ανεπαρκής. Το μοντέλο SARIMA, που επεκτείνει το ARIMA ενσωματώνοντας εποχικές παραμέτρους, αποδίδεται με τη μορφή SARIMA(p,d,q)(P,D,Q)_s. Οι παράμετροι P, D και Q αντιστοιχούν στις εποχικές συνιστώσες του αυτοπαλίνδρομου, της ολοκλήρωσης και του κινητού μέσου όρου, ενώ το s εκφράζει την εποχική περιοδικότητα (π.χ. s=12 για μηνιαία δεδομένα με ετήσια περιοδικότητα). Η ορθή εκτίμηση αυτών των παραμέτρων βασίζεται και πάλι στις ACF και PACF συναρτήσεις, καθώς και στη διάγνωση λευκού θορύβου στα residuals (Deretic, Stanimirovic, Awadh, Vujanovic, & Djukic, 2022).

Μαθηματικά, το μοντέλο ARIMA(p,d,q) περιγράφεται ως εξής:

$$\Phi(B)(1 - B)^d y_t = \Theta(B)\varepsilon_t \quad (2.1)$$

Όπου:

y_t είναι η τιμή της χρονοσειράς στη χρονική στιγμή t,

B είναι ο τελεστής οπισθοδρόμησης (backshift operator), δηλαδή $B^k y_t = y_{t-k}$,

$\Phi(B) = 1 - \phi_1 B - \dots - \phi_p B^p$ είναι το πολυώνυμο του αυτοπαλίνδρομου μέρους,

$\Theta(B) = 1 + \theta_1 B + \dots + \theta_q B^q$ είναι το πολυώνυμο του κινητού μέσου όρου,

ε_t είναι ο λευκός θόρυβος με μηδενική μέση τιμή και σταθερή διασπορά.

Για το εποχικό μοντέλο SARIMA(p,d,q)(P,D,Q)_s, η μορφή του δίνεται από τη σχέση:

$$\Phi_P(B^s)\Phi(B)(1 - B)^d (1 - B^s)^D y_t = \Theta_Q(B^s)\Theta(B)\varepsilon_t \quad (2.2)$$

Όπου:

$\Phi_P(B^s)$ και $\Theta_Q(B^s)$ είναι τα εποχικά πολυώνυμα των συνιστωσών AR και MA αντίστοιχα,

$(1 - B^s)^D$ είναι ο εποχικός διαφοριστής.

Η εφαρμογή των μοντέλων ARIMA/SARIMA προϋποθέτει συγκεκριμένες υποθέσεις: η χρονοσειρά (ή η διαφορισμένη εκδοχή της) πρέπει να είναι στάσιμη, τα residuals πρέπει να αντιπροσωπεύουν λευκό θόρυβο, και το μοντέλο να έχει καλή προσαρμογή χωρίς υπερπροσαρμογή (overfitting). Στην πράξη, η στασιμότητα ελέγχεται με στατιστικά τεστ όπως το ADF και το KPSS τεστ. Η ύπαρξη αυτοσυσχέτισης στα residuals ελέγχεται με τα Ljung–Box και Box-Pierce τεστ. Η ποιότητα πρόβλεψης αξιολογείται με βάση δείκτες σφάλματος όπως RMSE, MAE και MAPE.

Η αποτελεσματικότητα των ARIMA και SARIMA μοντέλων έχει επιβεβαιωθεί από πολυάριθμες εμπειρικές εφαρμογές στον τομέα της οδικής ασφάλειας. Για παράδειγμα, σε μελέτη για την πρόβλεψη θανάτων από τροχαία δυστυχήματα στην Κίνα, το μοντέλο SARIMA(0,1,1)(0,1,1)[12] παρουσίασε χαμηλό σφάλμα πρόβλεψης και υψηλή προσαρμοστικότητα (Qian, et al., 2020). Παρόμοια, στο Ιράν και στη Μαλαισία, τα μοντέλα ARIMA(2,0,0) και SARIMA(1,0,2)(1,0,0)[12] αντίστοιχα χρησιμοποιήθηκαν επιτυχώς για τη μοντελοποίηση των τάσεων θανάτων και τραυματισμών από τροχαία (Parvareh, et al., 2018) (Azhari, Zaidi, Anuar, & Othman, 2022).

Συμπερασματικά, τα μοντέλα ARIMA και SARIMA, αν και θεωρούνται «κλασικά», παραμένουν βασικά εργαλεία για τη στατιστική μοντελοποίηση χρονοσειρών. Η ισχύς τους έγκειται στην ικανότητά τους να προσφέρουν ερμηνεύσιμες παραμέτρους, να αποτυπώνουν τη χρονική εξάρτηση και την εποχικότητα, και να παρέχουν αξιόπιστες βραχυπρόθεσμες προβλέψεις.

2.4.2 ΜΕΤΡΙΚΕΣ ΑΞΙΟΛΟΓΗΣΗΣ ΣΤΗΝ ΠΡΟΒΛΕΨΗ ΧΡΟΝΟΣΕΙΡΩΝ (ARIMA / SARIMA)

Η διαδικασία αξιολόγησης μοντέλων χρονοσειρών όπως τα ARIMA και SARIMA είναι κρίσιμη για την εξαγωγή έγκυρων και αξιόπιστων συμπερασμάτων. Η αξιολόγηση αυτή δεν περιορίζεται μόνο στην ακρίβεια των προβλέψεων, αλλά περιλαμβάνει επίσης τη στατιστική εγκυρότητα της καταλληλότητας του μοντέλου, τον έλεγχο της στασιμότητας των δεδομένων, την επάρκεια της δομής του μοντέλου, καθώς και τη συμπεριφορά των residuals. Για να επιτευχθεί αυτό, χρησιμοποιείται ένα σύνολο διαγνωστικών τεχνικών και μετρικών οι οποίες μπορούν να ταξινομηθούν σε πέντε βασικές κατηγορίες: εργαλεία ανάλυσης εξαρτήσεων (ACF/PACF), τεστ στασιμότητας (ADF, KPSS), κριτήρια επιλογής μοντέλων (AIC, AICc,

BIC), μετρικές πρόβλεψης (RMSE, MAE, MAPE), και στατιστικοί έλεγχοι για τα residuals (Ljung–Box, Box–Pierce).

Η διαδικασία μοντελοποίησης με τα ARIMA/SARIMA μοντέλα ξεκινά με τη διερεύνηση της εσωτερικής δομής εξάρτησης της χρονοσειράς μέσω των συναρτήσεων αυτοσυσχέτισης (Autocorrelation Function – ACF) και μερικής αυτοσυσχέτισης (Partial Autocorrelation Function – PACF). Η ACF μετρά την συνολική γραμμική συσχέτιση μιας χρονοσειράς με τις lagged τιμές της, επιτρέποντας την ανίχνευση MA (Moving Average) συμπεριφορών. Η PACF εξετάζει τη μερική εξάρτηση κάθε χρονικού σημείου από τα προηγούμενα, εξουδετερώνοντας την επίδραση των προηγούμενων καθυστερήσεων. Για παράδειγμα, αν η ACF πέφτει εκθετικά και η PACF κόβεται απότομα, αυτό μπορεί να υποδεικνύει αυτοπαλίνδρομο πρώτης τάξης (δηλαδή μοντέλο AR(1)) και άρα να καθοδηγήσει τον χρήστη στην κατάλληλη διαμόρφωση και την επιλογή των παραμέτρων p, d, q των μοντέλων ARIMA/SARIMA.

Πριν όμως από την επιλογή του κατάλληλου μοντέλου, απαιτείται έλεγχος της στασιμότητας της χρονοσειράς. Τα ARIMA μοντέλα προϋποθέτουν ότι η χρονοσειρά είναι στάσιμη ως προς μέση τιμή και διακύμανση. Για τον έλεγχο αυτό, εφαρμόζεται συνήθως το τεστ Dickey–Fuller ή το ενισχυμένο τεστ Augmented Dickey–Fuller (ADF), το οποίο βασίζεται στην εξίσωση:

$$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \sum_{i=1}^p \delta_i \Delta y_{t-i} + \varepsilon_t \quad (2.3)$$

Η απόρριψη της μηδενικής υπόθεσης με $p\text{-value} < 0.05$ υποδηλώνει στασιμότητα. Συμπληρωματικά, το KPSS test (Kwiatkowski–Phillips–Schmidt–Shin) εξετάζει την αντίθετη υπόθεση, δηλαδή ότι η χρονοσειρά είναι στάσιμη ως προς μέσο ή/και τάση. Η χρήση και των δύο τεστ μαζί (ADF + KPSS) παρέχει ένα αξιόπιστο στατιστικό εργαλείο για τη διάγνωση της στασιμότητας και αποφυγή ψευδών συμπερασμάτων.

Αφού επιβεβαιωθεί η στασιμότητα (ή εφαρμοστούν κατάλληλοι διαφορισμοί), περνάμε στην εκτίμηση και επιλογή μοντέλου. Τα κριτήρια AIC (Akaike Information Criterion), AICc (διορθωμένο AIC για μικρά δείγματα) και BIC (Bayesian Information Criterion) χρησιμοποιούνται για να συγκρίνουν υποψήφια μοντέλα και να επιλεγεί αυτό που επιτυγχάνει

τον βέλτιστο συμβιβασμό μεταξύ καλής προσαρμογής και παραμετρικής απλότητας. Οι μαθηματικοί τύποι είναι οι εξής:

$$AIC = 2k - 2\ln(L) \quad (2.4)$$

$$BIC = \ln(n)k - 2\ln(L) \quad (2.5)$$

$$AICc = AIC + \frac{2k(k+1)}{n-k-1} \quad (2.6)$$

Όπου k είναι ο αριθμός των παραμέτρων του μοντέλου, n το μέγεθος του δείγματος και L η μέγιστη τιμή της συνάρτησης πιθανοφάνειας. Μικρότερες τιμές AIC / BIC / $AICc$ υποδεικνύουν καλύτερο μοντέλο. Ιδιαίτερη προσοχή απαιτείται σε μικρά δείγματα, όπου η χρήση του $AICc$ είναι προτιμότερη.

Έπειτα, για την ποσοτική αξιολόγηση της προβλεπτικής ικανότητας των μοντέλων, χρησιμοποιούνται οι μετρικές σφάλματος $RMSE$ (Root Mean Squared Error), MAE (Mean Absolute Error) και $MAPE$ (Mean Absolute Percentage Error). Οι μαθηματικοί τύποι τους είναι:

$$RMSE = \sqrt{\left(\frac{1}{n} \sum_{t=1}^n (y_t - \hat{y}_t)^2\right)} \quad (2.7)$$

$$MAE = \frac{1}{n} \sum_{t=1}^n |y_t - \hat{y}_t| \quad (2.8)$$

$$MAPE = \frac{100\%}{n} \sum_{t=1}^n \left| \frac{y_t - \hat{y}_t}{y_t} \right| \quad (2.9)$$

Το $RMSE$ είναι ευαίσθητο σε ακραίες τιμές λόγω του τετραγώνου των σφαλμάτων, ενώ το MAE είναι πιο ανθεκτικό. Το $MAPE$ εκφράζει το σφάλμα ως ποσοστό, διευκολύνοντας την κατανόηση και την ερμηνεία της αποδοτικότητας της πρόβλεψης.

Τέλος, είναι απαραίτητο να εξετάζεται και αν τα residuals του μοντέλου αντιπροσωπεύουν λευκό θόρυβο, δηλαδή δεν φέρουν πληροφορία. Για τον σκοπό αυτό χρησιμοποιείται συχνά τα τεστ $Ljung-Box$ και $Box-Pierce$, τα οποία ελέγχουν αν υπάρχουν

σημαντικές αυτοσυσχετίσεις στα residuals σε προκαθορισμένα lags. Ο μαθηματικός τύπος του Ljung–Box είναι ο εξής:

$$Q = n(n+2) \sum_{k=1}^h \frac{\hat{\rho}_k^2}{n-k} \quad (2.10)$$

Όπου $\hat{\rho}_k$ είναι ο δείκτης αυτοσυσχετίσης των residuals στην καθυστέρηση k και h ο αριθμός των lags που εξετάζονται. Μη στατιστικά σημαντικές τιμές ($p\text{-value} > 0.05$) υποδεικνύουν ότι τα residuals δεν περιέχουν περαιτέρω συστηματικότητα και συνεπώς ότι το μοντέλο είναι πλήρες.

Σε εφαρμοσμένο επίπεδο, η χρήση των ARIMA και SARIMA μοντέλων σε μελέτες πρόβλεψης ατυχημάτων έχει αποδειχθεί εξαιρετικά χρήσιμη. Για παράδειγμα, στην ανάλυση των τροχαίων ατυχημάτων στην Αιθιοπία, εφαρμόστηκαν μοντέλα ARIMA για την πρόβλεψη των μηνιαίων θυμάτων και τραυματισμών στην περιοχή Amhara, με τα κριτήρια AIC και BIC να καθοδηγούν την επιλογή των βέλτιστων μοντέλων (ARIMA(2,0,0)(1,0,0) για τραυματισμούς και ARIMA(2,0,0)(1,1,0) για συνολικά ατυχήματα). Το Ljung–Box test εφαρμόστηκε για να επιβεβαιώσει την ανεξαρτησία των residuals και την καταλληλότητα των μοντέλων (Getahun, 2021).

Σε μια αντίστοιχη μελέτη για το Βελιγράδι, οι (Qian, et al., 2020) χρησιμοποίησαν το μοντέλο SARIMA για την πρόβλεψη τροχαίων ατυχημάτων, επισημαίνοντας την εποχικότητα των δεδομένων. Η επιλογή του βέλτιστου μοντέλου βασίστηκε στην ελαχιστοποίηση του MAPE (5.22%) και του RMSE, ενώ χρησιμοποιήθηκαν τα AIC / BIC για την αποφυγή υπερπροσαρμογής του μοντέλου. Η μελέτη αυτή απέδειξε τη χρησιμότητα της SARIMA ανάλυσης σε πολυπαραγοντικές, εποχιακές χρονοσειρές, όπως αυτές των οδικών ατυχημάτων (Qian, et al., 2020).

Τέλος, στην περίπτωση της Κίνας, οι (Qian, et al., 2020) σύγκριναν μοντέλα SARIMA και τεχνητών νευρωνικών δικτύων (Elman RNN) για την πρόβλεψη θανάτων από τροχαία, χρησιμοποιώντας τα RMSE, MAE και MAPE ως κριτήρια αξιολόγησης.

2.5 Μηχανική Μάθηση για Προβλέψεις Χρονοσειρών

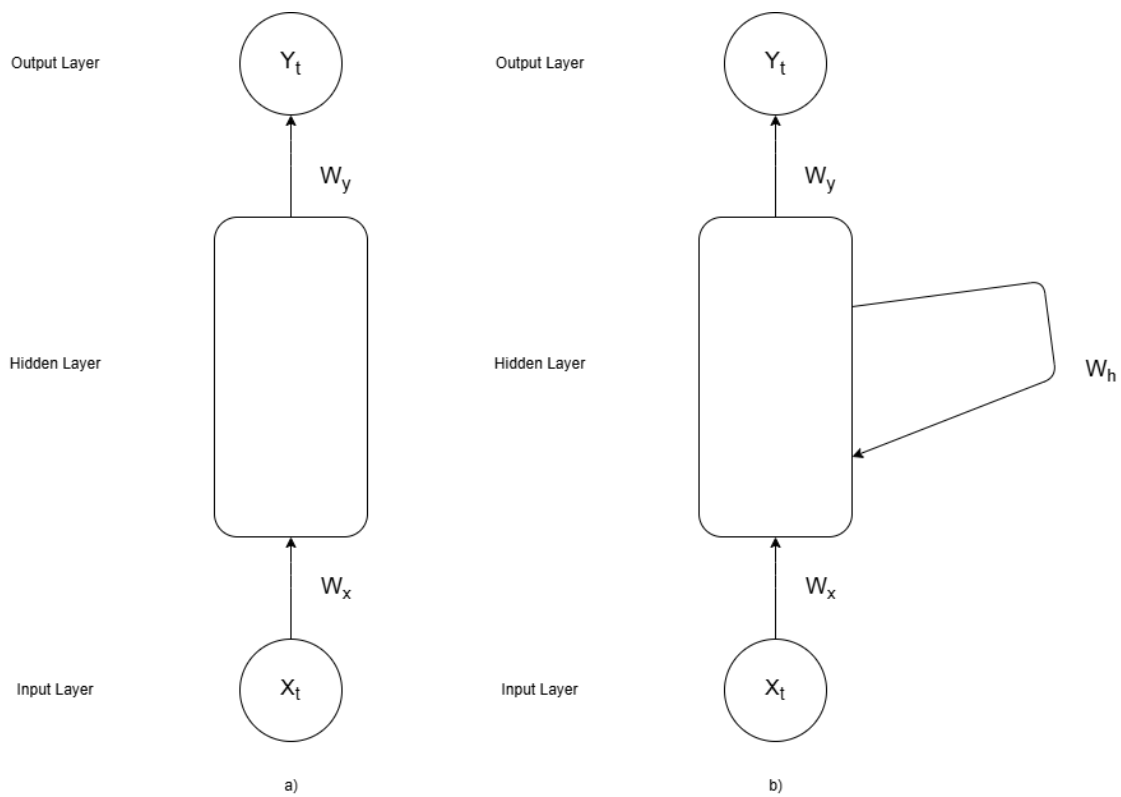
Η πρόβλεψη χρονοσειρών με μεθόδους μηχανικής μάθησης (Machine Learning - ML) έχει αποκτήσει ιδιαίτερη σημασία τα τελευταία χρόνια, ιδίως στον τομέα της οδικής ασφάλειας, όπου η ακρίβεια πρόβλεψης μπορεί να υποστηρίξει τη λήψη κρίσιμων αποφάσεων πολιτικής. Σε αντίθεση με τα στατιστικά μοντέλα όπως τα ARIMA/SARIMA, τα μοντέλα ML μπορούν να χειριστούν μη γραμμικές σχέσεις, να αξιοποιήσουν μεγάλους όγκους ετερογενών δεδομένων (π.χ. καιρικές συνθήκες, ώρα, τύπος οχήματος) και να προσαρμόζονται δυναμικά σε πολύπλοκα πρότυπα. Τεχνικές όπως τα Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM), Gated Recurrent Units (GRU), καθώς και υβριδικά μοντέλα (π.χ. CNN+LSTM ή LSTM-GBRT), έχουν ήδη εφαρμοστεί με επιτυχία σε μελέτες πρόβλεψης τροχαίων ατυχημάτων, επιτυγχάνοντας υψηλή ακρίβεια, όπως φαίνεται και σε πρόσφατες ερευνητικές εργασίες από περιοχές όπως η Κίνα, η Ινδία, το Λονδίνο και το Ιράν (Qian, et al., 2020) (Zhang, Yang, & Wushour, 2020) (Angadi & Halyal, 2024) (Balawi & Tenekeci, 2024). Οι μέθοδοι αυτές συνδυάζουν την ικανότητα εντοπισμού χωροχρονικών προτύπων με την ευελιξία ενσωμάτωσης εξωγενών μεταβλητών, καθιστώντας τις κατάλληλες για προβλέψεις υψηλής αξιοπιστίας σε πολύπλοκα συστήματα όπως το οδικό δίκτυο.

2.5.1 LSTM KAI RNNS

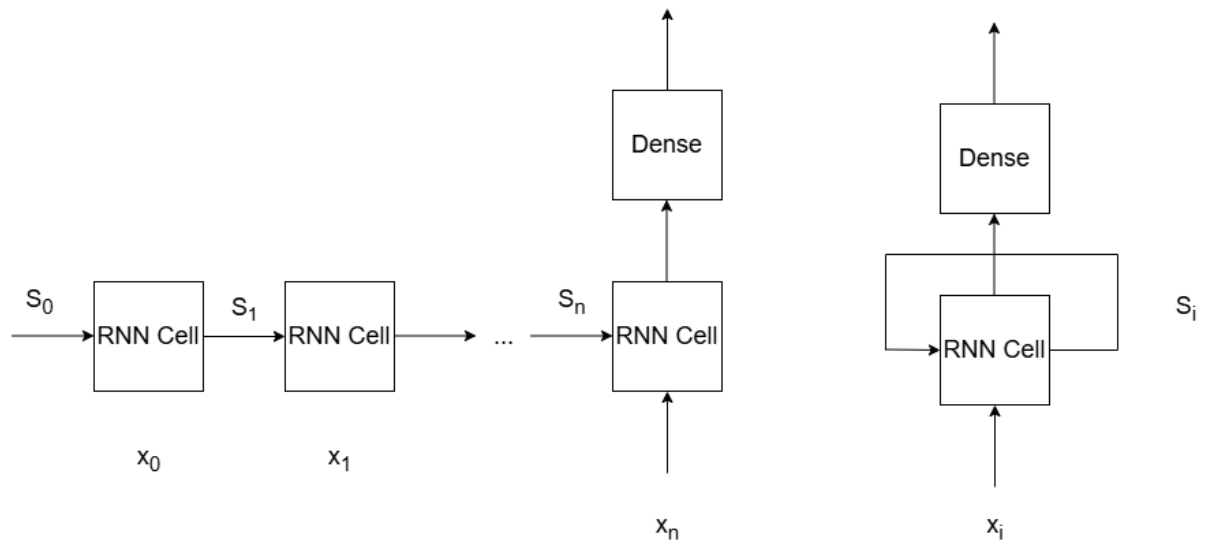
Τα Recurrent Neural Networks (RNNs) αποτελούν έναν ιδιαίτερα σημαντικό τύπο τεχνητών νευρωνικών δικτύων που σχεδιάστηκαν για την επεξεργασία και ανάλυση ακολουθιακών δεδομένων, όπως είναι οι χρονοσειρές. Σε αντίθεση με τα παραδοσιακά feedforward networks, τα RNNs διαθέτουν μια εσωτερική αναδρομική δομή η οποία τους επιτρέπει να διατηρούν μνήμη προηγούμενων εισόδων κατά την επεξεργασία της τρέχουσας πληροφορίας. Ουσιαστικά, κάθε νευρώνας RNN λαμβάνει όχι μόνο την τρέχουσα είσοδο x_t αλλά και την προηγούμενη κρυφή κατάσταση h_{t-1} , δημιουργώντας μια επαναλαμβανόμενη ροή πληροφορίας στον χρόνο. Ο βασικός μαθηματικός τύπος που περιγράφει την λειτουργία ενός RNN είναι:

$$h_t = \sigma(Wx_t + U h_{t-1} + b) \quad (2.11)$$

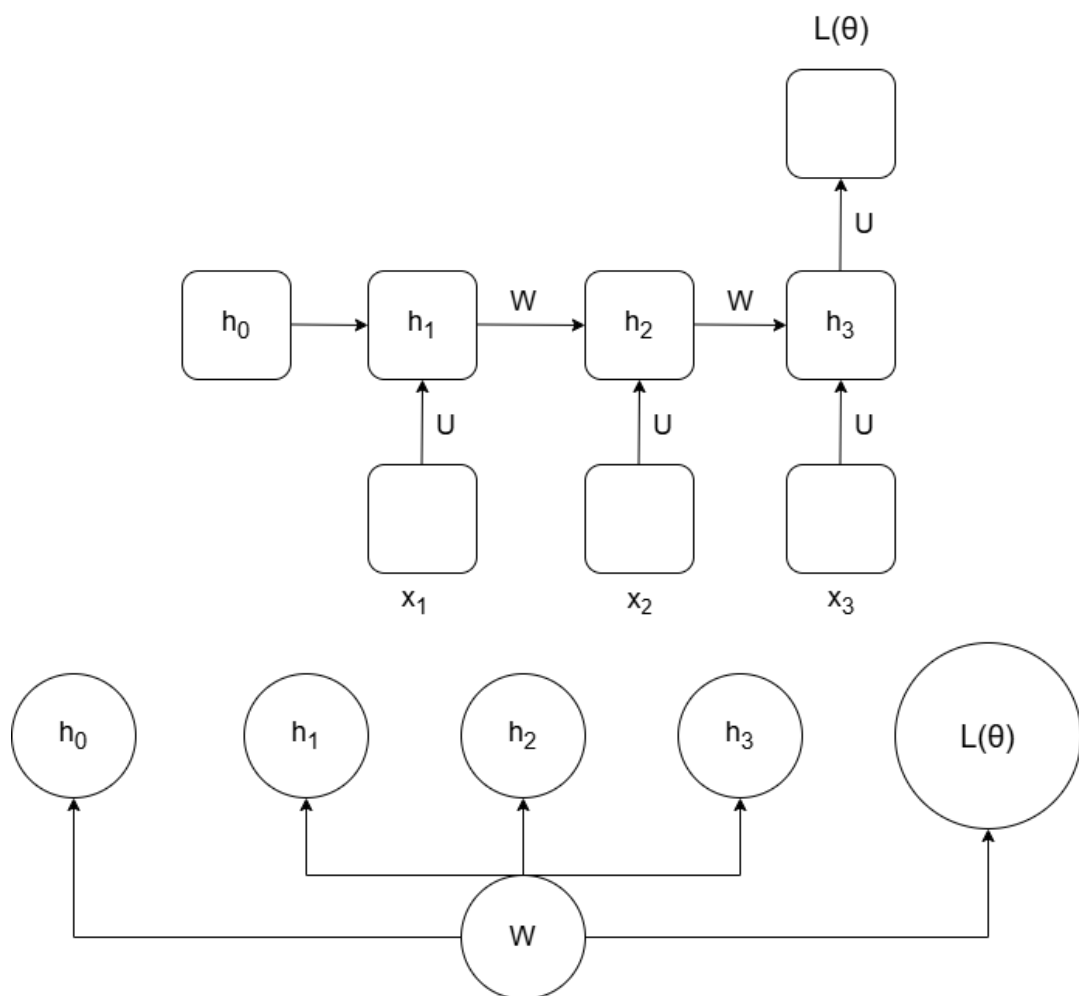
Όπου W και U είναι οι πίνακες βαρών, b είναι το διάνυσμα μετατόπισης (bias) και σ συνήθως μια μη γραμμική συνάρτηση ενεργοποίησης όπως η $\tanh$ ή η ReLU . Παρότι η αρχιτεκτονική αυτή είναι αρκετά ευέλικτη, στην πράξη τα RNNs παρουσιάζουν σοβαρά προβλήματα όταν καλούνται να μάθουν μακροπρόθεσμες εξαρτήσεις, κυρίως εξαιτίας της αποδυνάμωσης (ή εξαφάνισης) του gradient κατά την οπισθοδιάδοση μέσω του χρόνου (Backpropagation Through Time - BPTT), περιορίζοντας την αποδοτικότητά τους στην πρόβλεψη πολύπλοκων χρονικών φαινομένων όπως τα τροχαία ατυχήματα τα οποία έχουν μακροχρόνιες συσχετίσεις (Angadi & Halyal, 2024).



Σχήμα 2.5.1-1: Αρχιτεκτονικές των FeedForward Neural Network (FNN) (a) και Recurrent Neural Network (RNN) (b)



Σχήμα 2.5.1-2: Ολική Συνάρτηση του RNN



Σχήμα 2.5.1-3: Backpropagation Through Time (BPTT) in RNNs

Για να ξεπεραστούν οι περιορισμοί των απλών RNNs, αναπτύχθηκαν τα Long Short-Term Memory (LSTM) νευρωνικά δίκτυα, τα οποία εισάγουν μια πιο πολύπλοκη αλλά σταθερότερη δομή με μηχανισμούς ελέγχου της πληροφορίας μέσω πυλών (gates). Κάθε κυψέλη μνήμης LSTM περιλαμβάνει τρεις βασικές πύλες: την πύλη λήθης (forget gate), την πύλη εισόδου (input gate) και την πύλη εξόδου (output gate), οι οποίες επιτρέπουν στο δίκτυο να διατηρεί, να ενημερώνει ή να απορρίπτει πληροφορίες από την κυψέλη μνήμης. Οι αντίστοιχες μαθηματικές εξισώσεις είναι:

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f) \quad (2.12)$$

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i) \quad (2.13)$$

$$\tilde{c}_t = \tanh(W_c x_t + U_c h_{t-1} + b_c) \quad (2.14)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \quad (2.15)$$

$$o_t = \sigma(W_o x_t + U_o h_{t-1} + b_o) \quad (2.16)$$

$$h_t = o_t \odot \tanh(c_t) \quad (2.17)$$

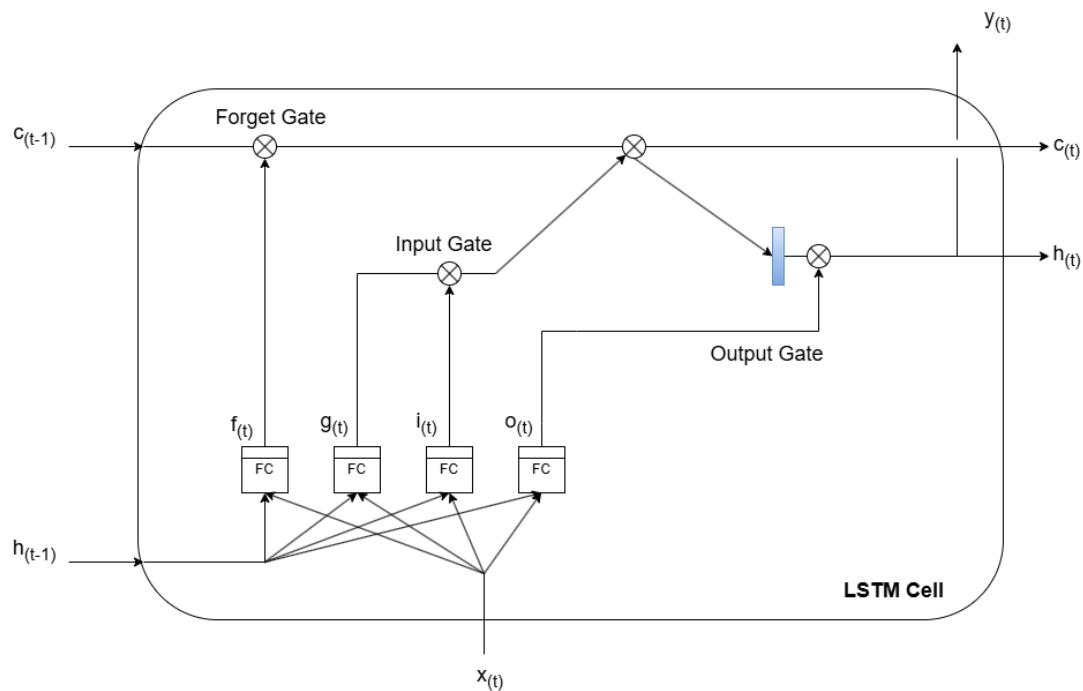
Η κυψέλη μνήμης c_t λειτουργεί ως φορέας μακροπρόθεσμης πληροφορίας, επιτρέποντας στο μοντέλο να διατηρεί κρίσιμες πληροφορίες ακόμη και σε πολύ μεγάλες χρονικές κλίμακες. Η δυνατότητα αυτή καθιστά τα LSTM κατάλληλα για την πρόβλεψη φαινομένων με έντονη χρονική εξάρτηση, όπως είναι η συχνότητα και η βαρύτητα των τροχαίων ατυχημάτων, τα οποία επηρεάζονται από μακροχρόνιες κοινωνικές, περιβαλλοντικές και εποχικές συνιστώσες (Angadi & Halyal, 2024).

Η πρακτική εφαρμογή των LSTM σε δεδομένα πρόβλεψης τροχαίων ατυχημάτων έχει αποδειχθεί πολλές φορές στην επιστημονική βιβλιογραφία. Σε μελέτες που αξιολόγησαν διάφορα deep learning μοντέλα, συμπεριλαμβανομένων των RNN, GRU, CNN-LSTM, Transformer και MLP, το LSTM κατέγραψε τις χαμηλότερες τιμές σφαλμάτων, όπως MAPE (5.850%) και RMSE (0.186) σε μηνιαία πρόβλεψη περιστατικών. Τα αποτελέσματα αυτά τεκμηριώνουν την ικανότητα του LSTM να μαθαίνει επαναλαμβανόμενα μοτίβα σε πολύπλοκες χρονοσειρές, όπως οι εβδομαδιαίες και εποχικές διακυμάνσεις των ατυχημάτων, οι οποίες επηρεάζονται από τη ροή κυκλοφορίας, τις καιρικές συνθήκες και τις κοινωνικοοικονομικές δραστηριότητες. Πιο συγκεκριμένα, σε δεδομένα από έξυπνες πόλεις όπως το Hubballi-Dharwad, η μοντελοποίηση με LSTM νευρωνικά δίκτυα επέτρεψε τον

εντοπισμό κρίσιμων χρονικών παραθύρων υψηλού κινδύνου, επιτρέποντας την έγκαιρη λήψη προληπτικών μέτρων που αφορούσαν την οδική ασφάλεια (Angadi & Halyal, 2024).

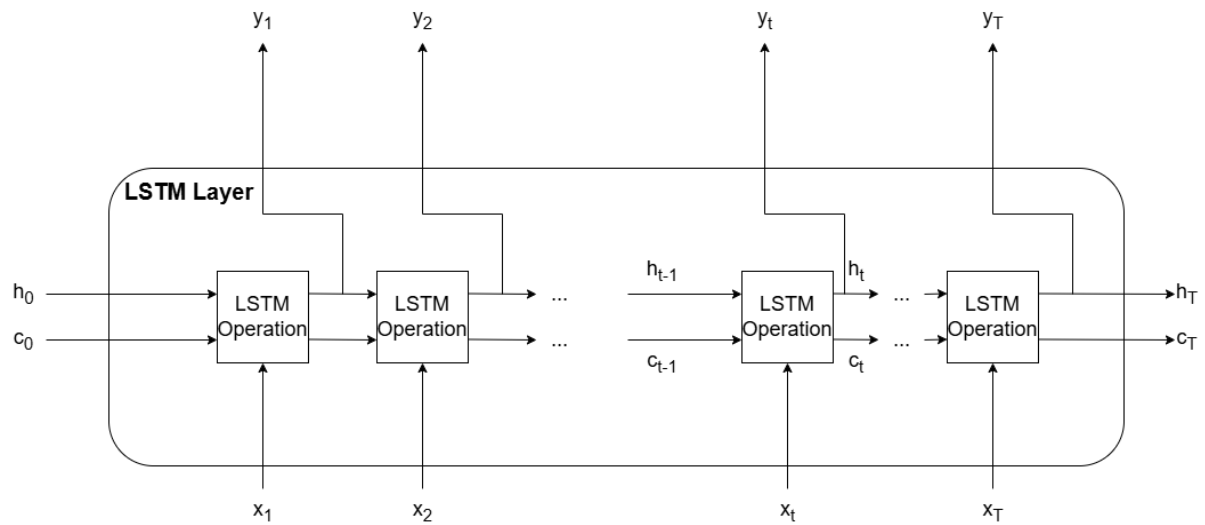
Η περαιτέρω ενίσχυση των αποτελεσμάτων πρόβλεψης επιτεύχθη μέσω της υιοθέτησης υβριδικών μοντέλων που συνδυάζουν τα πλεονεκτήματα των LSTM με εκείνα των παραδοσιακών μοντέλων μηχανικής μάθησης, όπως τα Gradient Boosted Regression Trees (GBRT). Στο μοντέλο LSTM-GBRT, η LSTM μονάδα εξάγει τα χρονικά χαρακτηριστικά των δεδομένων, τα οποία στη συνέχεια αξιοποιούνται από το GBRT για τη βελτιστοποίηση της πρόβλεψης μέσω ενός ισχυρού ensemble regression framework. Σε σύγκριση με μεμονωμένα μοντέλα (RNN, MLP, SVR, ARIMA), το LSTM-GBRT μοντέλο πέτυχε υψηλότερη ακρίβεια με $R^2 = 0.9967$ και $RMSLE = 0.0172$. Η ενσωμάτωσή του σε πλατφόρμες επίβλεψης της οδικής ασφάλειας επέτρεψε σε διαχειριστικούς φορείς μεταφορών να εντοπίσουν επικίνδυνες περιόδους και να αναπτύξουν στρατηγικές παρεμβάσεις με βάση προβλέψεις υψηλής ακρίβειας (Zhang, Yang, & Wushour, 2020).

Παρότι τα RNN αποτέλεσαν την αφετηρία για τη μοντελοποίηση χρονικών εξαρτήσεων, τα LSTM αποδείχθηκαν ανώτερα στις περισσότερες εφαρμογές λόγω της ικανότητάς τους να διαχειρίζονται μακροχρόνιες συσχετίσεις χωρίς απώλεια πληροφορίας. Ωστόσο, μια σύγχρονη εναλλακτική είναι τα GRU (Gated Recurrent Units), τα οποία ενοποιούν την πύλη λήθης και την πύλη εισόδου σε μία “update gate”, μειώνοντας την πολυπλοκότητα του μοντέλου και το χρόνο εκπαίδευσης. Όμως, παρά την χρησιμότητα των GRU, τα LSTM συνεχίζουν να υπερέχουν σε εφαρμογές όπως η πρόβλεψη τροχαίων ατυχημάτων, ιδίως όταν αυτές σχετίζονται με δυναμικά μεταβαλλόμενους παράγοντες, όπως η κινητικότητα, η εποχικότητα και η εφαρμογή πολιτικών ασφαλείας (Angadi & Halyal, 2024).



Σχήμα 2.5.1-4: Αρχιτεκτονική ενός LSTM Cell

Συνολικά, τα LSTM και RNNs έχουν συμβάλει σημαντικά στην εξέλιξη της πρόβλεψης χρονοσειρών στον τομέα της οδικής ασφάλειας. Η ικανότητά τους να μαθαίνουν και να γενικεύουν πολύπλοκα χρονικά μοτίβα έχει οδηγήσει στη βελτίωση των μεθόδων πρόληψης και στον έγκαιρο εντοπισμό επικίνδυνων περιόδων. Η βιβλιογραφία καταδεικνύει πως η εφαρμογή των LSTM σε δεδομένα τροχαίων ατυχημάτων, σε συνδυασμό με επιπλέον κοινωνικοοικονομικούς και περιβαλλοντικούς παράγοντες, οδηγεί σε μοντέλα υψηλής προβλεπτικής ισχύος, ικανά να υποστηρίξουν στρατηγικές δημόσιας πολιτικής. Η ενσωμάτωσή τους σε έξυπνες πλατφόρμες πρόβλεψης ενισχύει τον σχεδιασμό μέτρων πρόληψης, την αποτελεσματική κατανομή πόρων (όπως μονάδες πρώτης βοήθειας) και τη χάραξη στοχευμένων πολιτικών που βασίζονται σε δεδομένα υψηλής αξιοπιστίας (Angadi & Halyal, 2024).



Σχήμα 2.5.1-5: Αρχιτεκτονική ενός LSTM Layer

2.5.2 ΜΕΤΡΙΚΕΣ ΑΞΙΟΛΟΓΗΣΗΣ ΣΤΗΝ ΠΡΟΒΛΕΨΗ ΧΡΟΝΟΣΕΙΡΩΝ (LSTM)

Στο πλαίσιο πρόβλεψης χρονοσειρών με LSTM, το MAE (Mean Absolute Error) αποτελεί έναν από τους πιο απλούς αλλά ισχυρούς δείκτες απόδοσης. Ορίζεται ως ο μέσος όρος των απόλυτων διαφορών μεταξύ προβλεπόμενων (F_i) και πραγματικών (A_i) τιμών:

$$MAE = \frac{1}{n} \sum_{i=1}^n |A_i - F_i| \quad (2.18)$$

Το RMSE (Root Mean Square Error) υπολογίζεται ως η τετραγωνική ρίζα του μέσου τετραγωνικού σφάλματος και δίνει μεγαλύτερη έμφαση σε μεγάλες αποκλίσεις:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (A_i - F_i)^2} \quad (2.19)$$

Αντίθετα, ο συντελεστής προσδιορισμού R^2 μετρά την αναλογία της διακύμανσης των δεδομένων που εξηγείται από το μοντέλο:

$$R^2 = 1 - \frac{\sum (A_i - F_i)^2}{\sum (A_i - \bar{A})^2} \quad (2.20)$$

Η χρήση και των τριών μετρικών παρέχει πλήρη εικόνα για την ακρίβεια, σταθερότητα και την ερμηνευτική ικανότητα ενός LSTM μοντέλου.

Το MAPE (Mean Absolute Percentage Error) αποτελεί αναλογική εκτίμηση του σφάλματος, διευκολύνοντας συγκρίσεις μεταξύ μοντέλων και datasets διαφορετικής κλίμακας:

$$MAPE = \frac{1}{n} \sum_{i=1}^n \left| \frac{A_i - F_i}{A_i} \right| \times 100 \quad (2.21)$$

Ωστόσο, η ευαισθησία του σε μικρές ή μηδενικές πραγματικές τιμές οδηγεί στη χρήση του SMAPE (Symmetric MAPE):

$$SMAPE = \frac{100\%}{n} \sum_{i=1}^n \frac{|F_i - A_i|}{(|A_i| + |F_i|)/2} \quad (2.22)$$

Το WAPE (Weighted Absolute Percentage Error) βασίζεται στο συνολικό πραγματικό άθροισμα και είναι χρήσιμο όταν το συνολικό μέγεθος έχει σημασία, όπως στην πρόβλεψη συνολικών ατυχημάτων:

$$WAPE = \frac{\sum |A_i - F_i|}{\sum A_i} \times 100 \quad (2.23)$$

Το MASE (Mean Absolute Scaled Error) αποτελεί normalized μετρική και συγκρίνει την επίδοση του μοντέλου με εκείνη ενός baseline μοντέλου (όπως η naïve πρόβλεψη). Ορίζεται ως:

$$MASE = \frac{MAE(\text{μοντέλου})}{MAE(\text{naïve πρόβλεψης})} \quad (2.24)$$

Ένα μοντέλο με $MASE < 1$ θεωρείται καλύτερο από την naïve πρόβλεψη. Στην πρόβλεψη τροχαίων, όπου η χρονική δυναμική παρουσιάζει εποχικότητα, το MASE προσφέρει καλύτερη αξιολόγηση σε σύγκριση με παραδοσιακές μετρικές (Angadi & Halyal, 2024).

Η επίδοση ενός LSTM μοντέλου εξαρτάται επίσης έντονα από τις υπερπαραμέτρους του. Τα time steps ορίζουν το παράθυρο εισόδου και επηρεάζουν το πόσο μακροπρόθεσμες εξαρτήσεις το δίκτυο μπορεί να μάθει. Περισσότερα steps οδηγούν σε καλύτερη χρονική κατανόηση αλλά αυξάνουν το κόστος υπολογισμού. Ο αριθμός των LSTM units και dense units καθορίζει τη χωρητικότητα του δικτύου. Ένας μικρός αριθμός υπονοεί περιορισμένη εκμάθηση, ενώ υπερβολικά μεγάλος οδηγεί σε overfitting και μεγαλύτερους χρόνους εκπαίδευσης (Angadi & Halyal, 2024).

Το dropout εφαρμόζεται για την πρόληψη overfitting και αποτελεί τεχνική regularization, συνήθως με τιμές 0.2–0.5. Η επιλογή loss function (π.χ. MAE ή Huber loss) καθορίζει το τι τιμωρείται πιο αυστηρά κατά την εκπαίδευση. Το learning rate είναι κρίσιμη παράμετρος στον αλγόριθμο βελτιστοποίησης (π.χ. Adam). Χαμηλή τιμή επιβραδύνει τη σύγκλιση, ενώ υψηλή ενδέχεται να παρακάμψει το βέλτιστο σημείο (Angadi & Halyal, 2024).

Το batch size καθορίζει τον αριθμό δειγμάτων που προωθούνται ανά βήμα εκπαίδευσης. Μικρά batches παρέχουν στατιστική ποικιλία, ενώ μεγάλα batches ενισχύουν τη σταθερότητα του gradient. Τα epochs (π.χ. 100 ή 200) καθορίζουν πόσες φορές το δίκτυο βλέπει τα δεδομένα.

Σύμφωνα με τη μελέτη των (Angadi & Halyal, 2024), επιλέχθηκαν 2 LSTM layers με 50 μονάδες η καθεμία, Dense layer με 1 μονάδα και dropout 0.2. Το learning rate ήταν 0.0004 με optimizer Adam, και loss function το MAE. Τα epochs ήταν 100. Η επιλογή έγινε με βάση την απόδοση στο validation set και τα γραφήματα σύγκλισης της loss function ανά epoch.

Επιπλέον, στην μελέτη του (Zhang, Yang, & Wushour, 2020), δοκιμάστηκαν διάφορα πλήθη κρυφών μονάδων (5–13), και διαπιστώθηκε ότι 11 μονάδες με 2 LSTM layers έδωσαν το μικρότερο RMSE (2378.01), επιβεβαιώνοντας τη σημασία του fine-tuning μέσω grid search και cross-validation.

Κεφάλαιο 3 Συλλογή και Διαχείριση Δεδομένων

Η ορθολογική συλλογή και διαχείριση δεδομένων αποτελεί το θεμέλιο κάθε επιστημονικής μελέτης που αποσκοπεί στην εξαγωγή αξιόπιστων συμπερασμάτων και στη δημιουργία ακριβών μοντέλων πρόβλεψης. Στο πλαίσιο της παρούσας διπλωματικής εργασίας, η ανάλυση και πρόβλεψη των οδικών ατυχημάτων και θανάτων στην Ελλάδα βασίζεται σε χρονοσειρές υψηλής ανάλυσης, οι οποίες συλλέχθηκαν από επίσημες πηγές όπως το Εργαστήριο Οδικής Ασφάλειας του ΕΜΠ. Η διαχείριση αυτών των δεδομένων περιλαμβάνει τη διαδικασία καθαρισμού, μετασχηματισμού, αναδόμησης και εμπλουτισμού των αρχικών παρατηρήσεων, με σκοπό την εξαγωγή χρήσιμων features για στατιστική ανάλυση και μηχανική μάθηση. Επιπλέον, η κατανόηση της χρονικής δυναμικής και η διατήρηση της συνέπειας και της πληρότητας των δεδομένων κρίνονται ζωτικής σημασίας, τόσο για την εγκυρότητα της ανάλυσης όσο και για την αποτελεσματικότητα των μελλοντικών προβλέψεων. Η ενότητα αυτή τεκμηριώνει τη μεθοδολογία που ακολουθήθηκε για τη συλλογή, αποθήκευση και επεξεργασία των δεδομένων, ενώ παράλληλα περιγράφει τις βασικές αρχές προτυποποίησης και επιλεγμένων τεχνικών προεπεξεργασίας, θέτοντας τις βάσεις για την εφαρμογή μοντέλων πρόβλεψης όπως τα ARIMA, SARIMA και LSTM που εξετάζονται σε επόμενα κεφάλαια.

3.1 Ανάκτηση Δεδομένων από την ΒΔ SANTRA

Το αρχικό Dataset με δεδομένα προήλθε από τη βάση δεδομένων SANTRA του Εθνικού Μετσόβιου Πολυτεχνείου (ΕΜΠ). Η βάση δεδομένων SANTRA (Statistical Analysis of Road Traffic Accidents) αποτελεί ένα εξειδικευμένο εθνικό αποθετήριο δεδομένων που αναπτύχθηκε και συντηρείται από το Εργαστήριο Οδικής Ασφάλειας του ΕΜΠ. Στόχος της είναι η συλλογή, αποθήκευση και ανάλυση στατιστικών στοιχείων οδικών ατυχημάτων στην Ελλάδα.

Η βάση περιλαμβάνει δεδομένα υψηλής χρονικής ανάλυσης (εβδομαδιαία, μηνιαία και ετήσια) που αφορούν τον αριθμό ατυχημάτων, τις απώλειες ζωών, τους τραυματισμούς, καθώς

και ενδεχομένως άλλες παραμέτρους όπως τοποθεσία, τύπος εμπλεκόμενων οχημάτων και καιρικές συνθήκες.

3.1.1 ΣΧΟΛΙΑ ΚΑΙ ΠΕΡΙΓΡΑΦΗ ΤΩΝ ΔΕΔΟΜΕΝΩΝ ΑΠΟ ΤΗΝ ΒΔ SANTRA

Τα δεδομένα από τη βάση SANTRA είχαν αρχικά τη μορφή Count Panel Data και παρείχαν εβδομαδιαία καταγραφή του πλήθους των οδικών ατυχημάτων και των θανάτων που επήλθαν από αυτά, καλύπτοντας γεωγραφικά το σύνολο της Ελληνικής επικράτειας. Το αρχείο δόθηκε σε μορφή Excel και ήταν οργανωμένο σε δύο βασικά φύλλα εργασίας. Το ένα περιλάμβανε τις εβδομαδιαίες τιμές των συνολικών ατυχημάτων (accidents) και το άλλο τις αντίστοιχες τιμές των θανάτων (fatalities). Η χρονική κάλυψη των εν λόγω πινάκων εκτείνεται από το 1996 έως και το 2022.

Το πρώτο μέρος των δεδομένων (1996 – 2014), ήταν διατεταγμένο κατά τέτοιον τρόπο ώστε κάθε γραμμή να αντιπροσωπεύει μια συγκεκριμένη εβδομάδα του μήνα, ενώ οι στήλες αντιστοιχούν στα ημερολογιακά έτη από το 1996 έως το 2014. Οι μήνες εμφανίζονται διαδοχικά και συνοδεύονται από αριθμητική ένδειξη εβδομάδας (π.χ. ΙΑΝΟΥΑΡΙΟΣ – 1η έως 6η εβδομάδα). Το ίδιο ακριβώς μοτίβο τηρείται και για τα δεδομένα των θανάτων.

Αντίθετα, από το έτος 2015 και μετά, τα δεδομένα έχουν επεξεργαστεί διαφορετικά και παρέχονται σε μορφή pivot tables. Συγκεκριμένα, στο φύλλο εργασίας κάθε έτους εμφανίζονται δύο ξεχωριστοί πίνακες: ο πρώτος για τα ατυχήματα και ο δεύτερος για τους θανάτους, με τις εβδομάδες να είναι οργανωμένες σε γραμμές και τους μήνες σε στήλες.

Με μια πρώτη ματιά, ήταν εμφανές ότι τα δεδομένα δεν ήταν έτοιμα για άμεση χρήση σε μορφή χρονοσειράς. Η αρχική τους δομή, ειδικά για τα έτη 1996–2014, δεν είχε κάποια ενιαία γραμμική διάταξη του χρόνου (δηλαδή “date-indexed” μορφή). Αντί για μία στήλη που να περιλαμβάνει τις ημερομηνίες και μία άλλη τα αντίστοιχα δεδομένα (π.χ. ατυχήματα, θάνατοι), υπήρχαν πολλαπλές στήλες για κάθε έτος και πολλές γραμμές για κάθε μήνα, γεγονός που καθιστούσε την άμεση ανάλυση ή οπτικοποίηση μέσω μοντέλων χρονοσειράς ανέφικτη χωρίς προηγούμενο μετασχηματισμό.

Η μορφή των pivot tables για τα έτη μετά το 2015 παρουσίαζε επίσης δομικές ιδιαιτερότητες, καθώς οι εβδομάδες απεικονίζονταν ως γραμμές και οι μήνες ως στήλες, δημιουργώντας έναν πίνακα δύο διαστάσεων που απέκρυπτε τη χρονολογική ακολουθία των τιμών. Συνεπώς, ήταν σαφές από την αρχή ότι για να μπορέσουν τα δεδομένα να

χρησιμοποιηθούν αποτελεσματικά στις αναλύσεις που ακολουθούν, θα έπρεπε να προηγηθεί προεπεξεργασία, “ξεδίπλωμα” (unpivoting), και ανακατασκευή του χρονικού δείκτη, ώστε κάθε παρατήρηση να συνοδεύεται από ένα μοναδικό χρονικό στιγμιότυπο (timestamp) και ένα μοναδικό target value.

του ατυχήματος του ατυχήματος		1996	1997	1998	1999	2000	2001	2002	2003
ΙΑΝΟΥΑΡΙΟΣ	1	313	176	163	88	55	325	165	232
ΙΑΝΟΥΑΡΙΟΣ	2	322	347	387	363	346	407	257	293
ΙΑΝΟΥΑΡΙΟΣ	3	299	352	410	461	306	304	277	289
ΙΑΝΟΥΑΡΙΟΣ	4	327	413	379	443	369	338	270	281
ΙΑΝΟΥΑΡΙΟΣ	5	160	337	367	406	389	115	148	102
ΙΑΝΟΥΑΡΙΟΣ	6				48	116			
ΦΕΒΡΟΥΑΡΙΟΣ	5	154	67				237	178	152
ΦΕΒΡΟΥΑΡΙΟΣ	6	326	387	418	308	294	360	320	243
ΦΕΒΡΟΥΑΡΙΟΣ	7	346	408	365	375	429	350	323	221
ΦΕΒΡΟΥΑΡΙΟΣ	8	386	405	415	399	390	368	313	224
ΦΕΒΡΟΥΑΡΙΟΣ	9	256	338	417	376	318	181	143	112
ΦΕΒΡΟΥΑΡΙΟΣ	10				70	138			
ΜΑΡΤΙΟΣ	9	95	45				221	220	149
ΜΑΡΤΙΟΣ	10	328	437	434	375	254	402	360	240
ΜΑΡΤΙΟΣ	11	353	434	371	456	440	361	296	228
ΜΑΡΤΙΟΣ	12	316	376	384	350	406	411	321	240
ΜΑΡΤΙΟΣ	13	365	412	328	336	355	373	252	203
ΜΑΡΤΙΟΣ	14	55	127	155	170	337			

Σχήμα 3.1.1-1: Δεδομένα Ατυχημάτων Ιανουάριος – Μάρτιος από τα Έτη 1996 – 2003
(Μορφή Count Panel Data)

Accidents												
Week/Month	1	2	3	4	5	6	7	8	9	10	11	12
1	113											
2	182											
3	218											
4	182											
5	67	112										
6		160										
7		177										
8		176										
9		75	118									
10			154									
11			176									
12			183									
13			139	29								
14				233								
15				216								
16				225								
17				206								
18				33	176							
19					260							
20					242							
21					257							
22					130	93						
23						244						
24						241						
25						248						
26						180	33					

Σχήμα 3.1.1-2: Δεδομένα Ατυχημάτων Εβδομάδων 1-26 και Μηνών 1-12 του Έτους 2015
(Μορφή Pivot Table)

Επιπλέον, κατά την εξερεύνηση των δεδομένων για την περίοδο 1996–2014, προέκυψε ένα δομικό πρόβλημα σχετικό με την κατανομή των εβδομάδων εντός των μηνών. Συγκεκριμένα, διαπιστώθηκε ότι η έννοια της "εβδομάδας" στο dataset δεν ακολουθούσε το διεθνώς αποδεκτό ημερολογιακό πρότυπο, κατά το οποίο η εβδομάδα ξεκινά συνήθως από Δευτέρα ή Κυριακή και περιλαμβάνει 7 ημέρες. Αντίθετα, σε περιπτώσεις όπου ένας μήνας δεν ξεκινούσε την Κυριακή, ο καταμερισμός των ημερών οδηγούσε σε δημιουργία "σπασμένων" εβδομάδων (π.χ. από Τετάρτη έως Σάββατο), οι οποίες καταγράφονταν ως πλήρεις εβδομάδες. Αυτό είχε ως αποτέλεσμα την ασυνεπή καταγραφή συγκεκριμένων εβδομάδων και έτσι αντί για 4 που θα έπρεπε να υπάρχουν σε κάθε μήνα, το σύννηθος ήταν να υπάρχουν 5, 6 ή και 7.

Calendar for Year 1996

Ιανουάριος							Φεβρουάριος							Μάρτιος						
Δε	Τρ	Τε	Πέ	Πα	Σά	Κυ	Δε	Τρ	Τε	Πέ	Πα	Σά	Κυ	Δε	Τρ	Τε	Πέ	Πα	Σά	Κυ
1	2	3	4	5	6	7				1	2	3	4				1	2	3	
8	9	10	11	12	13	14	5	6	7	8	9	10	11	4	5	6	7	8	9	10
15	16	17	18	19	20	21	12	13	14	15	16	17	18	11	12	13	14	15	16	17
22	23	24	25	26	27	28	19	20	21	22	23	24	25	18	19	20	21	22	23	24
29	30	31					26	27	28	29				25	26	27	28	29	30	31
5:○ 13:● 20:● 27:○							4:○ 12:● 19:● 26:○							5:○ 12:● 19:● 27:○						

Σχήμα 3.1.1-3: Φεβ. 1996, 1^η «εβδομάδα» είναι η 01/02 – 03/02

3.2 Συλλογή Δεδομένων από Διάφορες Πηγές για Κοινωνικοοικονομικούς Παράγοντες

Για την πληρέστερη κατανόηση των μεταβλητών που ενδέχεται να επηρεάζουν τον αριθμό των τροχαίων ατυχημάτων και των θανάτων, κρίθηκε απαραίτητη η ενσωμάτωση κοινωνικοοικονομικών δεικτών στο σύνολο των δεδομένων. Οι παράγοντες αυτοί περιλαμβάνουν μεταβλητές όπως το Ακαθάριστο Εγχώριο Προϊόν (ΑΕΠ), η ανεργία, η τιμή καυσίμων, ο αριθμός των κυκλοφορούντων οχημάτων, καθώς και η κατανάλωση αλκοόλ, οι οποίες έχουν κατά καιρούς τεκμηριωθεί στη διεθνή βιβλιογραφία ως σημαντικά συνδεδεμένες με την οδική ασφάλεια. Η συλλογή των σχετικών δεδομένων πραγματοποιήθηκε από πολλαπλές επίσημες πηγές, όπως η ΕΛΣΤΑΤ, η Eurostat, το Υπουργείο Οικονομικών, καθώς και διεθνείς βάσεις δεδομένων (OECD, World Bank), με στόχο τη δημιουργία ενός εμπλουτισμένου πίνακα χαρακτηριστικών που να επιτρέπει την πολυπαραγοντική ανάλυση και πρόβλεψη των ατυχημάτων.

3.2.1 ΣΧΟΛΙΑ ΚΑΙ ΠΕΡΙΓΡΑΦΗ ΤΩΝ ΚΟΙΝΩΝΙΚΟΟΙΚΟΝΟΜΙΚΩΝ ΔΕΔΟΜΕΝΩΝ ΑΠΟ ΔΙΑΦΟΡΕΣ ΠΗΓΕΣ

Όσον αφορά τους κοινωνικοοικονομικούς παράγοντες, δεν χρησιμοποιήθηκαν όλα τα δεδομένα στην ανάλυση, είτε λόγω ελλιπούς χρονικής κάλυψης είτε λόγω περιορισμένης συσχέτισης με τον υπό μελέτη στόχο. Παρακάτω παρουσιάζεται αναλυτικός πίνακας, στον οποίο καταγράφονται τα διαθέσιμα δεδομένα, η πηγή προέλευσης, καθώς και το αν εντάχθηκαν τελικά στη βάση δεδομένων για ανάλυση.

Πίνακας 3.2.1-1: Κοινωνικοοικονομικά Δεδομένα και Πηγές (α)

Δεδομένα	Διαθέσιμα στην Eurostat	Διαθέσιμα στην ΕΛΣΤΑΤ	Διαθέσιμα από άλλες Πηγές	Επιλέχθηκαν για Εισαγωγή στην ΒΔ
Τέλη Κυκλοφοριακής Συμφόρησης	✗	✗	✓	✗
Βελτιώσεις στην Ιατρική Τεχνολογία/Περίθαλψη	✓	✗	✓	✗
Θέσεις Εργασίας στον Τομέα των Μεταφορών	✗	✓	✗	✗
Διενυθέντα Οχηματοχιλιόμετρα	✗	✗	✓	✗
Μέτρα Οδικής Ασφάλειας	✗	✗	✓	✗
Νόμος για Χρήση Ζώνης Ασφαλείας	✗	✗	✓	✗
Βαθμοί Ποινής για Απρόσεκτη Οδήγηση	✗	✗	✓	✗
Οδήγηση με Ασφάλιση	✗	✓	✓	✗
Κατάσταση του οδικού δικτύου	✗	✓	✗	✗
Νόμος για Χρήση Ζώνης Ασφαλείας για Παιδιά	✗	✓	✓	✗
Τιμές καυσίμων	✗	✗	✓	✗

Πίνακας 3.2.1-2: Κοινωνικοοικονομικά Δεδομένα και Πηγές (β)

Δεδομένα	Διαθέσιμα στην Eurostat	Διαθέσιμα στην ΕΛΣΤΑΤ	Διαθέσιμα από άλλες Πηγές	Επιλέχθηκαν για Εισαγωγή στην ΒΔ
Ζήτηση Μεταφοράς	✗	✗	✗	✗
Όγκος Οδικής Κυκλοφορίας	✗	✗	✓	✗
Συμπεριφορά των Οδηγών	✗	✗	✓	✗
Επιλογή Μέσου Κυκλοφορίας	✗	✗	✓	✗
Εκπομπές καυσαερίων Λόγω Κυκλοφοριακού Φόρτου	✗	✗	✓	✗
Χιλιόμετρα Οδικής Κυκλοφορίας	✗	✗	✓	✗
ΑΕΠ	✓	✓	✗	✓
Ποσοστό Ανεργίας	✓	✓	✗	✓
Πληθυσμός	✓	✓	✗	✓
Νέα αυτοκίνητα		✓	✗	✗
Μήκος οδικού δικτύου	✗	✓	✗	✗
Υπάρχοντα Οχήματα	✓	✓	✓	✓

Μετά από αξιολόγηση της διαθεσιμότητας, της πληρότητας και της χρονικής συνέπειας των δεδομένων, αποφασίστηκε να ενσωματωθούν 16 μεταβλητές που καλύπτουν ένα ευρύ φάσμα οικονομικών, πληθυσμιακών, ενεργειακών και κοινωνικών χαρακτηριστικών. Οι μεταβλητές αυτές καταγράφηκαν αποκλειστικά σε ετήσια βάση, γεγονός που περιορίζει, αλλά ταυτόχρονα καθιστά πιο σταθερό το αναλυτικό επίπεδο σε αυτή τη φάση της έρευνας.

Συγκεκριμένα, τα πεδία που εντάχθηκαν στη βάση δεδομένων είναι τα εξής: `total_gdp`, `gdp_per_capita`, `unemployment`, `population`, `environmental_taxes`, `vehicles`, `petrol_cons`, `ng_cons`, `el_cons`, `fatal_per_million_veh`, `speed_infr`, `drink_infr`, `belt_infr`, `helmet_infr`, `inflation_rate`, και `gini_index`. Οι μεταβλητές αυτές επιλέχθηκαν βάσει της θεωρητικής τους συνάφειας με τη συχνότητα και τη σοβαρότητα των τροχαίων ατυχημάτων, καθώς και βάσει της πληρότητας των εγγραφών τους κατά την εξεταζόμενη χρονική περίοδο.

Πιο συγκεκριμένα, οι ενεργειακές μεταβλητές `el_cons` (κατανάλωση ηλεκτρικής ενέργειας), `ng_cons` (κατανάλωση φυσικού αερίου) και `petrol_cons` (κατανάλωση πετρελαίου) αντλήθηκαν από την επίσημη βάση δεδομένων της Eurostat. Οι οικονομικοί και πληθυσμιακοί δείκτες όπως `environmental_taxes` (περιβαλλοντικοί φόροι), `gdp_per_capita` (κατά κεφαλήν ΑΕΠ), `population` (πληθυσμός), `unemployment` (ποσοστό ανεργίας), `total_gdp` (συνολικό ΑΕΠ), `vehicles` (συνολικός αριθμός οχημάτων), `gini_index` (δείκτης ανισότητας) και `inflation_rate` (πληθωρισμός) λήφθηκαν από την ΕΛΣΤΑΤ, η οποία αποτελεί τον κεντρικό στατιστικό φορέα της Ελλάδας. Ο δείκτης `fatal_per_million_veh` (θάνατοι ανά εκατομμύριο οχημάτων) υπολογίστηκε με βάση υπάρχοντα δεδομένα ατυχημάτων και οχημάτων. Τέλος, οι δείκτες οδικής συμπεριφοράς `speed_infr`, `drink_infr`, `belt_infr` και `helmet_infr` προήλθαν από στοιχεία του Εργαστηρίου Οδικής Ασφάλειας του ΕΜΠ (NTUA) και από την έκδοση Road Safety Figures 2017, η οποία παρέχει επιμέρους στατιστικά για παραβάσεις σχετιζόμενες με την οδική ασφάλεια. Το σύνολο αυτών των μεταβλητών σχημάτισε ένα Dataset με τιμές από το 1996 έως το 2023, το οποίο ονομάστηκε `yearly_indicators.csv`, ώστε αργότερα να εισαχθεί στην ΒΔ με περισσότερη ευκολία.

Η απόφαση να διατηρηθεί το επίπεδο της ανάλυσης σε ετήσιο για αυτές τις μεταβλητές βασίστηκε στο γεγονός ότι η πλειονότητα των σχετικών δεδομένων είτε δεν ήταν διαθέσιμη σε υψηλότερη χρονική ανάλυση (μηνιαία/εβδομαδιαία) είτε παρουσίαζε σημαντικά κενά. Η επιλογή της ετήσιας κλίμακας διασφαλίζει τη στατιστική συνέπεια και επιτρέπει την απρόσκοπτη σύνδεση με τα αντίστοιχα ετήσια aggregates των οδικών ατυχημάτων και

θανάτων. Επιπλέον, στην παρούσα εργασία, τα δεδομένα αυτά αποτέλεσαν το πρώτο επίπεδο εμπλουτισμού της βάσης, πάνω στο οποίο θα οικοδομηθεί η επέκταση προς πιο granular μοντελοποίηση σε επόμενα στάδια της μελέτης.

3.3 Αρχικοί Μετασχηματισμοί στα Δεδομένα Πριν την Εισαγωγή τους στη ΒΔ

Πέραν της δημιουργίας του αρχείου yearly_indicators.csv, το οποίο περιλαμβάνει τους βασικούς κοινωνικοοικονομικούς δείκτες σε ετήσια ανάλυση, κρίθηκε απαραίτητο να πραγματοποιηθούν επιπλέον μετασχηματισμοί στα δεδομένα των τροχαίων ατυχημάτων και των θανάτων. Ειδικότερα, τα δεδομένα από τη ΒΔ SANTRA για την περίοδο 1996–2014 παρείχονταν σε count panel data μορφή, με εβδομάδες ενσωματωμένες στους μήνες και τα έτη ως ξεχωριστές στήλες. Η μορφή αυτή, αν και κατάλληλη για οπτική ανάγνωση, δεν προσέφερε την επιθυμητή συνέπεια και ενιαία δομή με τα δεδομένα της περιόδου 2015 και μετά, τα οποία είχαν ήδη οργανωθεί σε μορφή pivot tables.

Προκειμένου να επιτευχθεί ομοιομορφία στη βάση δεδομένων και ολοκληρωσιμότητα της χρονικής σειράς, αποφασίστηκε να μετατραπούν και τα παλαιότερα δεδομένα (1996–2014) στην ίδια pivot μορφή. Η νέα δομή προσφέρει πυκνότερη και πιο "συμπαγέμενη" απεικόνιση, διαχωρίζοντας τις εβδομάδες ως γραμμές και τους μήνες ως στήλες, επιτρέποντας έτσι την καλύτερη ενσωμάτωσή τους σε στατιστικά μοντέλα και σε διαδικασίες προεπεξεργασίας. Επιπλέον, με αυτόν τον τρόπο διατηρείται δομική συνέχεια στο dataset, χωρίς μεταβολές στη μορφή των εγγραφών κατά τη μετάβαση από το 2014 στο 2015.

Για την επίτευξη αυτής της μετατροπής, αναπτύχθηκαν εξειδικευμένα Python scripts, τα οποία εκτελούσαν ενέργειες όπως unpivot/reshape των αρχικών πινάκων, κατανομή "σπασμένων" εβδομάδων στους κατάλληλους μήνες και δημιουργία των τελικών πινάκων ανά έτος σε pivot tables μορφή. Το τελικό αποτέλεσμα ήταν ένα χρονικά συνεπές σύνολο δεδομένων για τα έτη 1996–2022, έτοιμο προς εισαγωγή στη ΒΔ.

Α	Β	Γ	Δ	Ε	ΣΤ	Ζ	Η	Θ	Ι	Κ	Λ	Μ	Ν	Ξ	Ο	Π	Ρ	Σ	Τ	Υ
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
ΙΑΝΟΥΑΡΙΟΣ	1	41	19	16	6	7	35	24	35	20	28	18	34	29	32	21	17	13	22	17
ΙΑΝΟΥΑΡΙΟΣ	2	32	31	41	28	32	46	18	28	28	34	19	21	20	21	14	16	15	11	12
ΙΑΝΟΥΑΡΙΟΣ	3	29	27	25	38	19	21	31	26	36	21	27	30	30	21	21	16	18	19	18
ΙΑΝΟΥΑΡΙΟΣ	4	32	36	43	39	29	28	13	16	17	20	20	29	16	18	18	13	29	14	5
ΙΑΝΟΥΑΡΙΟΣ	5	12	24	31	26	21	9	11	6	10	13	21	12	9	6	11	5		3	2
ΙΑΝΟΥΑΡΙΟΣ	6				5	16														
ΦΕΒΡΟΥΑΡΙΟΣ	5	15	6				27	14	16	19	13	17	18	23	16	16	8	5	11	6
ΦΕΒΡΟΥΑΡΙΟΣ	6	32	20	51	27	23	35	22	33	31	29	28	23	26	25	20	27	9	16	18
ΦΕΒΡΟΥΑΡΙΟΣ	7	32	36	29	26	46	26	21	18	11	30	22	30	12	22	19	18	16	12	14
ΦΕΒΡΟΥΑΡΙΟΣ	8	34	39	36	35	26	50	30	42	37	17	25	23	21	28	22	12	25	10	10
ΦΕΒΡΟΥΑΡΙΟΣ	9	28	22	31	29	32	17	10	10	16	17	12	11	10	17	2	5	5	6	3
ΦΕΒΡΟΥΑΡΙΟΣ	10				9	15														
ΜΑΡΤΙΟΣ	9	16	2				20	24	15	16	16	17	18	13	17	15	9	5	4	7
ΜΑΡΤΙΟΣ	10	32	40	41	32	14	25	24	26	28	27	30	31	27	20	24	17	8	7	16
ΜΑΡΤΙΟΣ	11	19	32	27	39	38	22	23	25	20	27	32	40	20	15	18	19	22	14	9
ΜΑΡΤΙΟΣ	12	26	20	30	37	40	22	36	33	30	24	17	18	35	25	25	24	12	14	11
ΜΑΡΤΙΟΣ	13	19	37	27	22	17	37	18	16	28	24	22	26	25	30	21	16	24	16	7
ΜΑΡΤΙΟΣ	14	4	7	12	16	29														
ΑΠΡΙΛΙΟΣ	13						15	4	1		5	4	7			1		2		1
ΑΠΡΙΛΙΟΣ	14	25	14	15	16	5	33	26	25	38	13	40	36	25	20	29	17	23	17	17
ΑΠΡΙΛΙΟΣ	15	47	39	23	42	38	35	28	41	34	26	24	28	32	24	18	25	17	19	16
ΑΠΡΙΛΙΟΣ	16	39	37	46	48	47	31	24	25	34	30	34	34	31	27	25	24	15	21	17
ΑΠΡΙΛΙΟΣ	17	28	61	54	41	40	33	24	24	23	20	26	31	31	39	17	17	14	12	5
ΑΠΡΙΛΙΟΣ	18	18	37	25	49	42	5	3	3	8	4	6	1	8	3	7		1		
ΑΠΡΙΛΙΟΣ	19					8														
ΜΑΪΟΣ	18	25	11	11	4		38	49	22	25	25	24	24	31	18	15	15	14	6	11
ΜΑΪΟΣ	19	46	33	38	28	49	18	23	33	42	25	25	33	20	19	31	20	24	18	21

Σχήμα 3.2.1-1: Δεδομένα Τροχαίων Ατυχημάτων και Θανάτων πριν την Μεταμόρφωση

Α	Β	Γ	Δ	Ε	ΣΤ	Ζ	Η	Θ	Ι	Κ	Λ	Μ	Ν	Ξ	Ο	Π	Ρ	Σ	Τ	Υ	Φ	Χ
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
Accidents	Week/Month	1	2	3	4	5	6	7	8	9	10	11	12									
1	313																					
2	322																					
3	299																					
4	327																					
5	160	154																				
6		326																				
7		346																				
8		386																				
9		256	95																			
10			328																			
11			353																			
12			316																			
13			365																			
14			55	422																		
15				481																		
16				362																		
17				409																		
18				196	312																	
19					512																	
20					456																	
21					539																	
22					380	75																
23						579																
24						568																
25						635																
26						534																
27						75	532															
28							587															
29							547															
30							516															
31							292	254														
32								597														
33								510														
34								495														
35								524														
36									534													
37									526													
38									536													
39									510													
40									342													
41									483													
42									476													
43									481													

Σχήμα 3.2.1-2: Δεδομένα Τροχαίων Ατυχημάτων και Θανάτων μετά την Μεταμόρφωση

3.4 Δημιουργία Βάσης Δεδομένων

Για την αποτελεσματική οργάνωση, αποθήκευση και αξιοποίηση των δεδομένων που συγκεντρώθηκαν στα προηγούμενα στάδια, υλοποιήθηκε σχεσιακή βάση δεδομένων με την ονομασία RoadAccidents, χρησιμοποιώντας την PostgreSQL με το UI του pgAdmin4. Η βάση σχεδιάστηκε με τρόπο που να υποστηρίζει πολλαπλά επίπεδα χρονικής ανάλυσης, καθώς και τον εμπλουτισμό των χρονοσειρών με κοινωνικοοικονομικούς δείκτες.

3.4.1 ΣΧΕΣΕΙΣ ΜΕΤΑΞΥ ΜΕΤΑΒΛΗΤΩΝ ΚΑΙ ΔΙΑΚΡΙΣΗ ΤΩΝ DATA TYPES

Η δομή της ΒΔ αποτελείται από πέντε βασικούς πίνακες: `weekly_accidents`, `monthly_accidents`, `yearly_accidents`, `yearly_indicators` και `yearly_indicators_filled`. Οι πρώτοι τρεις περιέχουν τα δεδομένα των τροχαίων ατυχημάτων και θανάτων σε εβδομαδιαία, μηνιαία και ετήσια ανάλυση αντίστοιχα, διατηρώντας τις μεταβλητές `accidents` και `fatalities`, με χρονικά κλειδιά (`year`, `month`, `week`) ανά περίπτωση. Οι δύο τελευταίοι πίνακες (`yearly_indicators` και `yearly_indicators_filled`) αποθηκεύουν τις 16 κοινωνικοοικονομικές μεταβλητές που αναφέρθηκαν προηγουμένως για κάθε έτος.

Κατά τη σχεδίαση της ΒΔ, ιδιαίτερη έμφαση δόθηκε στη σωστή επιλογή τύπων δεδομένων για κάθε μεταβλητή, καθώς και στη δήλωση των κατάλληλων `primary` και `foreign keys` με σκοπό τη διασφάλιση της μοναδικότητας και της ακεραιότητας των εγγραφών. Οι πίνακες `weekly_accidents`, `monthly_accidents`, και `yearly_accidents` περιλαμβάνουν αριθμητικά πεδία για τα έτη, τους μήνες, τις εβδομάδες, καθώς και για τις τιμές των `accidents` και `fatalities`. Όλες αυτές οι μεταβλητές αποθηκεύονται ως `integers`, καθώς αντιστοιχούν σε διακριτές ποσοτικές μετρήσεις. Οι χρονικές παράμετροι (`year`, `month`, `week`) δηλώνονται ως `NOT NULL` και αποτελούν συνδυαστικά τα `primary keys` στους αντίστοιχους πίνακες.

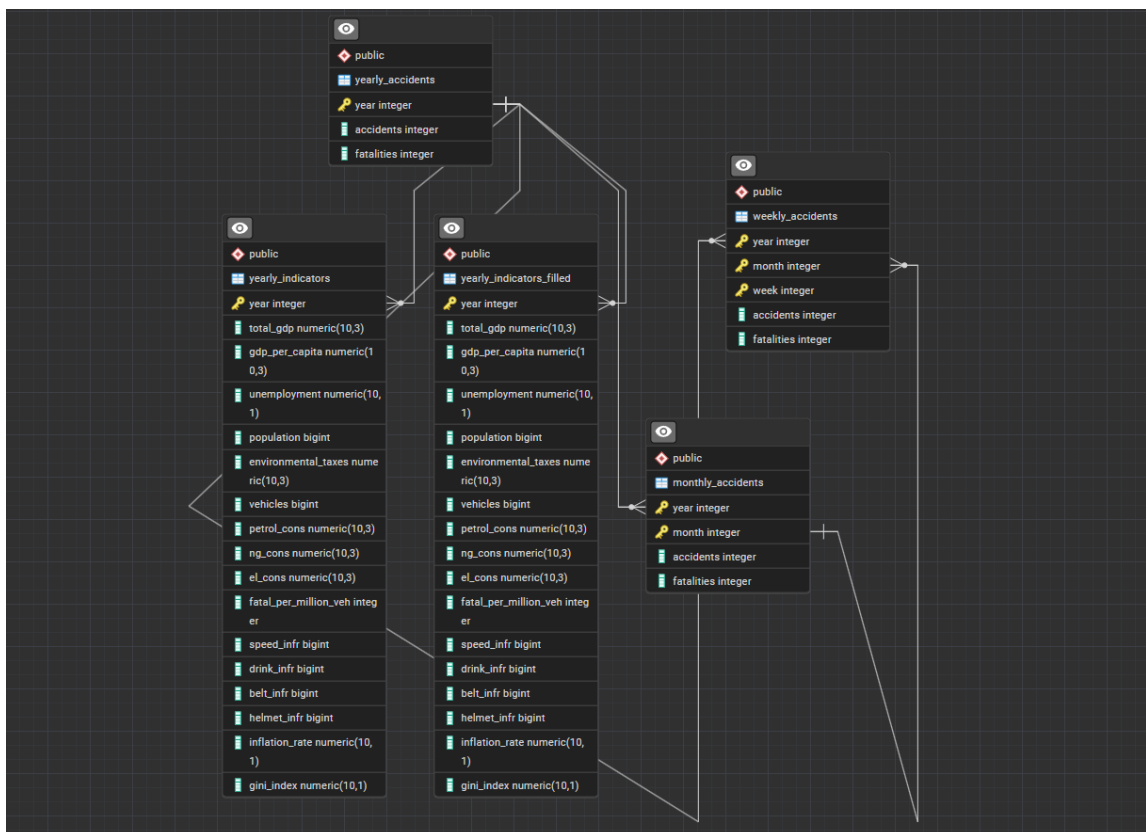
Αναφορικά με τις κοινωνικοοικονομικές μεταβλητές του πίνακα `yearly_indicators`, έγινε προσεκτική επιλογή μεταξύ `integer`, `bigint` και `numeric data types`, ανάλογα με τη φύση και την ακρίβεια των δεδομένων. Δείκτες όπως `total_gdp`, `gdp_per_capita`, `unemployment`, `environmental_taxes`, και `inflation_rate` αποθηκεύονται ως `numeric(10,3)` ή `numeric(10,1)`, επιτρέποντας την αναπαράσταση δεκαδικών τιμών με ακρίβεια έως και τρία δεκαδικά ψηφία. Από την άλλη, μεταβλητές όπως `vehicles`, `belt_infr`, και `drink_infr`, οι οποίες έχουν μεγάλο

ακέραιο εύρος, επιλέχθηκαν ως bigint ώστε να διατηρούν την αριθμητική ακρίβεια χωρίς περιορισμούς μεγέθους.

Η χρήση του πεδίου year ως primary key στον πίνακα yearly_indicators διασφαλίζει ότι κάθε εγγραφή αντιστοιχεί μοναδικά σε ένα συγκεκριμένο έτος, επιτρέποντας συγχρονισμό με τους αντίστοιχους πίνακες ατυχημάτων σε ετήσια βάση. Αυτό καθιστά δυνατή τη μελλοντική δημιουργία JOIN queries για συσχετιστική ανάλυση μεταξύ π.χ. αριθμού τροχαίων και οικονομικών μεταβλητών.

3.4.2 ΔΙΑΓΡΑΜΜΑ ΟΝΤΟΤΗΤΩΝ-ΣΥΣΧΕΤΙΣΕΩΝ (ER) ΤΗΣ ΒΑΣΗΣ ΔΕΔΟΜΕΝΩΝ

Η εσωτερική σχεδίαση της βάσης παρουσιάζεται στο διάγραμμα οντοτήτων-συσχετίσεων, όπου φαίνεται η σύνδεση όλων των πινάκων μέσω της μεταβλητής year, διασφαλίζοντας τη χρονολογική ευθυγράμμιση των δεδομένων.



Σχήμα 3.4.2-1: ER Schema της ΒΔ RoadAccidents.

3.4.3 ΕΙΣΑΓΩΓΗ ΔΕΔΟΜΕΝΩΝ ΣΤΗ ΒΔ ROADACCIDENTS

Η εισαγωγή των εβδομαδιαίων τιμών για τις μεταβλητές accidents και fatalities στον πίνακα weekly_accidents πραγματοποιήθηκε μέσω χειροκίνητης διαδικασίας με τη χρήση SQL UPDATE queries, για κάθε έτος, μήνα και εβδομάδα από το 1996 έως και το 2022. Για κάθε χρόνο δημιουργήθηκε ξεχωριστό query με τη χρήση του CASE WHEN, το οποίο καθόριζε με ακρίβεια τις τιμές των πεδίων accidents και fatalities για τις έξι εβδομάδες του μήνα. Οι τιμές αυτές αντλήθηκαν απευθείας από τα αρχικά Excel αρχεία της βάσης SANTRA, κατόπιν επεξεργασίας και καθαρισμού. Η προσέγγιση αυτή παρείχε πλήρη αντιστοίχιση εβδομάδων, μηνών και ετών, επιτρέποντας τον ακριβή συγχρονισμό των δεδομένων μεταξύ της βάσης και των πρωτογενών πινάκων.

Η επιλογή της χειροκίνητης μεθόδου αντί για αυτοματοποιημένα scripts έγινε συνειδητά, καθώς προσέφερε αυξημένο έλεγχο και ακρίβεια στην εισαγωγή. Λόγω της ιδιαίτερης μορφής των εβδομάδων σε ορισμένα έτη (π.χ. μη πλήρεις εβδομάδες στην αρχή ή στο τέλος του μήνα, εβδομάδες που εκτείνονται σε δύο μήνες), η αυτοματοποίηση παρουσίαζε αυξημένο ρίσκο ασυνεπειών. Επομένως, κάθε τιμή εισήχθη προσεκτικά, εβδομάδα προς εβδομάδα, με την υποβοήθηση του ημερολογιακού ελέγχου και της προεπεξεργασίας των αρχείων Excel σε pivot μορφή. Η χρήση της εντολής UPDATE ανά μήνα, με ξεκάθαρες συνθήκες WHERE year = ... AND month = ..., διασφάλισε την στοχευμένη ενημέρωση των εγγραφών χωρίς απώλειες ή επικαλύψεις. Ένα παράδειγμα query για συμπλήρωση ενός μήνα ενός χρόνου για τα τροχαία ατυχήματα έχει ως εξής:

```
UPDATE weekly_accidents
SET accidents =
CASE
    WHEN week = 1 THEN ...
    WHEN week = 2 THEN ...
    WHEN week = 3 THEN ...
    WHEN week = 4 THEN ...
    WHEN week = 5 THEN ...
    WHEN week = 6 THEN NULL
ELSE accidents
END
```

```
WHERE year = ...  
AND month = ...  
AND week BETWEEN 1 AND 6;
```

Αφού συμπληρώθηκαν πλήρως οι εβδομαδιαίες εγγραφές, ακολούθησε το aggregation των δεδομένων σε μηνιαίο και ετήσιο επίπεδο, με χρήση SQL εντολών που βασίστηκαν σε queries όπως SELECT SUM(...). Για παράδειγμα, για την ενημέρωση του πίνακα monthly_accidents, χρησιμοποιήθηκε το ακόλουθο query:

```
UPDATE monthly_accidents  
SET fatalities = (  
    SELECT SUM(COALESCE(wd.fatalities, 0))  
    FROM weekly_accidents wd  
    WHERE wd.year = monthly_accidents.year  
    AND wd.month = monthly_accidents.month  
)  
WHERE year = ...;
```

Το query αυτό εξασφάλιζε ότι κάθε μηνιαία εγγραφή λάμβανε το συνολικό άθροισμα των θανάτων από τις αντίστοιχες εβδομάδες. Η χρήση της συνάρτησης COALESCE εγγυόταν ότι οι κενές τιμές (NULL) δεν επηρέαζαν τα αποτελέσματα. Αντίστοιχα, για την ενημέρωση του πίνακα yearly_accidents, αθροίστηκαν οι τιμές από τον monthly_accidents μέσω παρόμοιου μηχανισμού aggregation.

Όσον αφορά τον πίνακα yearly_indicators, η εισαγωγή πραγματοποιήθηκε με τη χρήση της εντολής COPY, η οποία επέτρεψε την μαζική εισαγωγή των ετήσιων μεταβλητών από το αντίστοιχο αρχείο CSV. Η εντολή που χρησιμοποιήθηκε ήταν:

```
COPY public.yearly_indicators  
FROM 'file path'  
DELIMITER ';' CSV HEADER;
```

Τέλος, μετά την ολοκλήρωση όλων των εισαγωγών, πραγματοποιήθηκε έλεγχος ακεραιότητας (integrity check). Ο έλεγχος αυτός περιλάμβανε διασταυρώσεις τιμών μεταξύ εβδομαδιαίων, μηνιαίων και ετήσιων πινάκων ώστε να επιβεβαιωθεί ότι τα σύνολα συμφωνούν, καθώς και σύγκριση των τελικών τιμών με τα πρωτογενή Excel αρχεία. Η διαδικασία αυτή απέδειξε ότι οι τιμές είχαν εισαχθεί με συνέπεια και ακρίβεια.

3.4.4 ΑΞΙΟΛΟΓΗΣΗ ΤΗΣ ΣΥΜΠΛΗΡΩΣΗΣ ΕΛΛΙΠΩΝ ΤΙΜΩΝ ΚΑΙ ΑΙΤΙΟΛΟΓΗΣΗ ΤΗΣ ΑΠΟΡΡΙΨΗΣ ΤΗΣ

Στο πλαίσιο της προετοιμασίας του πίνακα `yearly_indicators_filled`, έγινε αρχικά μια προσπάθεια να συμπληρωθούν οι ελλειπείς τιμές ορισμένων κοινωνικοοικονομικών μεταβλητών, οι οποίες παρουσίαζαν μικρά ή μεγαλύτερα κενά στη χρονική τους κάλυψη. Για τον σκοπό αυτό αξιοποιήθηκε μια απλή μεθοδολογική προσέγγιση μέσω γραμμικής παλινδρόμησης (linear regression), υλοποιημένη με Python, με στόχο την πρόβλεψη των ελλειπουσών τιμών βάσει της σχέσης τους με άλλες διαθέσιμες μεταβλητές. Αν και η μέθοδος αυτή φάνηκε αρχικά ελκυστική λόγω της απλότητάς της, η εφαρμογή της ανέδειξε προβλήματα. Συγκεκριμένα, οι προβλεπόμενες τιμές παρουσίαζαν υπερβολική εξομάλυνση (over-smoothing), αποτυγχάνοντας να αποδώσουν τις πραγματικές διακυμάνσεις ή διαρθρωτικές μεταβολές που ενδέχεται να υπήρχαν στα ιστορικά δεδομένα.

Επιπλέον, για κάποιες μεταβλητές με χαμηλή συσχέτιση με τις υπόλοιπες ή με διακεκομμένες χρονικές σειρές (π.χ. `environmental_taxes`, `ng_cons`, `gini_index`), η συμπλήρωση βάσει παλινδρόμησης οδήγησε σε ανακριβείς ή τεχνητά "λογικοφανείς" τιμές, που όμως δεν στηρίζονταν σε εμπειρικά δεδομένα. Μετά από αξιολόγηση των αποτελεσμάτων και σύγκριση με τις αρχικές πηγές, αποφασίστηκε ότι η συγκεκριμένη προσέγγιση δεν εγγυάται επιστημονική εγκυρότητα, ιδιαίτερα σε ένα σύνολο δεδομένων που προορίζεται να τροφοδοτήσει μοντέλα πρόβλεψης. Κατά συνέπεια, επιλέχθηκε να μη συμπληρωθούν και συμπεριληφθούν τεχνητά οι τιμές, αλλά αντίθετα να παραμείνουν κενές, ώστε να φιλτραριστούν στο επόμενο στάδιο του Data Preparation, κατά τη φάση καθαρισμού και επιλογής μεταβλητών στον τελικό κώδικα.

3.4.5 QUERIES ΕΞΑΓΩΓΗΣ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΤΕΛΙΚΑ DATASETS

Στο τελευταίο στάδιο της διαχείρισης της βάσης δεδομένων RoadAccidents, πραγματοποιήθηκε η εξαγωγή των τελικών, επεξεργασμένων datasets που θα χρησιμοποιηθούν για στατιστική ανάλυση, εποχική μοντελοποίηση και συστήματα πρόβλεψης. Η διαδικασία βασίστηκε σε SQL queries, που επέτρεψαν την αναλυτική εξαγωγή δεδομένων σε εβδομαδιαίο, μηνιαίο και ετήσιο επίπεδο, καθώς και τον εμπλουτισμό τους με κοινωνικοοικονομικούς δείκτες. Η αρχή έγινε με την εξαγωγή του dataset weekly_accidents, μέσω του παρακάτω ερωτήματος:

```
SELECT *  
FROM weekly_accidents  
ORDER BY year, month, week;
```

Ακολούθησε η εξαγωγή των μηνιαίων δεδομένων από τον πίνακα monthly_accidents, ο οποίος προήλθε από aggregation των εβδομαδιαίων τιμών μέσω SQL queries. Η αντίστοιχη εντολή εξαγωγής ήταν:

```
SELECT *  
FROM monthly_accidents  
ORDER BY year, month;
```

Το τρίτο σύνολο εξήχθη από τον πίνακα yearly_accidents με την εντολή:

```
SELECT *  
FROM yearly_accidents  
ORDER BY year;
```

Και το Dataset με τις κοινωνικοοικονομικές μεταβλητές με την εντολή:

```
SELECT *  
FROM yearly_indicators  
ORDER BY year;
```

Για να συνδυαστούν τα δεδομένα ατυχημάτων με τους κοινωνικοοικονομικούς δείκτες, σχεδιάστηκαν queries με JOIN μεταξύ των πινάκων weekly_accidents και yearly_indicators, βάσει του κοινού πεδίου year. Το query:

```
SELECT wi.*,  
       yi.total_gdp, yi.gdp_per_capita, yi.unemployment, yi.population,  
       yi.environmental_taxes, yi.vehicles, yi.petrol_cons, yi.ng_cons,  
       yi.el_cons, yi.fatal_per_million_veh, yi.speed_infr, yi.drink_infr,  
       yi.belt_infr, yi.helmet_infr, yi.inflation_rate, yi.gini_index  
FROM weekly_accidents wi  
JOIN yearly_indicators yi ON wi.year = yi.year  
ORDER BY wi.year, wi.month, wi.week;
```

παρήγαγε το dataset weekly_accidents_yearly_indicators, στο οποίο κάθε εβδομάδα εμπλουτίζεται με τα αντίστοιχα ετήσια socioeconomic χαρακτηριστικά. Παρόμοια λογική εφαρμόστηκε σε μηνιαίο και ετήσιο επίπεδο.

Συνοψίζοντας, η τελική λίστα των εξαγόμενων συνόλων δεδομένων περιλαμβάνει τα εξής επτά αρχεία:

- weekly_accidents.csv
- monthly_accidents.csv
- yearly_accidents.csv
- yearly_indicators.csv
- weekly_accidents_yearly_indicators.csv
- monthly_accidents_yearly_indicators.csv
- yearly_accidents_yearly_indicators.csv

Όλα τα παραπάνω datasets προέκυψαν από επαναλήψιμα SQL queries, τα οποία διατηρούν τόσο την ακεραιότητα της πληροφορίας όσο και τη λογική συνέχεια στη χρονική δομή.

The screenshot shows the pgAdmin 4 interface with a SQL query executed. The query selects various indicators from the 'weekly_accidents' table and joins them with 'yearly_indicators' on the 'year' field. The results are displayed in a table with 18 rows and 15 columns.

Query:

```

SELECT wi.*,
       yi.total_gdp, yi.gdp_per_capita, yi.unemployment, yi.population,
       yi.environmental_taxes, yi.vehicles, yi.petrol_cons, yi.ng_cons,
       yi.el_cons, yi.fatal_per_million_voh, yi.speed_infr, yi.drink_infr,
       yi.belt_infr, yi.helmet_infr, yi.inflation_rate, yi.gini_index
FROM weekly_accidents wi
JOIN yearly_indicators yi ON wi.year = yi.year
ORDER BY wi.year, wi.month, wi.week;

```

Data Output:

year	month	week	accidents	fatalities	total_gdp	gdp_per_capita	unemployment	population	environmental_taxes	vehicles	petrol_cons	ng_cons	el_cons	fatal_per_million_voh
integer	integer	integer	integer	integer	numeric (10,3)	numeric (10,3)	numeric (10,1)	bigint	numeric (10,3)	bigint	numeric (10,3)	numeric (10,3)	numeric (10,3)	integer
1996	1	1	313	41	100.664	9.489	426.8	10608821	[null]	3797234	[null]	[null]	[null]	[null]
1996	1	2	322	32	100.664	9.489	426.8	10608821	[null]	3797234	[null]	[null]	[null]	[null]
1996	1	3	299	29	100.664	9.489	426.8	10608821	[null]	3797234	[null]	[null]	[null]	[null]
1996	1	4	327	32	100.664	9.489	426.8	10608821	[null]	3797234	[null]	[null]	[null]	[null]
1996	1	5	160	12	100.664	9.489	426.8	10608821	[null]	3797234	[null]	[null]	[null]	[null]
1996	1	6	[null]	[null]	100.664	9.489	426.8	10608821	[null]	3797234	[null]	[null]	[null]	[null]
1996	2	1	154	15	100.664	9.489	426.8	10608821	[null]	3797234	[null]	[null]	[null]	[null]
1996	2	2	326	32	100.664	9.489	426.8	10608821	[null]	3797234	[null]	[null]	[null]	[null]
1996	2	3	346	32	100.664	9.489	426.8	10608821	[null]	3797234	[null]	[null]	[null]	[null]
1996	2	4	386	34	100.664	9.489	426.8	10608821	[null]	3797234	[null]	[null]	[null]	[null]
1996	2	5	256	28	100.664	9.489	426.8	10608821	[null]	3797234	[null]	[null]	[null]	[null]
1996	2	6	[null]	[null]	100.664	9.489	426.8	10608821	[null]	3797234	[null]	[null]	[null]	[null]
1996	3	1	95	16	100.664	9.489	426.8	10608821	[null]	3797234	[null]	[null]	[null]	[null]
1996	3	2	328	32	100.664	9.489	426.8	10608821	[null]	3797234	[null]	[null]	[null]	[null]
1996	3	3	353	19	100.664	9.489	426.8	10608821	[null]	3797234	[null]	[null]	[null]	[null]
1996	3	4	316	26	100.664	9.489	426.8	10608821	[null]	3797234	[null]	[null]	[null]	[null]
1996	3	5	385	19	100.664	9.489	426.8	10608821	[null]	3797234	[null]	[null]	[null]	[null]
1996	3	6	55	4	100.664	9.489	426.8	10608821	[null]	3797234	[null]	[null]	[null]	[null]

Successfully run. Total query runtime: 224 msec. 1944 rows affected.

Σχήμα 3.4.5-1: Περιβάλλον pgAdmin4 και Ενδεικτικό Query για Εξαγωγή των Datasets

Κεφάλαιο 4 Προεπεξεργασία Δεδομένων και Feature Engineering

Η διαδικασία της προεπεξεργασίας δεδομένων και του Feature Engineering αποτέλεσε βασικό στάδιο στη μετάβαση από τα ακατέργαστα datasets που εξήχθησαν από τη ΒΔ RoadAccidents προς μορφές συμβατές για στατιστικά μοντέλα αλλά και μοντέλα πρόβλεψης. Σκοπός ήταν η δημιουργία καθαρών, συνεπών και εμπλουτισμένων data frames που μπορούν να αξιοποιηθούν για ανάλυση χρονοσειρών, ανίχνευση συσχετίσεων και εκπαίδευση μοντέλων μηχανικής μάθησης. Όλη η διαδικασία υλοποιήθηκε σε περιβάλλον RStudio, αξιοποιώντας τη γλώσσα R.

Η επεξεργασία εφαρμόστηκε τόσο στα απλά όσο και στα συνδυαστικά datasets, ώστε να παραχθούν παράγωγες μεταβλητές που ενισχύουν την πληροφοριακή πυκνότητα κάθε εγγραφής. Μέσα από μια ακολουθία βημάτων, τα δεδομένα καθαρίστηκαν από ελλιπείς τιμές, επανατυποποιήθηκαν, και επαναδομήθηκαν χρονικά ώστε να επιτρέπουν την ακριβή ευθυγράμμιση μεταβλητών μεταξύ εβδομαδιαίων, μηνιαίων και ετήσιων επιπέδων.

4.1 Εισαγωγή Δεδομένων και ορισμός του Working Directory

Η αρχική φάση του Data Preparation και Feature Engineering περιλάμβανε την εισαγωγή όλων των τελικών datasets που είχαν εξαχθεί από την ΒΔ. Ορίστηκε το κατάλληλο working directory στο περιβάλλον RStudio ώστε όλα τα datasets .csv να είναι άμεσα προσβάσιμα χωρίς την ανάγκη επαναλαμβανόμενης δήλωσης όλων των file paths. Η ανάγνωση των αρχείων πραγματοποιήθηκε με χρήση της συνάρτησης read.csv(), με κάθε dataset να αποθηκεύεται σε ξεχωριστό dataframe (π.χ. acc.w, acc.m, ind.y, κ.λπ.), με σαφείς και συντομογραφημένες ονομασίες.

Η ονοματολογία των data frames σχεδιάστηκε συστηματικά. Δηλαδή, το πρόθεμα acc αναφέρεται στα datasets που περιέχουν τα τροχαία ατυχήματα και τους θανάτους, το ind στους κοινωνικοοικονομικούς δείκτες και οι καταλήξεις .w, .m, .y δηλώνουν χρονικό βάθος (weekly, monthly και yearly). Στα data frames που περιλαμβάνουν τα κοινωνικοοικονομικά δεδομένα

και τα δεδομένα τροχαίων ατυχημάτων και θανάτων μαζί, χρησιμοποιήθηκε η κατάληξη `.ind.y`, δηλώνοντας ενοποίηση με τον πίνακα `yearly_indicators`.

4.2 Βασικός Καθαρισμός και Μετατροπές Τύπων Μεταβλητών

Αμέσως μετά την εισαγωγή των δεδομένων, εφαρμόστηκε βασικός καθαρισμός σε όλα τα data frames. Οι μεταβλητές `accidents` και `fatalities` σε κάθε πίνακα μετατράπηκαν σε `numeric`, καθώς κατά την εισαγωγή από CSV ενδέχεται να εμφανιστούν ως `character` ή `factor`, ειδικά όταν υπάρχουν κενά ή συμβολικές τιμές. Οι χρονικές μεταβλητές (`year`, `month`, `week`) μετατράπηκαν σε αρχικό στάδιο σε `factor` με αντίστοιχα `levels`, ώστε να διασφαλιστεί η ορθή σειρά ταξινόμησης και η αποφυγή αυτόματης αναδιάταξης από την R.

Παράλληλα, στα data frames κοινωνικοοικονομικών δεικτών (Δηλαδή τα `ind.y` και τα συνδυαστικά data frames), εφαρμόστηκε μαζική μετατροπή όλων των πεδίων πλην της χρονικής μεταβλητής `year` σε `numeric`, μέσω της συνάρτησης `lapply()`. Επιπλέον, έγινε αρχικός έλεγχος και αφαίρεση των `NULL` και `NA` τιμών, σε όλα τα δεδομένα, ώστε να εξασφαλιστεί η πληρότητα και η ομοιομορφία των εγγραφών.

4.3 Δημιουργία Χρονικών Δεικτών και Ενοποίηση Ημερομηνιών

Για να επιτευχθεί χρονική συνέπεια και ευθυγράμμιση μεταξύ όλων των data frames, δημιουργήθηκαν νέες μεταβλητές `Date type` σε εβδομαδιαίο, μηνιαίο και ετήσιο επίπεδο. Οι μεταβλητές αυτές (`week_date`, `month_date`, `year_date`) συγκέντρωσαν σε μία στήλη τις πληροφορίες που προηγουμένως απεικονίζονταν σε τρεις στήλες (`year`, `month`, `week`).

Η πιο σύνθετη επεξεργασία εφαρμόστηκε στον εβδομαδιαίο μετασχηματισμό (`acc.w_mod` και `acc.w.ind.y_mod`), όπου χρησιμοποιήθηκε η `lubridate::wday()` και η συνάρτηση `case_when()` για τον εντοπισμό της πρώτης Κυριακής κάθε μήνα, και από εκεί υπολογίστηκαν οι ημερομηνίες αρχής κάθε εβδομάδας (και σε ειδικές περιπτώσεις οι "μισές" εβδομάδες στο τέλος του μήνα). Αντίστοιχα, στους πίνακες `monthly` και `yearly accidents`, οι ημερομηνίες κατασκευάστηκαν απλά με τη συνάρτηση `as.Date(paste(...))`, αφού τα δεδομένα ήταν εξαρχής ομογενοποιημένα.

4.4 Διαδικασία και Μεθοδολογία του Feature Engineering

Η διαδικασία του Feature Engineering αποτέλεσε έναν από τους σημαντικότερους άξονες της παρούσας επεξεργασίας, καθώς μέσω αυτής ενισχύθηκαν τα αρχικά δεδομένα με στόχο τη βελτίωση της ερμηνευσιμότητας και της προβλεπτικής τους ικανότητας. Η μεθοδολογία που ακολουθήθηκε περιλάμβανε τον σταδιακό εμπλουτισμό των datasets με νέα χαρακτηριστικά που αποτυπώνουν χρονικές, εποχικές, κοινωνικοοικονομικές και διάφορες άλλες πληροφορίες. Η ανάπτυξη αυτών των μεταβλητών δεν περιορίστηκε μόνο σε απλούς δείκτες, αλλά περιέλαβε και πιο σύνθετα μεγέθη όπως κυλιόμενους μέσους όρους, χρονικές καθυστερήσεις (lags), αθροιστικά χαρακτηριστικά (cumulative features), και μετασχηματισμούς που προέκυψαν από τον συνδυασμό εξωτερικών παραμέτρων.

Τα βήματα του Feature Engineering εφαρμόστηκαν με συνέπεια σε κάθε μορφή του διαθέσιμου υλικού, τόσο στα απλά datasets που περιλάμβαναν μόνο δεδομένα τροχαίων ατυχημάτων, όσο και στα σύνθετα, στα οποία είχαν συγχωνευθεί και οι κοινωνικοοικονομικοί δείκτες. Με αυτόν τον τρόπο, εξασφαλίστηκε η ενιαία αντιμετώπιση και επεξεργασία των δεδομένων σε όλα τα επίπεδα χρονικής ανάλυσης (εβδομαδιαίο, μηνιαίο, ετήσιο), επιτρέποντας συγκρίσιμα αποτελέσματα.

4.4.1 SEASON, QUARTER, ISSUMMER, ISWINTER

Η ένταξη εποχικών χαρακτηριστικών στα datasets των ατυχημάτων αποτέλεσε ένα από τα πρώτα βήματα στη διαδικασία της δημιουργίας επεξηγηματικών μεταβλητών. Αρχικά, δημιουργήθηκε μεταβλητή που προσδιορίζει το τρίμηνο στο οποίο ανήκει κάθε παρατήρηση (quarter), βάσει του μήνα. Η ταξινόμηση έγινε ως εξής: Ιανουάριος–Μάρτιος στο 1ο τρίμηνο, Απρίλιος–Ιούνιος στο 2ο, Ιούλιος–Σεπτέμβριος στο 3ο και Οκτώβριος–Δεκέμβριος στο 4ο.

Παράλληλα, έγινε κατηγοριοποίηση του μήνα και με βάση τις εποχές του έτους, με τη μεταβλητή season, ώστε να αποδοθεί στην κάθε παρατήρηση η τιμή "Winter", "Spring", "Summer" ή "Autumn", ανάλογα με τον μήνα αναφοράς. Στη συνέχεια, με σκοπό να διευκολυνθούν οι αναλύσεις που χρειάζονται binary variables, προέκυψαν δύο επιπλέον binary flags: is_summer και is_winter, τα οποία λαμβάνουν την τιμή 1 αν η παρατήρηση αντιστοιχεί σε καλοκαιρινό ή χειμερινό μήνα αντίστοιχα, και 0 διαφορετικά.

4.4.2 HOLIDAYS AND SPECIAL EVENTS

Η επίδραση των εθνικών, θρησκευτικών και ειδικών αργιών στα τροχαία ατυχήματα είναι ένα ευρέως διαδεδομένο φαινόμενο, καθώς οι ημέρες αυτές χαρακτηρίζονται από αυξημένη ή ελαττωμένη κυκλοφορία, αλλαγή συνηθειών οδήγησης και, πολλές φορές, μειωμένο έλεγχο από τις αρχές. Για τον λόγο αυτό κρίθηκε αναγκαίο να δημιουργηθεί μία σύνθετη binary μεταβλητή, `is_holiday`, η οποία υποδεικνύει αν μια ημερολογιακή εγγραφή αντιστοιχεί σε ημέρα σταθερής ή κινητής αργίας.

Αρχικά, δημιουργήθηκε σύνολο σταθερών ημερομηνιών, όπως η Πρωτοχρονιά (1 Ιανουαρίου), τα Θεοφάνια (6 Ιανουαρίου), η 25η Μαρτίου (Εθνική Επέτειος), η Πρωτομαγιά (1 Μαΐου), η Κοίμηση της Θεοτόκου (15 Αυγούστου), η 28η Οκτωβρίου και τα Χριστούγεννα (24–25 Δεκεμβρίου). Αυτές οι ημερομηνίες συγκρίνονται με την ημερομηνία της παρατήρησης (`week_date` ή `month_date`) και παράγουν την αρχική τιμή του `is_holiday`.

Πέρα από αυτές, έγινε επίσης υπολογισμός και των κινητών αργιών όπως το Πάσχα. Για την εξαγωγή της ημερομηνίας του Πάσχα χρησιμοποιήθηκε αλγόριθμος υπολογισμού βάσει του εκκλησιαστικού ημερολογίου, από την οποία προκύπτουν οι ημερομηνίες της Καθαράς Δευτέρας, του Μεγάλου Σαββάτου, της Κυριακής του Πάσχα, της Δευτέρας του Πάσχα, του Αγίου Πνεύματος κ.ά. Ειδικές μέρες όπως το τριήμερο της Αποκριάς, η Μεγάλη Παρασκευή και άλλες μεταβλητές αργίες υπολογίστηκαν δυναμικά για κάθε έτος, και συγκρίθηκαν αντίστοιχα με την ημερομηνία της κάθε παρατήρησης.

4.4.3 WEEK / MONTH / YEAR LAGS

Επόμενο βήμα στη διαδικασία του Feature Engineering αποτέλεσε η κατασκευή μεταβλητών χρονικών καθυστερήσεων (`lags`), με στόχο την ενσωμάτωση της πληροφορίας του παρελθόντος στη μοντελοποίηση του φαινομένου των τροχαίων ατυχημάτων. Οι μεταβλητές αυτές δημιουργήθηκαν για τα βασικά δεδομένα των `accidents` και `fatalities`, και επέτρεψαν τη συστηματική ενσωμάτωση ιστορικών τιμών στο ίδιο πλαίσιο ανάλυσης, ώστε να ληφθούν υπόψη υστερημένες επιδράσεις που μπορεί να εκδηλώνονται μετά από κάποιο χρονικό διάστημα.

Πιο συγκεκριμένα, στα εβδομαδιαία δεδομένα δημιουργήθηκαν μεταβλητές καθυστέρησης από 1 έως και 52 εβδομάδες, προκειμένου να καταγραφούν τόσο οι

βραχυπρόθεσμες μεταβολές (όπως π.χ. από εβδομάδα σε εβδομάδα) όσο και πιο μακροπρόθεσμες συσχετίσεις, όπως επαναλαμβανόμενα εποχικά μοτίβα που εκτείνονται σε βάθος ενός έτους. Στα μηνιαία datasets κατασκευάστηκαν αντίστοιχα lags από 1 έως 24 μήνες, ώστε να εξεταστούν ενδεχόμενες περιοδικότητες σε ετήσια ή διετή βάση, ενώ στα ετήσια δεδομένα εισήχθησαν αντίστοιχα καθυστερήσεις από 1 έως 10 χρόνια.

4.4.4 WEEK/MONTH/YEAR ROLLING AVERAGES

Στη συνέχεια, έγινε δημιουργία των κυλιόμενων μέσων όρων (rolling averages), με σκοπό την ανάδειξη πιο καθαρών και σταθερών τάσεων στην εξέλιξη του φαινομένου μέσα στον χρόνο. Στο πλαίσιο αυτό, εφαρμόστηκαν κυλιόμενοι μέσοι όροι σε διάφορα χρονικά παράθυρα, ανάλογα με τη συχνότητα των δεδομένων. Για τα εβδομαδιαία δεδομένα, υπολογίστηκαν rolling averages διάρκειας 4, 7, 12, 24 και 52 εβδομάδων, καλύπτοντας έτσι τόσο μηνιαία όσο και ετήσια δυναμική. Αντίστοιχα, στα μηνιαία datasets υπολογίστηκαν μέσοι όροι 3, 6, 12 και 24 μηνών, ενώ στα ετήσια δεδομένα δημιουργήθηκαν rolling averages για 3, 5 και 10 έτη.

4.4.5 ΕΠΙΠΛΕΟΝ ΔΕΙΚΤΕΣ ΚΑΙ BINARY FLAGS ΓΙΑ ΝΟΜΟΘΕΣΙΕΣ

Για τη βαθύτερη κατανόηση της δυναμικής του φαινομένου, δημιουργήθηκαν επιπλέον δείκτες τόσο στα απλά όσο και στα συνδυασμένα weekly / monthly και yearly data frames που βασίζονται σε ποσοστιαίες μεταβολές και ρυθμούς αύξησης. Οι δείκτες αυτοί προσφέρουν μια πιο κανονικοποιημένη και συγκρίσιμη θεώρηση των ατυχημάτων και των θανάτων, ιδίως σε περιόδους όπου οι απόλυτες τιμές ενδέχεται να παρουσιάζουν ακραίες διακυμάνσεις λόγω εξωτερικών παραγόντων (π.χ. μέτρα ασφαλείας, καιρικές συνθήκες, κοινωνικοοικονομικές εξελίξεις).

Συγκεκριμένα, υπολογίστηκαν οι ρυθμοί αύξησης (Growth Rates) τόσο για τα ατυχήματα όσο και για τους θανάτους. Ο δείκτης AccidentGrowthRate ορίστηκε ως ο λόγος της διαφοράς των ατυχημάτων από την προηγούμενη παρατήρηση προς την προηγούμενη τιμή, δηλαδή $(\text{accidents} - \text{lag}(\text{accidents})) / \text{lag}(\text{accidents})$. Αντίστοιχα, ο FatalityGrowthRate υπολογίστηκε με την ίδια λογική, εφαρμόζοντας το ίδιο μοντέλο στις τιμές των θανατηφόρων ατυχημάτων.

Πέραν αυτών, δημιουργήθηκε και ο δείκτης θνησιμότητας ανά ατύχημα (FatalityRate), ο οποίος αποτελεί το πηλίκο fatalities / accidents για κάθε χρονική παρατήρηση. Ο συγκεκριμένος δείκτης επιτρέπει τη μέτρηση της σχετικής σοβαρότητας των ατυχημάτων και μπορεί να αναδείξει κρίσιμες μεταβολές στη φύση ή τη βαρύτητα των περιστατικών. Επιπλέον, συμπεριλήφθηκαν και οι δείκτες σχετικής μεταβολής AccidentChangeRate και FatalityChangeRate, οι οποίοι μετρούν την ποσοστιαία αλλαγή σε σχέση με τη χρονική καθυστέρηση κατά μία περίοδο (Lag1): $(\text{accidents} - \text{Accidents_Lag1}) / \text{Accidents_Lag1}$ και $(\text{fatalities} - \text{Fatalities_Lag1}) / \text{Fatalities_Lag1}$, αντίστοιχα.

Εκτός από τους αριθμητικούς μετασχηματισμούς, δημιουργήθηκαν και νέες μεταβλητές που αποτυπώνουν αναλογίες, συσχετίσεις και την παρουσία συγκεκριμένων πολιτικών παρεμβάσεων ή ιστορικών περιόδων. Συγκεκριμένα, κατασκευάστηκαν δείκτες όπως AccidentsPerCapita, FatalitiesPerCapita, AccidentsPerVehicle, AccidentsUnemployment και FatalitiesInflation, με στόχο την αποτύπωση της σχέσης των βασικών μεγεθών με τον πληθυσμό, τα διαθέσιμα οχήματα ή τις μακροοικονομικές συνθήκες.

Επιπλέον, εισήχθησαν dummy variables για κρίσιμες περιόδους που θεωρείται ότι επιδρούν άμεσα ή έμμεσα στη συχνότητα και τη βαρύτητα των τροχαίων. Η μεταβλητή economic_crisis σημειώνει την περίοδο 2009–2018 ως περίοδο βαθιάς οικονομικής ύφεσης, ενώ η covid καταγράφει τις εβδομάδες, μήνες και χρονιές μεταξύ Μαρτίου 2020 και Μαΐου 2022, όπου τέθηκαν σε ισχύ περιοριστικά μέτρα λόγω της πανδημίας του COVID-19. Παράλληλα, εισήχθη και η μεταβλητή new_highways, η οποία σηματοδοτεί την ολοκλήρωση βασικών έργων υποδομής μετά το 2017, που θεωρείται έτος έναρξης ενός νέου πλαισίου οδικής κυκλοφορίας με βελτιωμένο δίκτυο.

Επιπρόσθετα, δημιουργήθηκαν βασικά features που δείχνουν πότε εφαρμόστηκαν μέτρα του ΚΟΚ ή ευρωπαϊκοί κανονισμοί, ώστε να ληφθεί υπόψη η πιθανή επίδρασή τους στα τροχαία ατυχήματα. Πιο συγκεκριμένα, εισήχθησαν οι μεταβλητές speed_limits, seat_belt, helmet, alcohol, phone, licensing, kid_seat, και kteo, όλες με έτος αφετηρίας το 1999 (βάσει του Νόμου 2696/1999 ΦΕΚ 57/Α'/23.3.1999).

Παράλληλα, η υιοθέτηση ευρωπαϊκών τεχνολογικών κανονισμών (Κανονισμός ΕΚ 661/2009) λήφθηκε υπ' όψιν μέσω των μεταβλητών airbags και abs, οι οποίες τέθηκαν σε ισχύ από το 2009 και μετά.

Ακολούθησε επίσης η προσθήκη των dummy variables fines και point_system που αποτυπώνουν αυστηρότερες πολιτικές επιβολής προστίμων και συστήματα ποινών μέσω ΦΕΚ

50/A'/2.3.2007 και ΦΕΚ 623/B'/27.4.2007 αντίστοιχα. Τέλος, η μεταβλητή pedestrians εισήχθη με έτος αναφοράς το 2011, βάσει της ΥΑ 43600/2011 (ΦΕΚ 1457/B'/16.6.2011), η οποία εστίασε στην προστασία ευάλωτων χρηστών του οδικού δικτύου. Όλες οι παραπάνω μεταβλητές αποτυπώθηκαν με δυαδικές τιμές (0 ή 1) ανάλογα με το αν η ρύθμιση ήταν ενεργή ή όχι για κάθε έτος, επιτρέποντας έτσι τη μοντελοποίηση της παρουσίας ή απουσίας συγκεκριμένων πολιτικών παρεμβάσεων σε κάθε παρατήρηση.

4.4.6 CUMULATIVE FEATURES

Συμπληρωματικά, ιδιαίτερη βαρύτητα δόθηκε και στη δημιουργία αθροιστικών μεταβλητών (Cumulative Features), οι οποίες αποτυπώνουν τη συνολική επιβάρυνση του φαινομένου των τροχαίων ατυχημάτων και των θυμάτων τους. Οι μεταβλητές CumulativeAccidents και CumulativeFatalities υπολογίστηκαν με χρήση της σωρευτικής συνάρτησης cumsum(), προσφέροντας έτσι μια συνεχή ένδειξη της συσσώρευσης περιστατικών από την αρχή της κάθε χρονοσειράς έως και κάθε χρονικό σημείο παρατήρησης. Οι δείκτες αυτοί είναι ιδιαίτερα χρήσιμοι για αναλύσεις μακροχρόνιων τάσεων και σύγκρισης περιόδων, τόσο για λόγους περιγραφικής στατιστικής όσο και για ενίσχυση προβλεπτικών μοντέλων.

Επιπρόσθετα, εφαρμόστηκαν κινούμενα αθροίσματα (Rolling Sums) σε κυλίσμενα χρονικά παράθυρα, διαφοροποιημένα ανά τύπο δεδομένων. Για τα εβδομαδιαία δεδομένα χρησιμοποιήθηκαν παράθυρα 4, 8, 12 και 24 εβδομάδων, ενώ στα μηνιαία εφαρμόστηκαν αθροίσματα για 3, 6, 12 και 24 μήνες. Στα ετήσια δεδομένα αντίστοιχα κατασκευάστηκαν rolling sums για περιόδους 3, 5 και 10 ετών.

Αξιοποιώντας αυτές τις μεταβλητές, τα μοντέλα μηχανικής μάθησης μπορούν να "μάθουν" από τη χρονική δομή και τη συσσωρευτική συμπεριφορά των δεδομένων, προσφέροντας καλύτερη κατανόηση της επίδρασης του παρελθόντος στις μελλοντικές προβλέψεις.

4.4.7 ΑΦΑΙΡΕΣΗ ΜΕΤΑΒΛΗΤΩΝ ΜΕ MISSING VALUES ΚΑΙ INCOMPLETE DATA ENTITIES

Στο τελικό στάδιο της διαδικασίας του Feature Engineering, πραγματοποιήθηκε μια κρίσιμη αξιολόγηση της πληρότητας και της ποιότητας των μεταβλητών που προέρχονταν από εξωτερικές πηγές. Η αξιολόγηση αυτή κατέληξε στο ότι ορισμένες από τις μεταβλητές

παρουσίαζαν σημαντικό ποσοστό ελλειπόν τιμών (missing values), είτε λόγω της μη διαθεσιμότητας ιστορικών δεδομένων για ολόκληρη την περίοδο 1996 - 2022, είτε εξαιτίας ασυνέπειας στη συχνότητα καταγραφής από τους αντίστοιχους οργανισμούς.

Ειδικότερα, μεταβλητές όπως οι `environmental_taxes`, `petrol_cons`, `ng_cons` και `el_cons` (κατανάλωση ενέργειας και περιβαλλοντικοί φόροι), αλλά και περισσότερο εξειδικευμένοι δείκτες όπως `fatal_per_million_veh`, `gini_index` και οι τέσσερις δείκτες παραβάσεων (`speed_infr`, `drink_infr`, `belt_infr`, `helmet_infr`), εμφάνιζαν σημαντικές ασυνέχειες. Για πολλές από αυτές, τα δεδομένα ήταν διαθέσιμα μόνο για μεμονωμένα ή πρόσφατα έτη, κάτι που περιόριζε τη χρησιμότητά τους στην παρούσα εργασία.

Ακόμη και στην περίπτωση που υπήρχαν διαθέσιμες τιμές για ορισμένα έτη, η παρουσία NA τιμών στις ενδιάμεσες χρονιές θα απαιτούσε τη χρήση τεχνικών συμπλήρωσης όπως παρεμβολή ή στατιστική πρόβλεψη, κάτι που όπως και προαναφέρθηκε, θα εισήγαγε θόρυβο και μη αξιόπιστες τάσεις στο dataset. Επιπλέον, η μερική καταγραφή αυτών των μεταβλητών θα μπορούσε να επηρεάσει τη σταθερότητα και τη γενίκευση των αλγορίθμων πρόβλεψης που θα εκπαιδεύονταν με βάση τα δεδομένα. Γι αυτόν τον λόγο, η απόφαση ήταν η αφαίρεση αυτών των μεταβλητών από τα τελικά `dataframes`, προκειμένου να διασφαλιστεί η ομοιογένεια και η καθαρότητα των δεδομένων.

Κεφάλαιο 5 Exploratory Data Analysis (EDA)

Η παρούσα ενότητα εστιάζει στην εξερεύνηση των δεδομένων που προηγήθηκε της εφαρμογής των στατιστικών και μηχανικών μοντέλων πρόβλεψης. Στόχος της Exploratory Data Analysis (EDA) είναι η σε βάθος κατανόηση της δομής και της συμπεριφοράς των δεδομένων, η αναγνώριση βασικών τάσεων, εποχικών μοτίβων και ασυνέπειας, καθώς και η εξαγωγή χρήσιμων υποθέσεων. Μέσω διαγραμματικών απεικονίσεων, βασικών περιγραφικών στατιστικών και συσχετίσεων μεταξύ μεταβλητών, η EDA λειτουργεί ως απαραίτητο ενδιάμεσο στάδιο ανάμεσα στην προετοιμασία και την προγνωστική ανάλυση.

Για την ανάλυση χρησιμοποιήθηκαν τα τελικά datasets που δημιουργήθηκαν στο προηγούμενο στάδιο επεξεργασίας και εμπλουτισμού των δεδομένων: `acc.w.ind.y_mod` για εβδομαδιαίο επίπεδο, `acc.m.ind.y_mod` για μηνιαίο και `acc.y.ind.y_mod` για ετήσιο.

5.1 Περιγραφική Στατιστική

Η πρώτη φάση της EDA επικεντρώθηκε στον υπολογισμό περιγραφικών στατιστικών για όλες τις αριθμητικές μεταβλητές των τελικών datasets σε εβδομαδιαίο, μηνιαίο και ετήσιο επίπεδο (`acc.w.ind.y_mod`, `acc.m.ind.y_mod`, `acc.y.ind.y_mod`). Για κάθε μεταβλητή, πραγματοποιήθηκε πλήρης στατιστική ανάλυση, η οποία περιλάμβανε βασικά μέτρα όπως η διάμεσος, τα τεταρτημόρια, τα δεκατημόρια και οι μέσες τιμές, αλλά και μέτρα διασποράς και μεταβλητότητας όπως το εύρος, η διακύμανση, η τυπική απόκλιση, ο συντελεστής μεταβλητότητας (CV) και το ενδοτεταρτημοριακό εύρος (IQR). Τα παραπάνω συνοδεύτηκαν και από μέτρα σχηματικότητας όπως η ασυμμετρία (skewness) και η κύρτωση (kurtosis), ώστε να αποτυπωθεί η μορφή κατανομής κάθε μεταβλητής.

Προκειμένου να αυτοματοποιηθεί η εξαγωγή των περιγραφικών στατιστικών, σχεδιάστηκε και υλοποιήθηκε η συνάρτηση `compute_statistics`. Η συνάρτηση εντοπίζει πρώτα

όλα τα αριθμητικά πεδία μέσω της `select(where(is.numeric))` και στη συνέχεια υπολογίζει τα στατιστικά μεγέθη για κάθε στήλη. Η εν λόγω συνάρτηση παρουσιάζεται παρακάτω:

```
compute_statistics <- function(df) {
  numeric_df <- df %>% select(where(is.numeric))
  stats <- data.frame(
    Variable = colnames(numeric_df),
    Std_Dev = apply(numeric_df, 2, sd, na.rm = TRUE),
    Count = colSums(!is.na(numeric_df)),
    Length = nrow(numeric_df),
    Max = apply(numeric_df, 2, max, na.rm = TRUE),
    Min = apply(numeric_df, 2, min, na.rm = TRUE),
    Mode = apply(numeric_df, 2, function(x) {
      ux <- unique(na.omit(x))
      ux[which.max(tabulate(match(x, ux)))]
    }),
    Median = apply(numeric_df, 2, median, na.rm = TRUE),
    Range = apply(numeric_df, 2, function(x) max(x, na.rm = TRUE) - min(x, na.rm = TRUE)),
    Variance = apply(numeric_df, 2, var, na.rm = TRUE),
    IQR = apply(numeric_df, 2, IQR, na.rm = TRUE),
    CV = apply(numeric_df, 2, function(x) sd(x, na.rm = TRUE) / mean(x, na.rm = TRUE)),
    Q1 = apply(numeric_df, 2, function(x) quantile(x, 0.25, na.rm = TRUE)),
    Q2 = apply(numeric_df, 2, function(x) quantile(x, 0.50, na.rm = TRUE)),
    Q3 = apply(numeric_df, 2, function(x) quantile(x, 0.75, na.rm = TRUE)),
    Skewness = apply(numeric_df, 2, function(x) skewness(x, na.rm = TRUE)),
    Kurtosis = apply(numeric_df, 2, function(x) kurtosis(x, na.rm = TRUE)),
    p10 = apply(numeric_df, 2, function(x) quantile(x, 0.10, na.rm = TRUE)),
    p25 = apply(numeric_df, 2, function(x) quantile(x, 0.25, na.rm = TRUE)),
    p50 = apply(numeric_df, 2, function(x) quantile(x, 0.50, na.rm = TRUE)),
    p75 = apply(numeric_df, 2, function(x) quantile(x, 0.75, na.rm = TRUE)),
    p90 = apply(numeric_df, 2, function(x) quantile(x, 0.90, na.rm = TRUE))
  )
}
```

```
return(stats)
}
```

5.1.1 ΕΒΔΟΜΑΔΙΑΙΑ ΔΕΔΟΜΕΝΑ

Όσον αφορά τα αποτελέσματα των περιγραφικών στατιστικών, μπορούν να εξαχθούν ορισμένα βασικά συμπεράσματα. Στο εβδομαδιαίο dataset, η μεταβλητότητα (τυπική απόκλιση) των ατυχημάτων και των θανάτων είναι ιδιαίτερα υψηλή, γεγονός που υποδηλώνει έντονες διακυμάνσεις στην εβδομαδιαία συχνότητά τους. Πιο συγκεκριμένα, τα ατυχήματα παρουσιάζουν τυπική απόκλιση περίπου 122, με εύρος τιμών από 15 έως 699, ενώ οι θάνατοι έχουν μέγιστη τιμή 78 και ελάχιστη 1. Ο δείκτης ασυμμετρίας (Skewness) είναι θετικός και για τις δύο μεταβλητές, κάτι που φανερώνει την ύπαρξη skewed κατανομών προς υψηλότερες τιμές, με πιο έντονο το φαινόμενο στους θανάτους.

Επιπλέον, οι τιμές των ποσοστιαίων δεικτών όπως τα τεταρτημόρια (Q1, Q2, Q3) και τα δεκατημόρια (p10, p90) δείχνουν ότι η πλειοψηφία των εβδομαδιαίων παρατηρήσεων συγκεντρώνεται σε χαμηλότερες τιμές, αλλά υπάρχουν και αρκετές ακραίες παρατηρήσεις, ιδίως στα ατυχήματα. Ενδεικτικά, το 75^ο εκατοστημόριο (p75) για τα ατυχήματα είναι 316, ενώ το 90^ο ανεβαίνει στις 412, υποδεικνύοντας σημαντική συγκέντρωση υψηλών τιμών στο πάνω μέρος της κατανομής. Αντίστοιχα, οι θάνατοι έχουν p75 ίσο με 30 και p90 ίσο με 40, γεγονός που τονίζει την ανάγκη για κατάλληλη διαχείριση αυτών των outliers κατά την μοντελοποίηση.

Τέλος, οι κοινωνικοοικονομικές μεταβλητές, όπως το ΑΕΠ (total_gdp) και ο πληθυσμός, εμφανίζουν μικρότερη διασπορά σε σύγκριση με τις μεταβλητές των ατυχημάτων, καθώς είναι ετήσιες και επομένως πιο σταθερές ανά εβδομάδα. Αυτό επιβεβαιώνεται από τις σχετικά χαμηλές τιμές του Συντελεστή Μεταβλητότητας (CV), καθώς και από τα περιορισμένα εύρη τιμών.

5.1.2 ΜΗΝΙΑΙΑ ΔΕΔΟΜΕΝΑ

Όσον αφορά τα μηνιαία aggregated δεδομένα, διακρίνονται πάλι παρόμοια περιγραφικά στατιστικά. Ειδικά, οι μεταβλητές accidents και fatalities παρουσιάζουν υψηλή διασπορά με σταθερή παρουσία ακραίων τιμών, όπως δείχνουν οι σημαντικές διαφορές μεταξύ ελάχιστων και μέγιστων τιμών (312–2776 για ατυχήματα και 23–292 για θανάτους). Ο συντελεστής μεταβλητότητας (CV) υποδηλώνει επίσης έντονες διακυμάνσεις σε σχέση με τη μέση τιμή,

ιδίως για τα ατυχήματα. Η δεξιά ασυμμετρία (skewness) της κατανομής των ατυχημάτων και θανάτων είναι θετική, κάτι που όπως και πριν υποδεικνύει ότι υπάρχουν περιπτώσεις με πολύ υψηλές τιμές σε σχέση με τον μέσο όρο. Αντίστοιχα, η τιμή της κυρτότητας (kurtosis) δείχνει ελαφρώς πλατύκαμπες κατανομές, ενισχύοντας την παρουσία outliers.

5.1.3 ΕΤΗΣΙΑ ΔΕΔΟΜΕΝΑ

Στα ετήσια δεδομένα αντίστοιχα, οι χρονοσειρές εμφανίζουν ακόμα μεγαλύτερες συγκεντρώσεις και εξάρσεις, όπως αναμενόταν λόγω του υψηλότερου επιπέδου ομαδοποίησης των δεδομένων. Οι μεταβλητές accidents και fatalities καταγράφουν πολύ μεγάλες διακυμάνσεις (9083–24819 για ατυχήματα, 584–2182 για θανάτους), επιβεβαιώνοντας την ετερογένεια μεταξύ των ετών.

Οι οικονομικοί δείκτες, όπως το total_gdp και το gdp_per_capita, διατηρούνται σχετικά σταθεροί με πιο μικρές αποκλίσεις, κάτι που αναμενόταν για ετήσια δεδομένα. Ωστόσο, η μεταβλητότητα της ανεργίας είναι αισθητή (range = 942.4), στοιχείο που σχετίζεται με την περίοδο της οικονομικής κρίσης. Οι τιμές skewness και kurtosis για τις περισσότερες μεταβλητές δείχνουν τάση για συμμετρικές κατανομές, αλλά με παρουσία λίγων εξαιρέσεων, όπως για παράδειγμα οι ακραίες τιμές στα ατυχήματα.

5.2 Επαγωγική Στατιστική

Για να ενισχυθεί η στατιστική εγκυρότητα της ανάλυσης και να ελεγχθούν ορισμένες υποθέσεις σχετικά με τη συμπεριφορά των τροχαίων ατυχημάτων και των σχετιζόμενων παραγόντων, πραγματοποιήθηκε ένα σύνολο δοκιμών επαγωγικής στατιστικής (inferential statistics) και στα τρία βασικά datasets: acc.w.ind.y_mod, acc.m.ind.y_mod και acc.y.ind.y_mod. Οι αναλύσεις αυτές στόχευσαν στο να εντοπίσουν σημαντικές διαφορές μεταξύ ομάδων, στατιστικά σημαντικές σχέσεις, καθώς και την κατανομή και τις ιδιότητες βασικών μεταβλητών.

Πιο συγκεκριμένα, δοκιμές όπως οι Shapiro-Wilk και Kolmogorov-Smirnov εφαρμόστηκαν για να εκτιμηθεί αν οι αριθμητικές μεταβλητές ακολουθούν κανονική κατανομή. Επιπλέον, έγιναν δοκιμές t-test, τόσο για σύγκριση με έναν θεωρητικό μέσο όρο (π.χ. για τα εβδομαδιαία ατυχήματα έναντι μιας τιμής αναφοράς), όσο και για συγκρίσεις μεταξύ δύο κατηγοριών (όπως is_summer), για την ανίχνευση εποχικών διαφορών. Στις

περιπτώσεις που οι προϋποθέσεις κανονικότητας δεν τηρούνταν, χρησιμοποιήθηκαν οι εναλλακτικές μη παραμετρικές μέθοδοι Wilcoxon και Kruskal-Wallis για την επιβεβαίωση των αποτελεσμάτων.

Τέλος, για τις πολυκατηγορικές μεταβλητές όπως η εποχικότητα (*season_numeric*), εφαρμόστηκαν αναλύσεις διασποράς (ANOVA) και ακολούθησαν Post-Hoc tests όπως το Tukey HSD και το Dunn Test για τον ακριβέστερο εντοπισμό ζευγών εποχών που παρουσιάζουν στατιστικά σημαντικές διαφορές. Επίσης, διεξήχθησαν δοκιμές ανεξαρτησίας όπως το Fisher's Exact Test και pairwise Wilcoxon tests με πολλαπλές συγκρίσεις, προσφέροντας έτσι μια πιο ολοκληρωμένη και αυστηρή στατιστική αποτίμηση της συμπεριφοράς των δεδομένων. Όπως και προαναφέρθηκε, οι δοκιμές αυτές εφαρμόστηκαν ισότιμα και στα τρία datasets, προκειμένου να εξαχθούν συγκρίσιμα και ερμηνεύσιμα συμπεράσματα.

5.2.1 ΕΒΔΟΜΑΔΙΑΙΑ ΔΕΔΟΜΕΝΑ

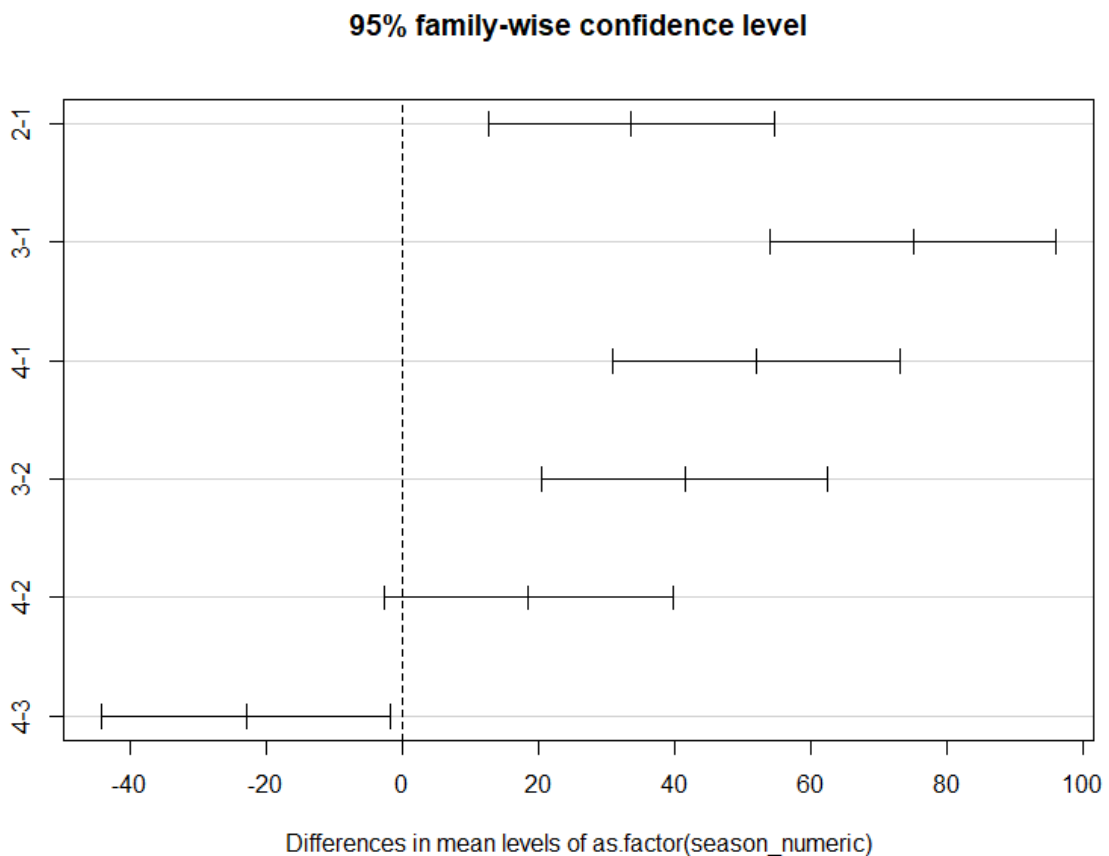
Τα αποτελέσματα της επαγωγικής στατιστικής για το εβδομαδιαίο dataset (*acc.w.ind.y_mod*) ανέδειξαν σημαντικά ευρήματα σχετικά με τη συμπεριφορά των βασικών μεταβλητών και τη συσχέτισή τους με εποχικούς και κοινωνικοοικονομικούς παράγοντες. Αρχικά, οι δοκιμές κανονικότητας Shapiro-Wilk και Kolmogorov-Smirnov κατέδειξαν ότι η πλειονότητα των αριθμητικών μεταβλητών αποκλίνει σημαντικά από την κανονική κατανομή, με εξαιρετικά χαμηλά p-values (συχνά κάτω από $1e-20$). Αυτό υποδεικνύει την παρουσία skewness, πιθανών outliers και μη συμμετρικών καμπυλών κατανομής.

Στο πλαίσιο συγκρίσεων μέσω τιμών, η δοκιμή t-test για τα εβδομαδιαία ατυχήματα αποκάλυψε ότι ο μέσος όρος των ατυχημάτων (247.4) διαφέρει στατιστικά σημαντικά από την τιμή αναφοράς των 200, με $p\text{-value} < 2.2e-16$. Η σύγκριση μέσω τιμών μεταξύ καλοκαιρινών και μη καλοκαιρινών εβδομάδων έδειξε επίσης ισχυρή διαφορά ($t = -6.23$, $p < 1e-9$), με τα καλοκαιρινά ατυχήματα να έχουν υψηλότερο μέσο όρο (282.5 έναντι 235.6), ενισχύοντας έτσι την υπόθεση για εποχική επίδραση στη συχνότητα των τροχαίων ατυχημάτων.

Η ανάλυση διασποράς (ANOVA) επιβεβαίωσε περαιτέρω την εποχική διαφοροποίηση, με εξαιρετικά σημαντική τιμή F (30.21, $p < 2e-16$) και στατιστικά σημαντικές διαφορές μεταξύ των εποχών. Τα αποτελέσματα των Tukey HSD και Dunn Test κατέδειξαν ότι οι μεγαλύτερες διαφορές παρατηρούνται μεταξύ Χειμώνα (1) και Άνοιξης/Καλοκαιριού (3), με όλα τα

αντίστοιχα ζεύγη να έχουν προσαρμοσμένα p-values κάτω από 0.001. Επιπλέον, το τεστ Kruskal-Wallis επιβεβαίωσε τα αποτελέσματα του ANOVA με σημαντικό χ^2 (74.29, $p < 5e-16$), αποδεικνύοντας ότι οι διαφορές στις κατανομές μεταξύ εποχών είναι στατιστικά σημαντικές, ακόμη και χωρίς την παραδοχή της κανονικότητας.

Επιπλέον, τα αποτελέσματα της μονοπαραγοντικής ανάλυσης διασποράς (ANOVA) για την επίδραση των αργιών (is_holiday) στον εβδομαδιαίο αριθμό τροχαίων ατυχημάτων αποκάλυψαν στατιστικά σημαντική διαφοροποίηση. Η ανάλυση έδειξε ότι η μεταβλητή is_holiday εξηγεί ένα μέρος της συνολικής διακύμανσης των ατυχημάτων, με $F = 10.84$ και $p\text{-value} = 0.00101$, που είναι σημαντικό με επίπεδο σημαντικότητας το 0.01. Αυτό σημαίνει ότι οι εβδομάδες που περιλαμβάνουν αργίες παρουσιάζουν σημαντικά διαφορετικό μέσο αριθμό ατυχημάτων σε σχέση με τις υπόλοιπες.



Σχήμα 5.2.1-1: Turkey HSD Αποτελέσματα για Σύγκριση Εβδομαδιαίων Τροχαίων Ατυχημάτων σε Σχέση με την Εποχή

Τέλος, οι μη παραμετρικές δοκιμές Wilcoxon και Fisher's Exact Test αποκάλυψαν σημαντικές συσχετίσεις μεταξύ των βασικών χαρακτηριστικών. Το Wilcoxon Test έδειξε σημαντική διαφορά μεταξύ των εβδομαδιαίων ατυχημάτων και θανάτων ($p < 2.2e-16$), ενώ το Fisher Test ανέδειξε ισχυρή εξάρτηση μεταξύ της παρουσίας καλοκαιρινής περιόδου και της εμφάνισης αργιών ($p = 3.25e-09$, odds ratio ≈ 0.13).

5.2.2 ΜΗΝΙΑΙΑ ΔΕΔΟΜΕΝΑ

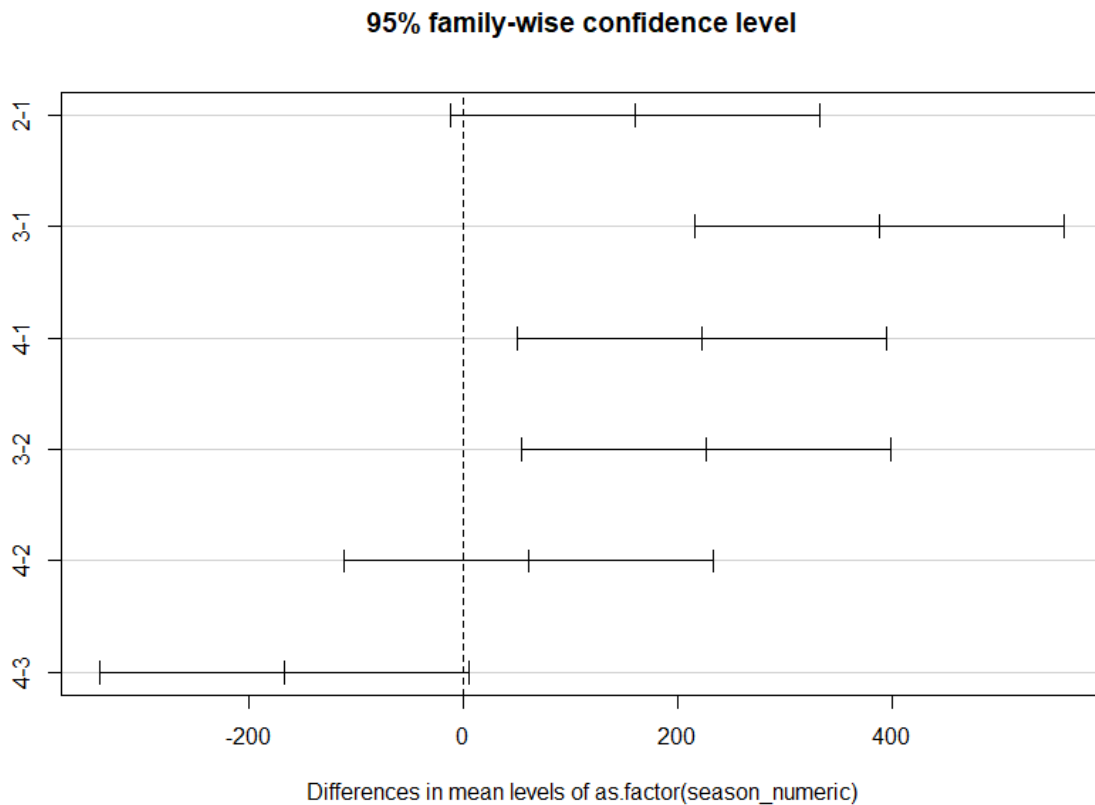
Η ανάλυση κανονικότητας μέσω των Test Shapiro-Wilk και Kolmogorov-Smirnov για το μηνιαίο dataset (acc.m.ind.y_mod) επιβεβαίωσε, όπως και στο εβδομαδιαίο δείγμα, την απόκλιση των περισσότερων μεταβλητών από την κανονική κατανομή. Συγκεκριμένα, σχεδόν όλες οι μεταβλητές παρουσίασαν p-values πολύ μικρότερες του 0.05, κάτι που καταδεικνύει ισχυρή απόρριψη της μηδενικής υπόθεσης περί κανονικότητας. Εξαίρεση αποτέλεσαν λίγες μεταβλητές όπως η FatalityRate, οι οποίες πλησίασαν ουδέτερες τιμές στο Kolmogorov-Smirnov, αλλά ακόμα και εκεί η απόδειξη κανονικότητας παρέμεινε ασθενής.

Η σύγκριση t-test μεταξύ θερινών και μη θερινών μηνών ανέδειξε στατιστικά σημαντική διαφορά, με τους θερινούς μήνες να εμφανίζουν σημαντικά υψηλότερο μέσο αριθμό ατυχημάτων (1480 έναντι 1220 περίπου, $p \approx 8e-5$). Τα αποτελέσματα αυτά επιβεβαιώνουν την εποχική επίδραση στο φαινόμενο.

Η ανάλυση διακύμανσης (ANOVA) επέτρεψε τη σύγκριση του μέσου αριθμού ατυχημάτων μεταξύ των τεσσάρων εποχών (season_numeric), αναδεικνύοντας ισχυρές στατιστικά διαφορές ($F = 11.61$, $p < 0.001$). Από τις πολλαπλές συγκρίσεις μέσω Tukey HSD, φάνηκε ότι το καλοκαίρι (εποχή 3) παρουσίασε σημαντικά υψηλότερες τιμές ατυχημάτων σε σχέση με τις άλλες εποχές, ειδικά σε σχέση με τον χειμώνα (season 1: +388 ατυχήματα, $p < 0.0001$). Αντίστοιχα και το φθινόπωρο παρουσίασε αυξημένα επίπεδα σε σχέση με τον χειμώνα. Το φαινόμενο αυτό ενισχύθηκε από τις μη παραμετρικές δοκιμές, όπως το Kruskal-Wallis test ($p \approx 2.3e-06$) και το Wilcoxon pairwise test, που συμφώνησαν ως προς την ύπαρξη στατιστικά σημαντικών διαφορών μεταξύ των εποχών.

Εκτός αυτού και όπως και προηγουμένως, η ανάλυση διασποράς (ANOVA) που πραγματοποιήθηκε για τη μεταβλητή season_numeric στο μηνιαίο dataset αποκάλυψε επίσης στατιστικά σημαντική επίδραση της εποχικότητας στον αριθμό των τροχαίων ατυχημάτων. Συγκεκριμένα, το τεστ έδωσε τιμή $F = 11.61$ με $p\text{-value} = 3.03e-07$, υποδεικνύοντας ότι

υπάρχει σημαντική διαφορά μεταξύ των μέσων τιμών ατυχημάτων στις τέσσερις εποχές του έτους. Η υψηλή στατιστική σημαντικότητα (***) στο 0.001) ενισχύει την υπόθεση ότι οι εποχές του έτους συνδέονται με διαφορετικά πρότυπα κινητικότητας και συνθηκών κυκλοφορίας (π.χ. καιρός, τουριστική περίοδος).



Σχήμα 5.2.2-1: Turkey HSD Αποτελέσματα για Σύγκριση Τροχαίων Μηνιαίων Ατυχημάτων σε Σχέση με την Εποχή

5.2.3 ΕΤΗΣΙΑ ΔΕΔΟΜΕΝΑ

Η αξιολόγηση της κανονικότητας μέσω των δοκιμών Shapiro-Wilk και Kolmogorov-Smirnov ανέδειξε σαφή απόκλιση της πλειοψηφίας των μεταβλητών από την κανονική κατανομή για τα ετήσια δεδομένα. Ειδικότερα, σχεδόν όλες οι αριθμητικές μεταβλητές όπως accidents, fatalities, unemployment, vehicles, καθώς και τα περισσότερα lags και rolling metrics (π.χ. RollingSum_12m_a) εμφάνισαν p-values πολύ μικρότερα του 0.05, απορρίπτοντας την υπόθεση κανονικότητας.

Επιπλέον, μέσω του Wilcoxon signed-rank test, πραγματοποιήθηκε σύγκριση μεταξύ των ετήσιων αριθμών ατυχημάτων και θανάτων. Το αποτέλεσμα ήταν εξαιρετικά σημαντικό ($p < 2.2e-16$), επιβεβαιώνοντας ότι υπάρχει συστηματική διαφορά μεταξύ των δύο μεταβλητών, κάτι αναμενόμενο δεδομένου ότι οι θάνατοι αποτελούν υποσύνολο των ατυχημάτων. Ωστόσο, η χρήση μη παραμετρικού τεστ εδώ ήταν απαραίτητη λόγω της μη κανονικότητας, διασφαλίζοντας έτσι την εγκυρότητα του αποτελέσματος.

5.3 Επιπλέον Στατιστικές Αναλύσεις και Plots

Για την πληρέστερη κατανόηση του φαινομένου των τροχαίων ατυχημάτων, κρίθηκε σκόπιμο να πραγματοποιηθούν επιπλέον στατιστικές αναλύσεις και Plots που επεκτείνουν την απλή περιγραφική και επαγωγική προσέγγιση. Οι αναλύσεις αυτές εστιάζουν στην εξερεύνηση συσχετίσεων, στην ανίχνευση μοτίβων και στην αποκάλυψη εσωτερικών δυναμικών των δεδομένων.

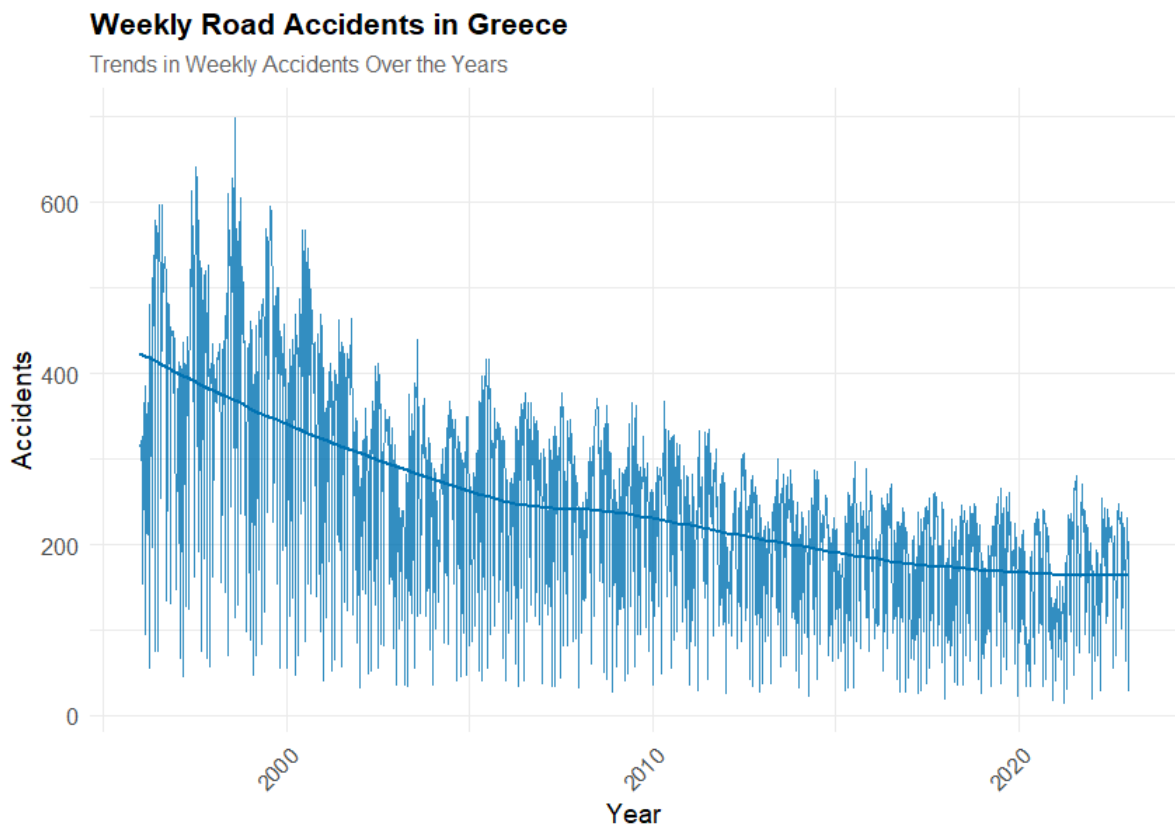
Παρόλο που πραγματοποιήθηκε ανάλυση και στα τρία χρονικά επίπεδα (εβδομαδιαίο, μηνιαίο και ετήσιο), στην παρούσα ενότητα δεν θα συμπεριληφθούν όλα τα διαγράμματα από κάθε επιμέρους dataset. Αντιθέτως, δίνεται έμφαση κυρίως στα εβδομαδιαία δεδομένα, καθώς αυτά προσφέρουν τη μεγαλύτερη χρονική ανάλυση και επιτρέπουν την πιο λεπτομερή διερεύνηση των διακυμάνσεων και των βραχυχρόνιων τάσεων των τροχαίων ατυχημάτων και των θανάτων.

5.3.1 TIME SERIES PLOTS

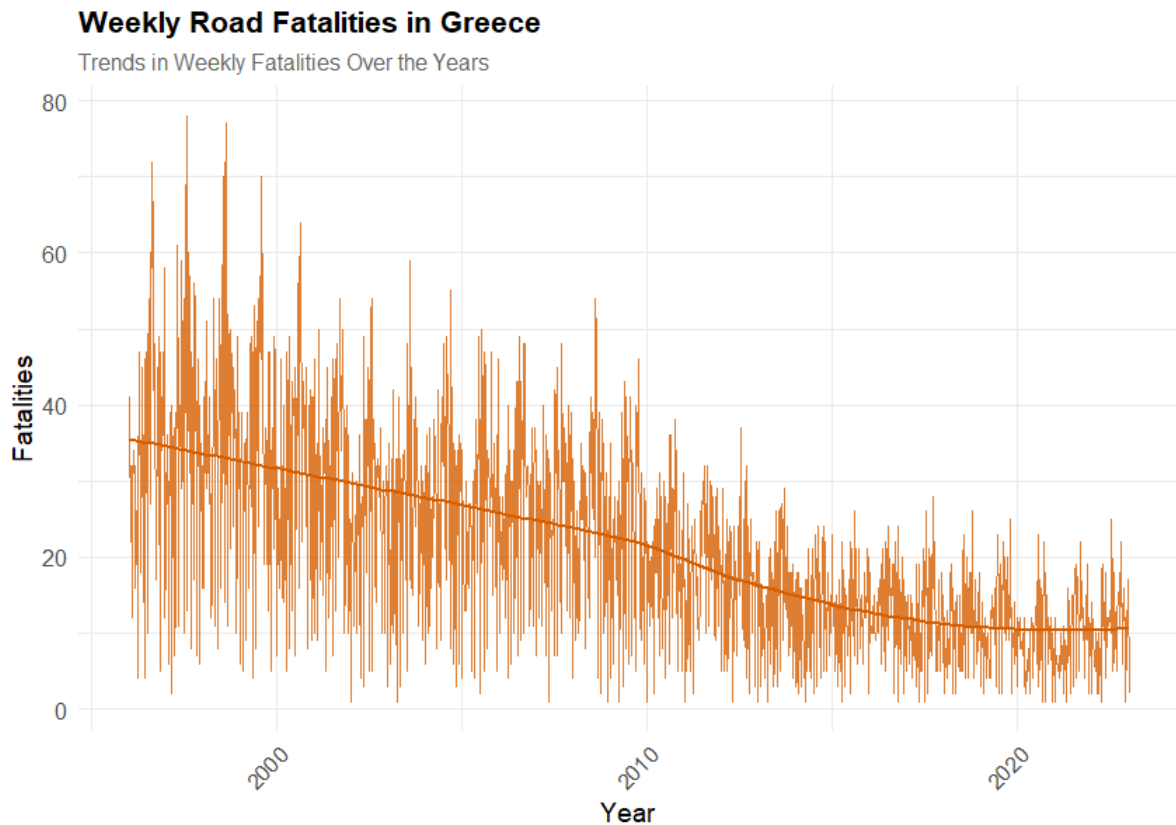
Η εξέλιξη των εβδομαδιαίων τροχαίων ατυχημάτων στην Ελλάδα από το 1996 έως το 2022 χαρακτηρίζεται από μια σταθερά καθοδική πορεία, η οποία αντικατοπτρίζει μια μακροχρόνια τάση μείωσης των συμβάντων. Παρά τις έντονες βραχυπρόθεσμες διακυμάνσεις που είναι αναμενόμενες λόγω της υψηλής χρονικής ανάλυσης, η συνολική εικόνα αποκαλύπτει σημαντική πρόοδο στον τομέα της οδικής ασφάλειας. Η πτωτική αυτή πορεία μπορεί να συνδέεται με παράγοντες όπως οι τεχνικές βελτιώσεις στις υποδομές, οι αυστηρότερες νομοθετικές ρυθμίσεις, η εντατικοποίηση των ελέγχων, αλλά και η σταδιακή ενίσχυση της οδηγικής συνείδησης. Σημειώνεται επίσης μια περίοδος σχετικής σταθεροποίησης των

επιπέδων ατυχημάτων μετά τα μέσα της δεκαετίας του 2010, που πιθανώς αντανακλά τον κορεσμό της επίδρασης ορισμένων πολιτικών ή την ανάγκη για νέες παρεμβάσεις.

Η πορεία των εβδομαδιαίων θανάτων από τροχαία ατυχήματα στην Ελλάδα κατά την περίοδο 1996–2022 εμφανίζει επίσης έντονα καθοδική τάση, αντανακλώντας μια συνεχή και σημαντική βελτίωση σε ζητήματα οδικής ασφάλειας. Παρά τις αυξομειώσεις που καταγράφονται σε εβδομαδιαίο επίπεδο λόγω συγκυριακών παραγόντων, το γενικό πρότυπο υποδηλώνει ότι οι θάνατοι έχουν μειωθεί περισσότερο από το ήμισυ μέσα σε περίπου δύο δεκαετίες. Η σταθερή αυτή μείωση είναι πιθανό να σχετίζεται με την πρόοδο στην ιατρική φροντίδα, την ενίσχυση των παρεμβάσεων πρόληψης (όπως ζώνες ασφαλείας και κράνη), καθώς και με την αυστηροποίηση των ελέγχων για παραβάσεις υψηλού κινδύνου, όπως η οδήγηση υπό την επήρεια αλκοόλ. Ιδιαίτερα μετά το 2010, οι θάνατοι φαίνεται να σταθεροποιούνται σε χαμηλότερα επίπεδα, γεγονός που δείχνει ότι, όπως και για τα ατυχήματα, οι συστημικές πολιτικές είχαν διαρκή αποτελέσματα στον περιορισμό των πιο σοβαρών συνεπειών.

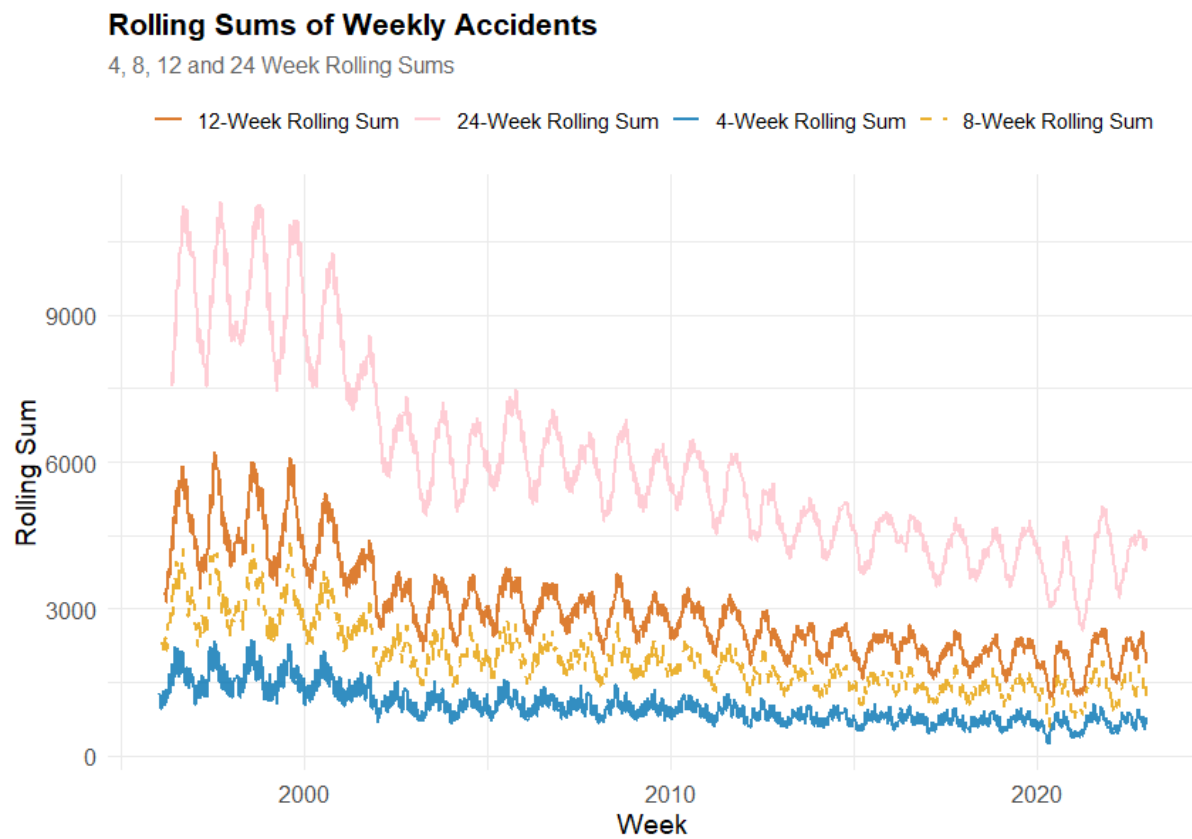


Σχήμα 5.3.1-1: Χρονοσειρά Τροχαίων Ατυχημάτων στην Ελλάδα 1996 – 2022

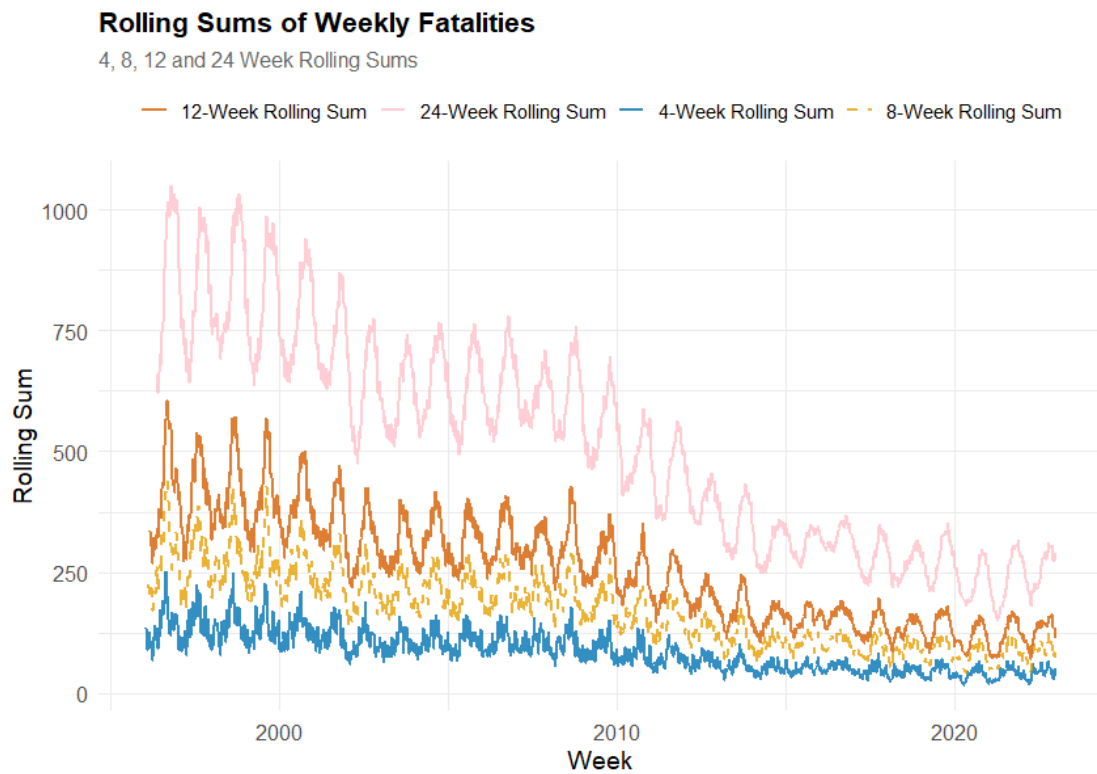


Σχήμα 5.3.1-2: Χρονοσειρά Θανάτων από Τροχαία Ατυχήματα στην Ελλάδα 1996 – 2022

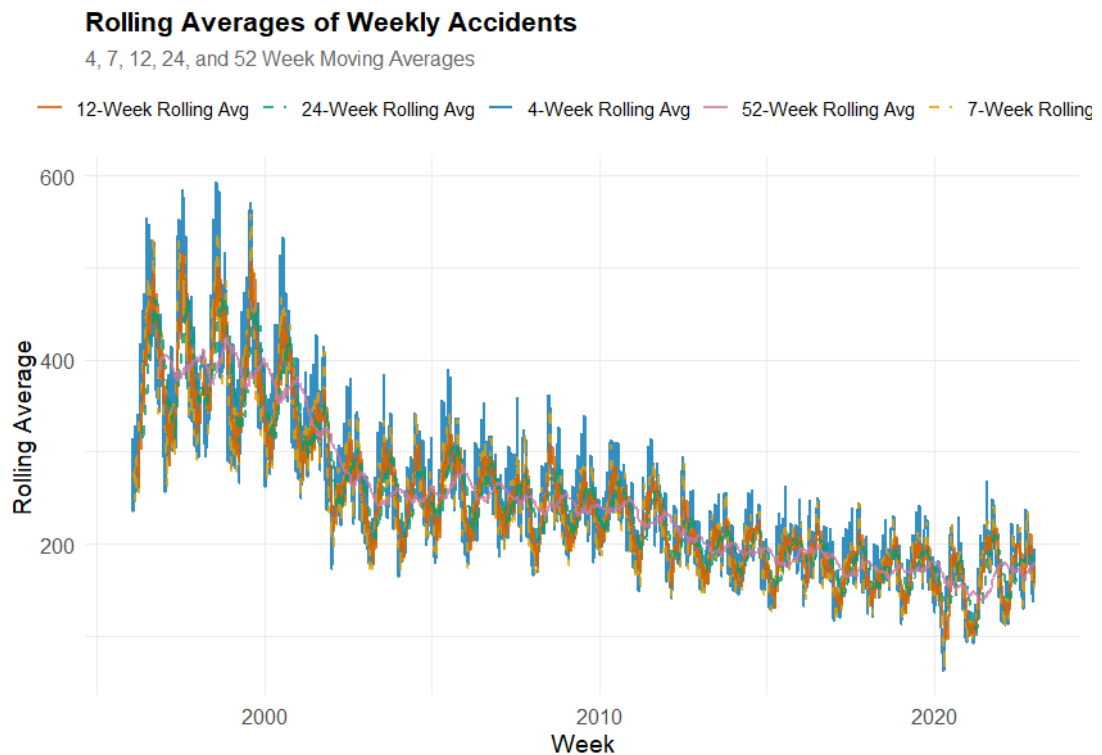
Συμπληρωματικά προς τα βασικά χρονοδιαγράμματα, τα διαγράμματα κυλιόμενων μέσων τιμών και κυλιόμενων αθροισμάτων προσφέρουν μια εξομαλυμένη οπτική της χρονικής εξέλιξης, διευκολύνοντας την ανίχνευση διαχρονικών τάσεων και περιοδικότητας. Οι κυλιόμενοι μέσοι όροι (Rolling Averages) βοηθούν στην εξάλειψη της θορυβώδους συμπεριφοράς των εβδομαδιαίων δεδομένων, αποτυπώνοντας με μεγαλύτερη καθαρότητα τη γενική κατεύθυνση της πορείας τόσο των ατυχημάτων όσο και των θανάτων. Από την άλλη, τα κυλιόμενα αθροίσματα (Rolling Sums) υποδεικνύουν τη σωρευτική επιβάρυνση σε συγκεκριμένα χρονικά παράθυρα, γεγονός που επιτρέπει την εκτίμηση της έντασης του φαινομένου σε μεσοπρόθεσμη βάση.



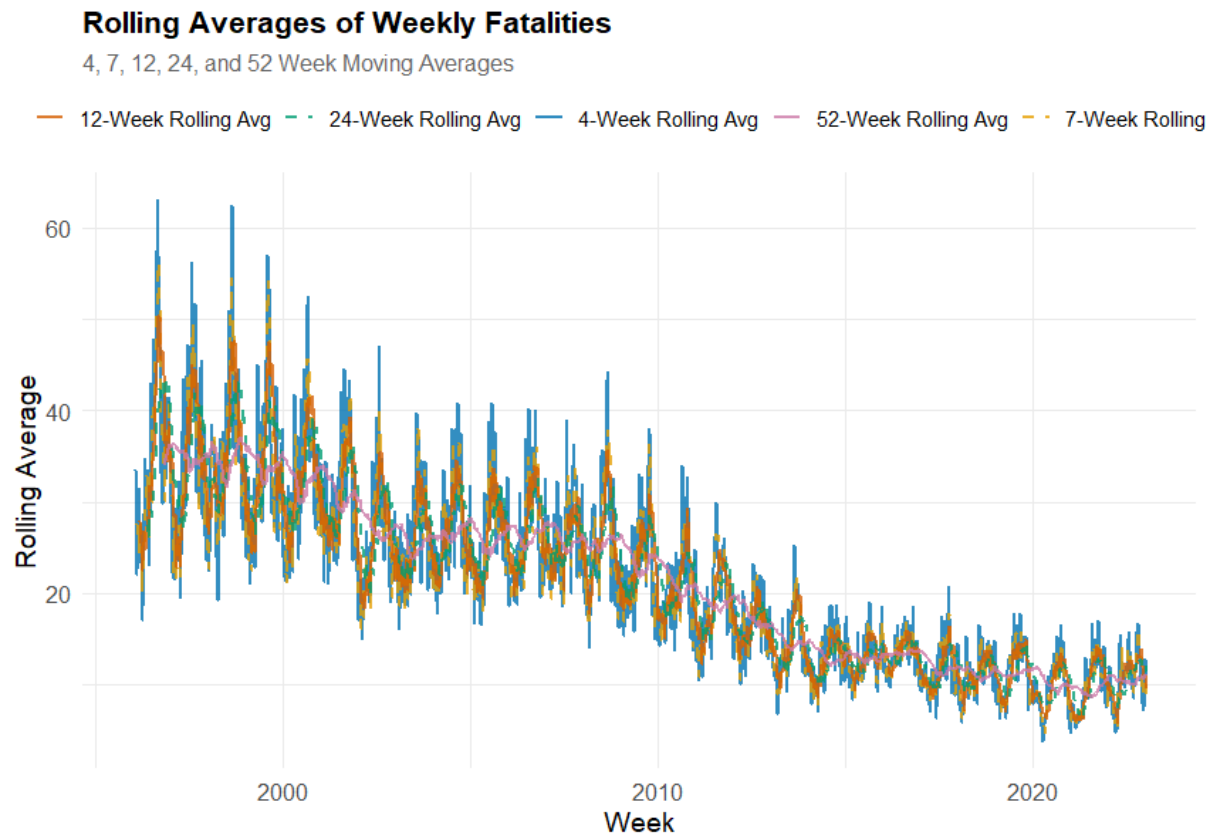
Σχήμα 5.3.1-3: Rolling Sums της Χρονοσειράς Εβδομαδιαίων Τροχαίων Ατυχημάτων



Σχήμα 5.3.1-4: Rolling Sums της Χρονοσειράς Εβδομαδιαίων Θανάτων από Τροχαία Ατυχήματα



Σχήμα 5.3.1-5: Rolling Averages της Χρονοσειράς Εβδομαδιαίων Τροχαίων Ατυχημάτων



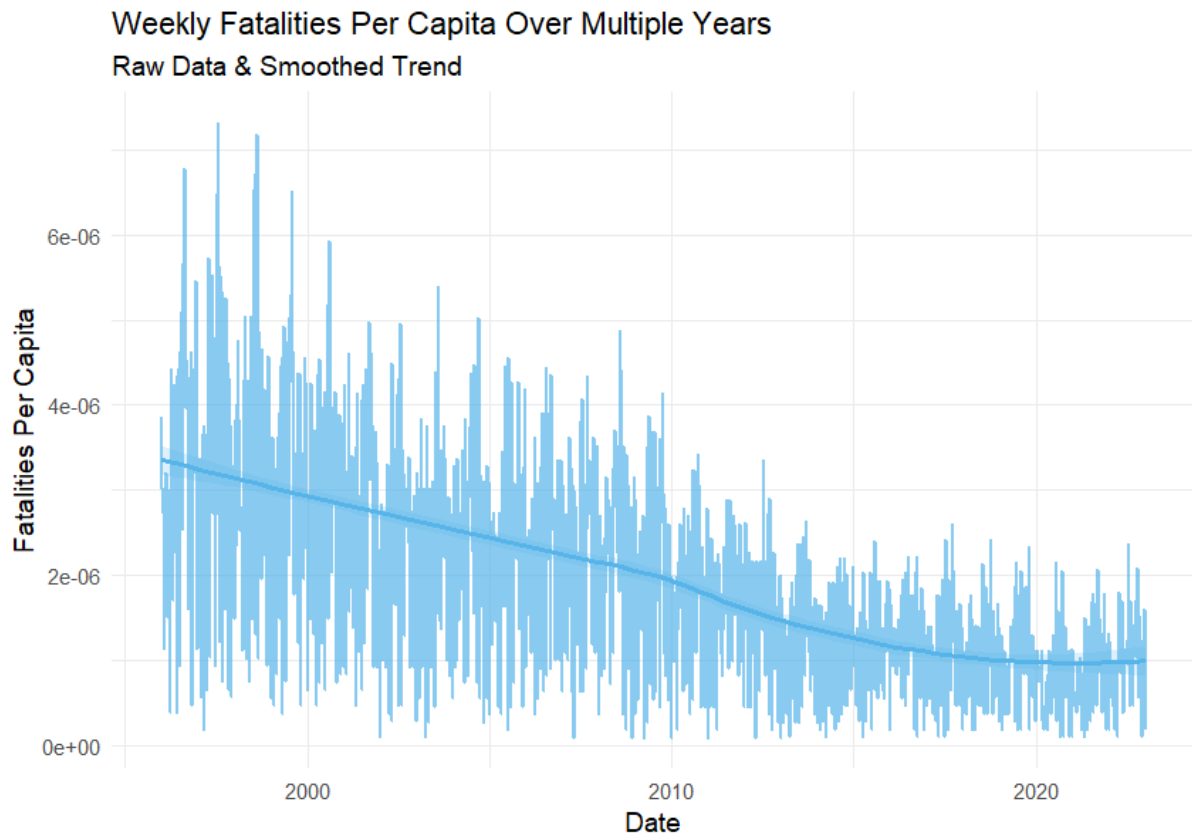
Σχήμα 5.3.1-6: Rolling Averages της Χρονοσειράς Εβδομαδιαίων Θανάτων από Τροχαία Ατυχήματα

Η ανάλυση των εβδομαδιαίων θανάτων ανά κάτοικο (per capita) προσφέρει μια πιο αντικειμενική αποτίμηση της σοβαρότητας του φαινομένου, λαμβάνοντας υπόψη τις πληθυσμιακές μεταβολές της χώρας μέσα στο χρόνο. Το γράφημα καταδεικνύει μια σταθερή πτωτική πορεία του δείκτη αυτού, με ορισμένες παροδικές αυξομειώσεις, κυρίως κατά τις πρώτες δεκαετίες της περιόδου.

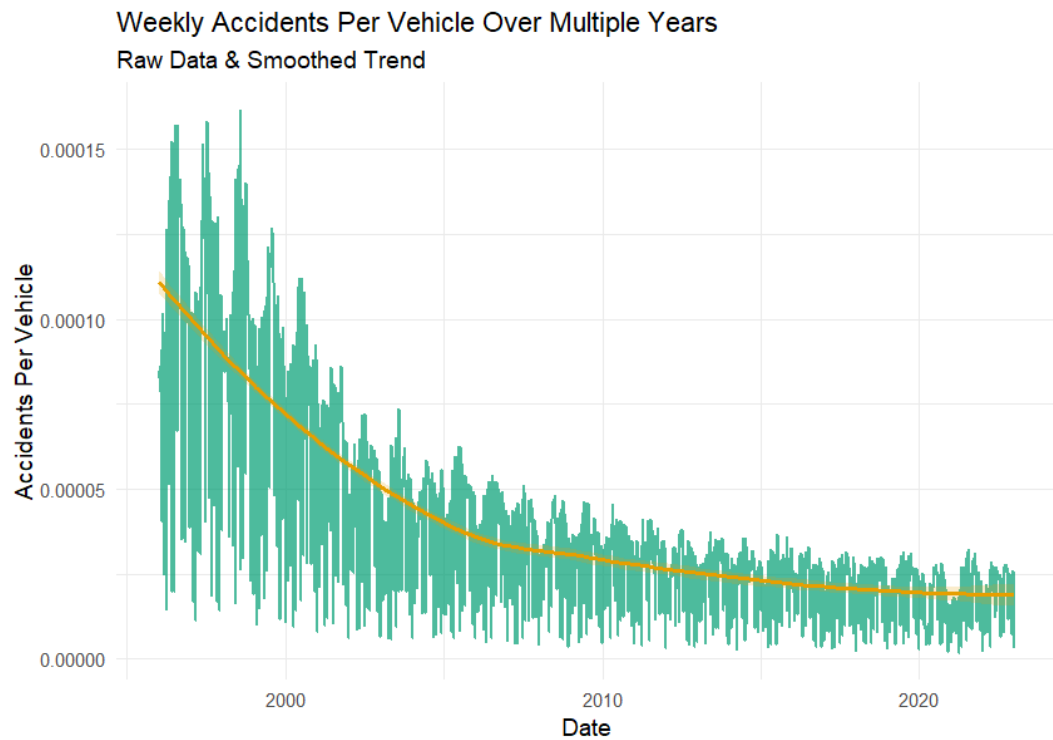
Αντίστοιχα, η ανάλυση των εβδομαδιαίων ατυχημάτων ανά όχημα επιτρέπει την αποσύνδεση της αυξημένης κινητικότητας ή της διεύρυνσης του στόλου οχημάτων από την πραγματική επίδοση σε επίπεδο οδικής ασφάλειας. Το γράφημα παρουσιάζει μια έντονα καθοδική πορεία του δείκτη, ιδιαίτερα κατά τις πρώτες δύο δεκαετίες, γεγονός που φανερώνει ότι, παρά την αύξηση του αριθμού των κυκλοφορούντων οχημάτων, ο σχετικός κίνδυνος εμπλοκής σε ατύχημα έχει μειωθεί σημαντικά.

Τέλος, ο δείκτης ποσοστού θνησιμότητας (fatality rate) αναλογεί τον αριθμό θανάτων προς τον συνολικό αριθμό ατυχημάτων και λειτουργεί ως δείκτης σοβαρότητας των

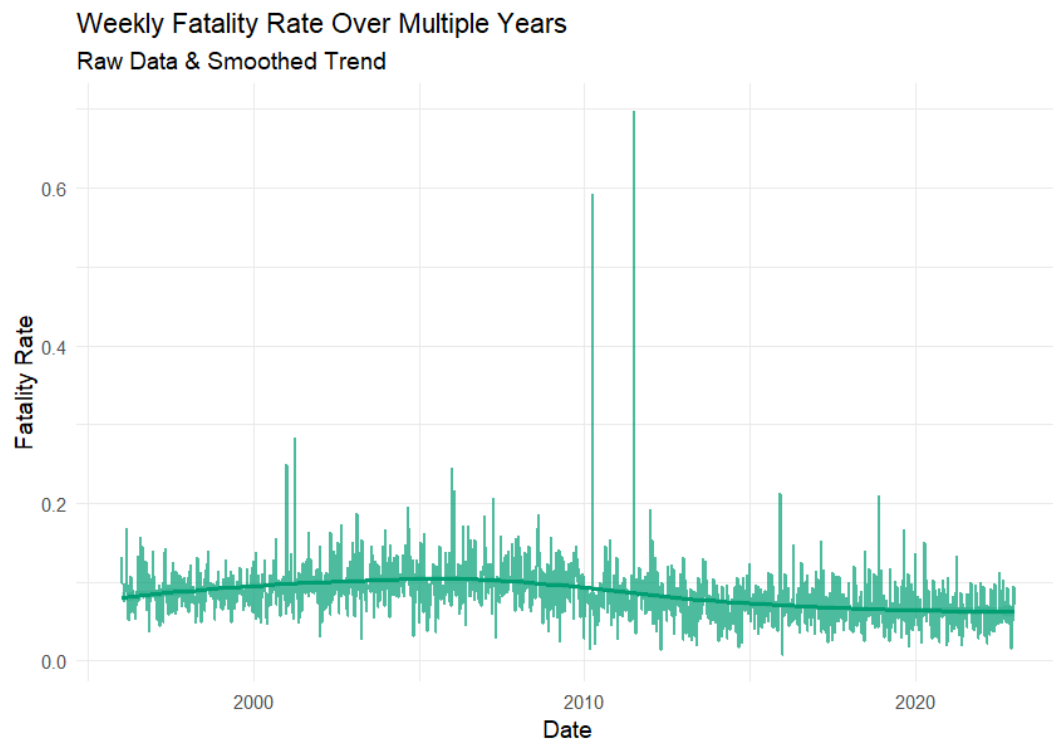
περιστατικών. Η παρατήρηση της σχετικά σταθερής πορείας του δείκτη με ελαφρώς φθίνουσα τάση υποδηλώνει ότι, παρόλο που τα ατυχήματα μειώνονται συνολικά, το ποσοστό εκείνων που είναι θανατηφόρα διατηρείται σε σχετικά σταθερό επίπεδο. Οι ακραίες τιμές που εμφανίζονται σε ορισμένα σημεία πιθανώς σχετίζονται με εβδομάδες που σημειώθηκαν λίγα ατυχήματα αλλά με σοβαρές συνέπειες.



Σχήμα 5.3.1-7: Time Series Plot για Θανάτους από Τροχαία Ατυχήματα ανά κάτοικο 1996 – 2022



Σχήμα 5.3.1-8: Time Series Plot για Τροχαία Ατυχήματα ανά Υπάρχοντα Οχήματα 1996 – 2022

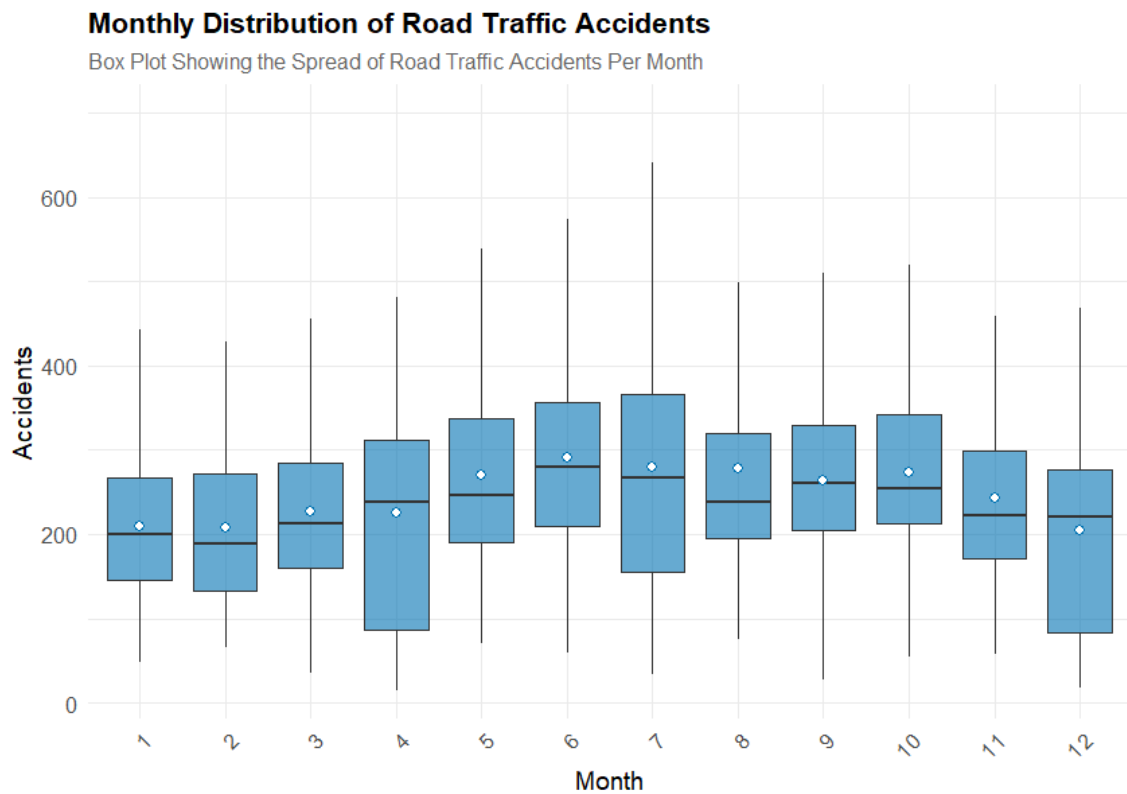


Σχήμα 5.3.1-9: Time Series Plot για Ποσοστό Θνησιμότητας των Τροχαίων Ατυχημάτων 1996 - 2022

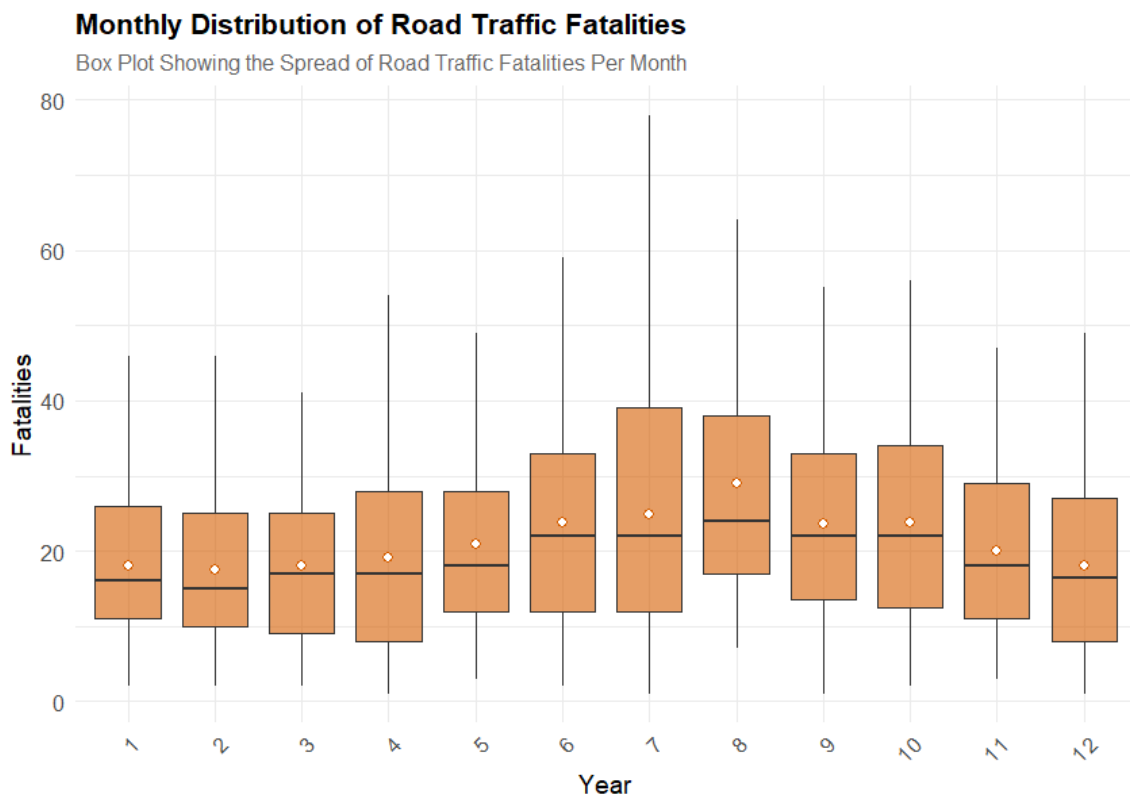
5.3.2 PLOTS ΕΠΙΠΛΕΟΝ ΠΕΡΙΓΡΑΦΙΚΩΝ ΣΤΑΤΙΣΤΙΚΩΝ

Οι κατανομές των ατυχημάτων ανά μήνα αποτυπώνονται μέσω boxplot, προσφέροντας μια περιγραφική εικόνα της εποχικότητας και της διασποράς των δεδομένων. Παρατηρείται ότι οι μέσες τιμές (λευκές κουκίδες) και οι διάμεσοι των ατυχημάτων αυξάνονται σταδιακά από τον Ιανουάριο έως τον Ιούλιο, φτάνοντας σε κορύφωση το καλοκαίρι, και στη συνέχεια μειώνονται προς το τέλος του έτους. Η μεταβλητότητα είναι επίσης αυξημένη κατά τους θερινούς μήνες, κάτι που πιθανόν σχετίζεται με την έντονη κινητικότητα λόγω διακοπών, τουρισμού και μεγαλύτερης έκθεσης στον δρόμο. Οι χαμηλότερες τιμές τον χειμώνα υποδηλώνουν περιορισμένη κυκλοφορία ή αυξημένη προσοχή λόγω δυσμενών καιρικών συνθηκών.

Αντίστοιχα, το boxplot της μηνιαίας κατανομής των θανάτων παρουσιάζει συγκρίσιμες εποχικές τάσεις. Οι μήνες από Ιούνιο έως Σεπτέμβριο καταγράφουν υψηλότερους μέσους όρους και ευρύτερη διασπορά στους θανάτους, γεγονός που ενισχύει την υπόθεση αυξημένου κινδύνου κατά τη θερινή περίοδο. Επιπλέον, οι ανώτερες τιμές (outliers) σε αυτούς τους μήνες υποδηλώνουν την ύπαρξη εβδομάδων με ακραία υψηλό αριθμό θανάτων, που πιθανώς οφείλονται σε σοβαρά δυστυχήματα ή σε αργίες με έντονη κυκλοφοριακή δραστηριότητα. Αντίθετα, οι χειμερινοί μήνες παρουσιάζουν μικρότερες διακυμάνσεις, στοιχείο που υποδηλώνει μια πιο σταθερή και προβλέψιμη δυναμική στα δεδομένα θανάτων από τροχαία ατυχήματα.



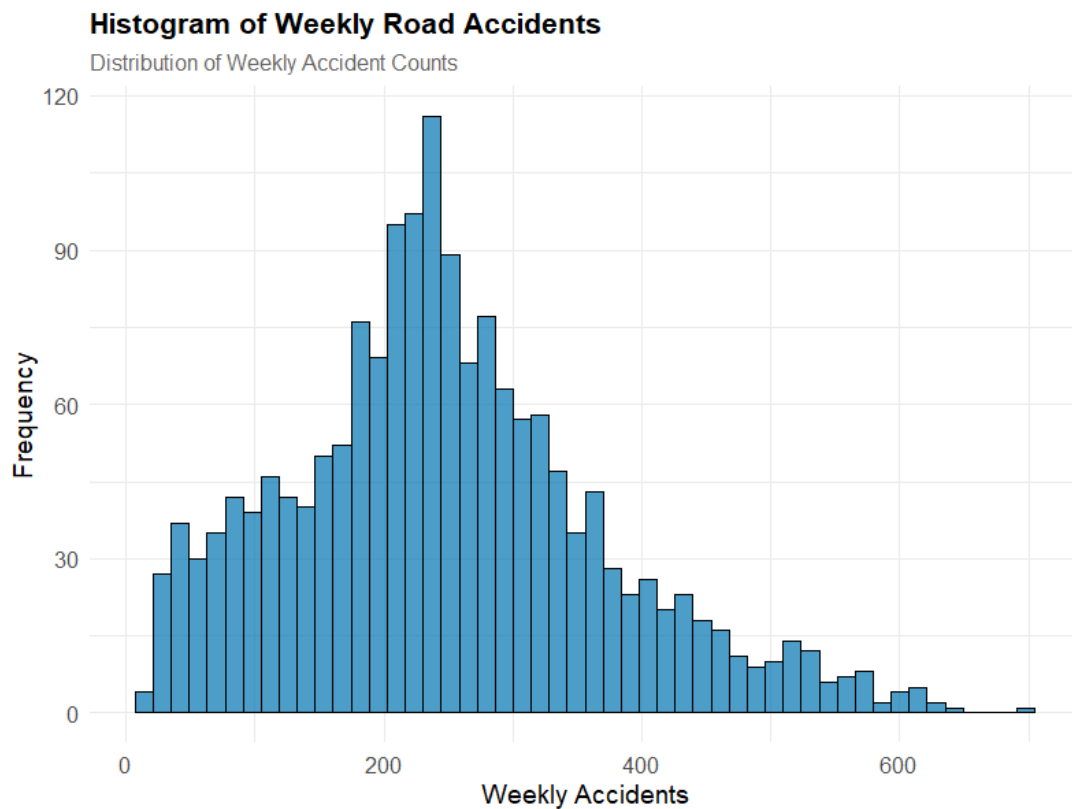
Σχήμα 5.3.2-1: Boxplot Κατανομής Τιμών Τροχαίων Ατυχημάτων ανά Μήνα



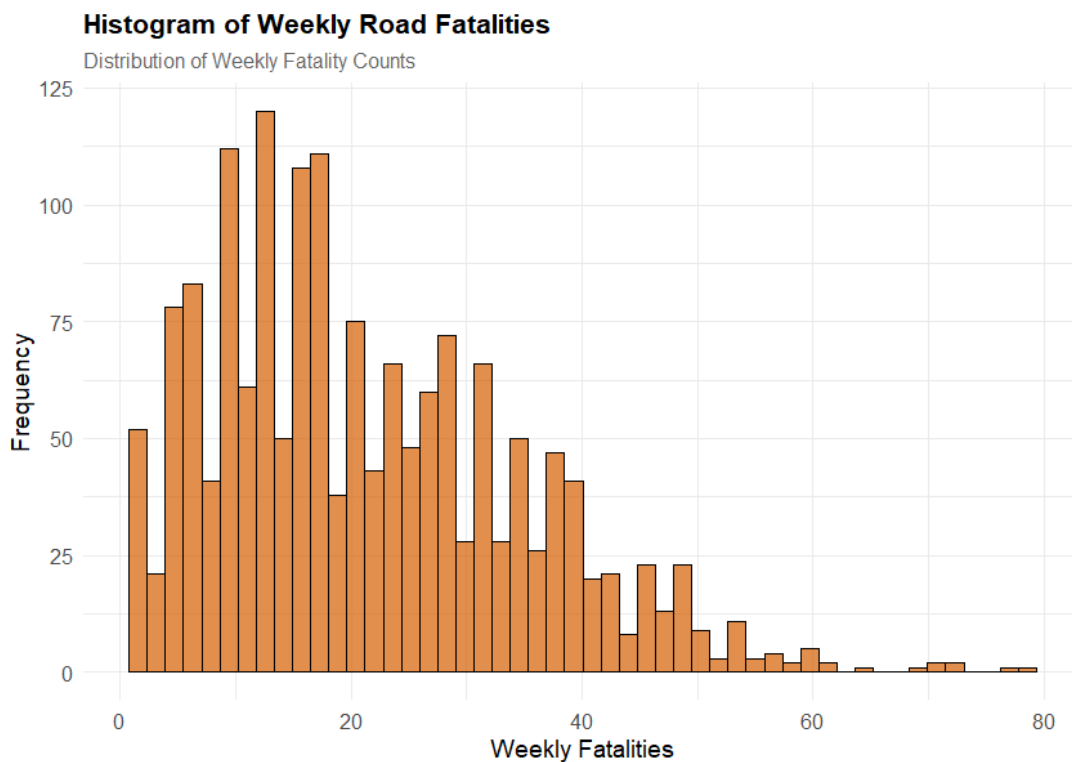
Σχήμα 5.3.2-2: Boxplot Κατανομής Τιμών Θανάτων από Τροχαία Ατυχήματα ανά Μήνα

Η κατανομή των εβδομαδιαίων τροχαίων ατυχημάτων και θανάτων αναδεικνύει κρίσιμες ιδιότητες της δυναμικής του φαινομένου, που δεν είναι εύκολα ορατές σε χρονοσειρές. Το ιστόγραμμα των εβδομαδιαίων ατυχημάτων δείχνει μια κατανομή με έντονη ασυμμετρία προς τα δεξιά, με συχνότερες τιμές να εντοπίζονται γύρω από τα 200 με 250 ατυχήματα ανά εβδομάδα. Υπάρχει όμως και μια ουρά υψηλών τιμών που φτάνουν έως και τα 600–700 ατυχήματα, γεγονός που υποδηλώνει την ύπαρξη περιόδων με εξαιρετικά αυξημένη επιβάρυνση. Αυτή η μορφή κατανομής δεν είναι τυχαία και πιθανώς συνδέεται με αιτίες, όπως η εποχικότητα και μεμονωμένα γεγονότα ή ακραίες καιρικές συνθήκες. Παράλληλα, το μεγάλο πλήθος εβδομάδων με μεσαία ή χαμηλά επίπεδα κατανομής επιβεβαιώνει την πτωτική μακροχρόνια τάση που παρατηρήθηκε και στα γραφήματα χρονοσειρών.

Αντίστοιχα, η κατανομή των εβδομαδιαίων θανάτων από τροχαία παρουσιάζει μεγαλύτερη συγκέντρωση σε χαμηλές τιμές, με τις περισσότερες εβδομάδες να κυμαίνονται κάτω από τους 30 θανάτους και σημαντικό ποσοστό ακόμα χαμηλότερα, κάτω από τους 15. Η κατανομή αυτή είναι επίσης έντονα ασύμμετρη προς τα δεξιά, όμως οι ακραίες τιμές (άνω των 50 θανάτων) είναι πολύ πιο σπάνιες συγκριτικά με τα ατυχήματα. Αυτό αποκαλύπτει ότι η θνησιμότητα έχει μειωθεί σημαντικά, παρά την επιμονή του φαινομένου των ατυχημάτων ως προς τη συχνότητα. Ταυτόχρονα, το γεγονός ότι η πλειονότητα των εβδομάδων βρίσκεται σε περιοχές χαμηλής επικινδυνότητας ενισχύει την υπόθεση ότι οι παρεμβάσεις στον τομέα της οδικής ασφάλειας είναι αποτελεσματικές, ειδικά ως προς την αποτροπή των θανατηφόρων συνεπειών.

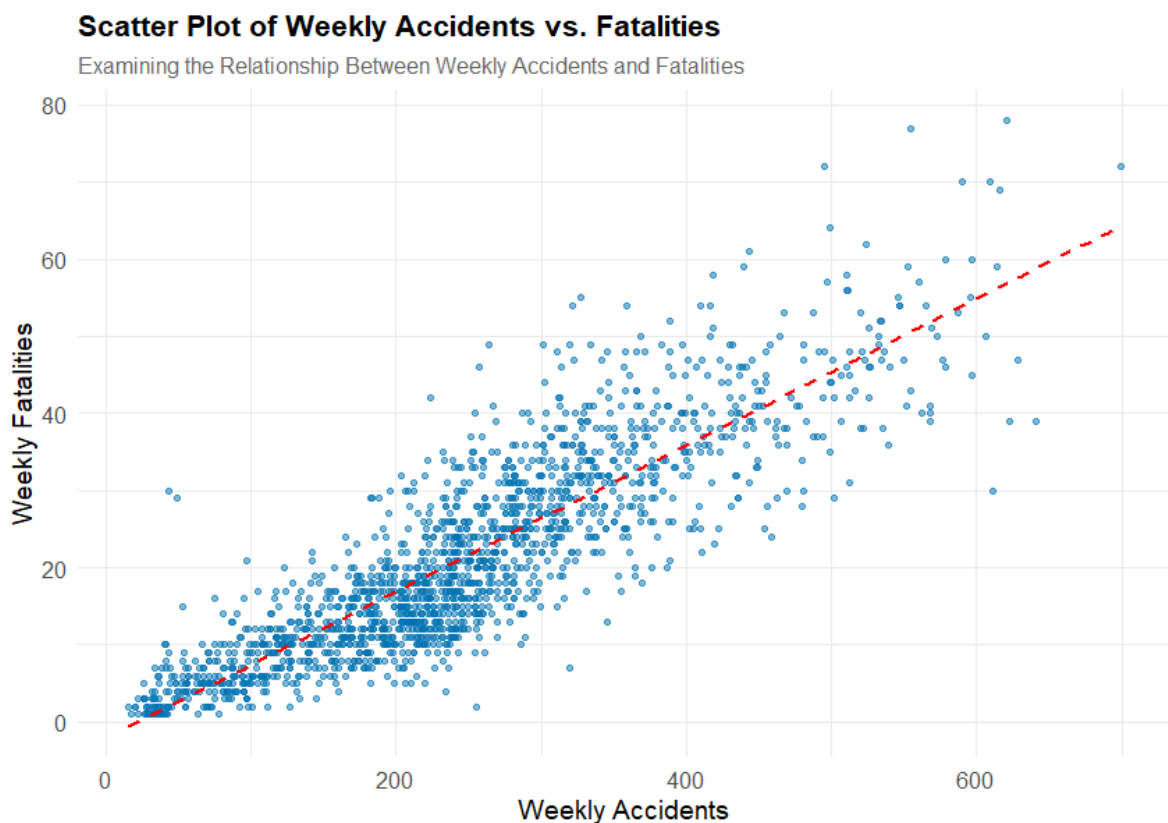


Σχήμα 5.3.2-3: Histogram Κατανομής Τιμών Τροχαίων Ατυχημάτων



Σχήμα 5.3.2-4: Histogram Κατανομής Τιμών Θανάτων από Τροχαία Ατυχήματα

Η συσχέτιση μεταξύ των εβδομαδιαίων ατυχημάτων και των θανάτων από τροχαία καθίσταται εμφανής μέσω της έντονης συγκέντρωσης σημείων κατά μήκος της ανοδικής ευθείας στο διάγραμμα διασποράς. Η ύπαρξη γραμμικής σχέσης υποδηλώνει ότι, σε γενικές γραμμές, οι εβδομάδες με αυξημένο αριθμό ατυχημάτων τείνουν να συνοδεύονται και από υψηλότερο αριθμό θανάτων. Η γραμμή παλινδρόμησης καταδεικνύει αυτή την τάση, ενισχύοντας την ερμηνεία πως τα δύο φαινόμενα είναι αλληλένδετα. Παρόλα αυτά, παρατηρούνται σημαντικές διακυμάνσεις γύρω από τη γραμμή, οι οποίες φανερώνουν ότι δεν εξαρτώνται όλοι οι θάνατοι μόνο από τον αριθμό των ατυχημάτων, αλλά και από παράγοντες όπως η σοβαρότητα των περιστατικών, οι συνθήκες του οδικού δικτύου και ίσως η άμεση ιατρική ανταπόκριση. Η κατανομή των σημείων σε υψηλότερες τιμές ατυχημάτων με σχετικά χαμηλότερους θανάτους υποδηλώνει βελτιώσεις στην οδική ασφάλεια ή στην αποτελεσματικότητα αντιμετώπισης, ενώ τα αντίθετα σενάρια δείχνουν πιθανές ακραίες ή δυσμενείς καταστάσεις.

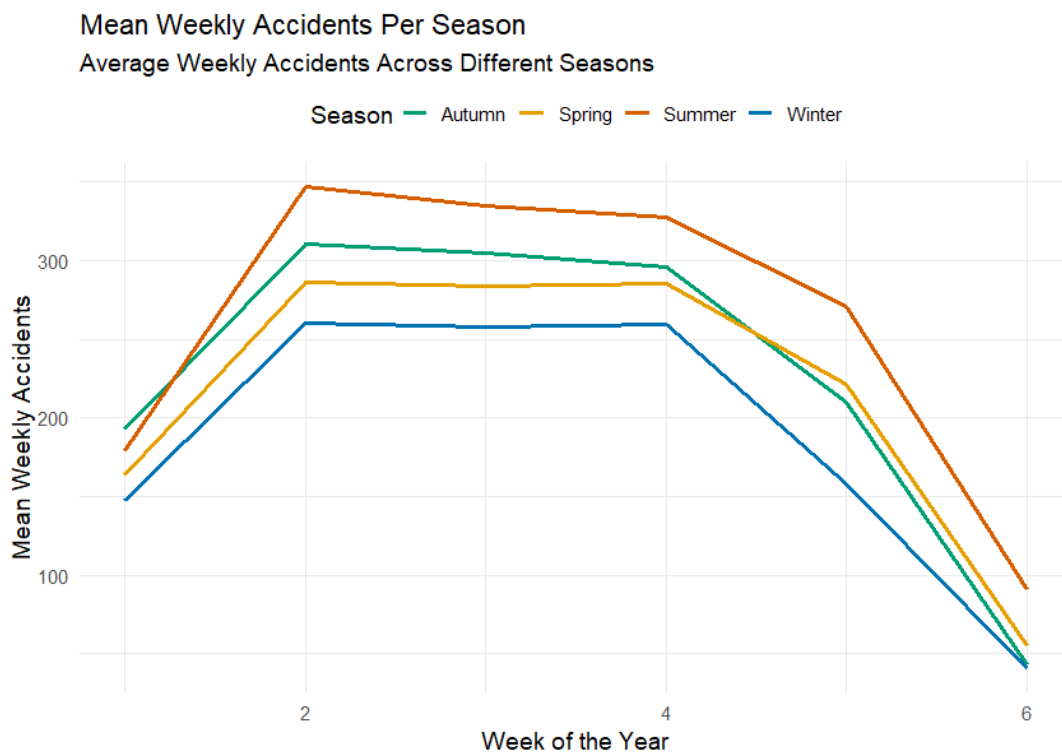


Σχήμα 5.3.2-5: Scatter Plot Τροχαίων Ατυχημάτων και Θανάτων

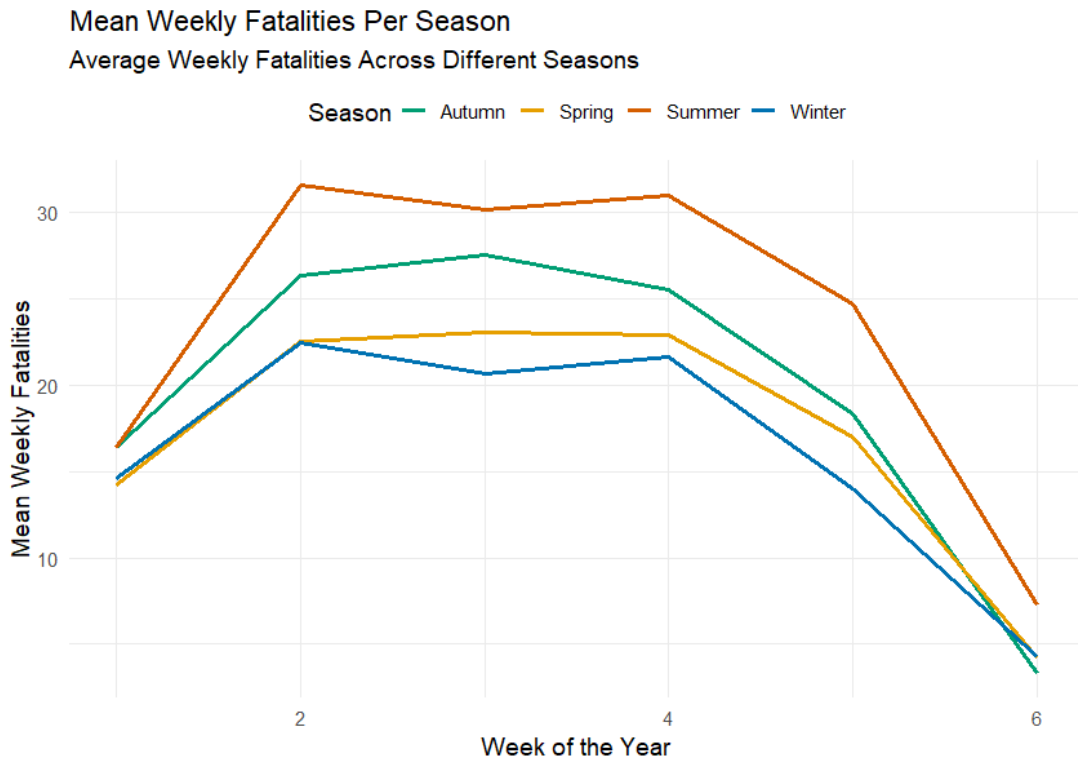
5.3.3 PLOTS ΑΤΥΧΗΜΑΤΩΝ ΣΕ ΣΥΣΧΕΤΙΣΗ ΜΕ ΤΙΣ ΕΠΟΧΕΣ ΚΑΙ ΤΙΣ ΑΡΓΙΕΣ

Η εποχική ανάλυση του μέσου αριθμού εβδομαδιαίων τροχαίων ατυχημάτων αναδεικνύει σαφείς διαφορές μεταξύ των τεσσάρων εποχών. Το καλοκαίρι καταγράφει σταθερά τα υψηλότερα επίπεδα, ιδιαίτερα κατά τις κεντρικές εβδομάδες της περιόδου, κάτι που ενδέχεται να οφείλεται στην αυξημένη κυκλοφορία λόγω διακοπών, τουρισμού και ταξιδιών. Η άνοιξη και το φθινόπωρο ακολουθούν με παρόμοια πρότυπα, ενώ ο χειμώνας εμφανίζει χαμηλότερους μέσους όρους, πιθανώς λόγω περιορισμένης κινητικότητας, κακών καιρικών συνθηκών και μειωμένων μετακινήσεων για αναψυχή.

Αντίστοιχα, στην εποχική κατανομή των μέσων εβδομαδιαίων θανάτων από τροχαία ατυχήματα φαίνεται επίσης μια κυριαρχία του καλοκαιριού, με τη μεγαλύτερη ένταση να εντοπίζεται στην τρίτη και τέταρτη εβδομάδα του κάθε μήνα. Οι τιμές είναι εμφανώς χαμηλότερες τον χειμώνα, όπως και στην περίπτωση των ατυχημάτων. Το φαινόμενο αυτό ενδέχεται να συνδέεται όχι μόνο με την ποσότητα των ατυχημάτων, αλλά και με τη σοβαρότητά τους, καθώς οι υψηλές ταχύτητες και η αυξημένη κυκλοφορία κατά τη θερινή περίοδο μπορεί να οδηγούν σε πιο θανατηφόρες συγκρούσεις.



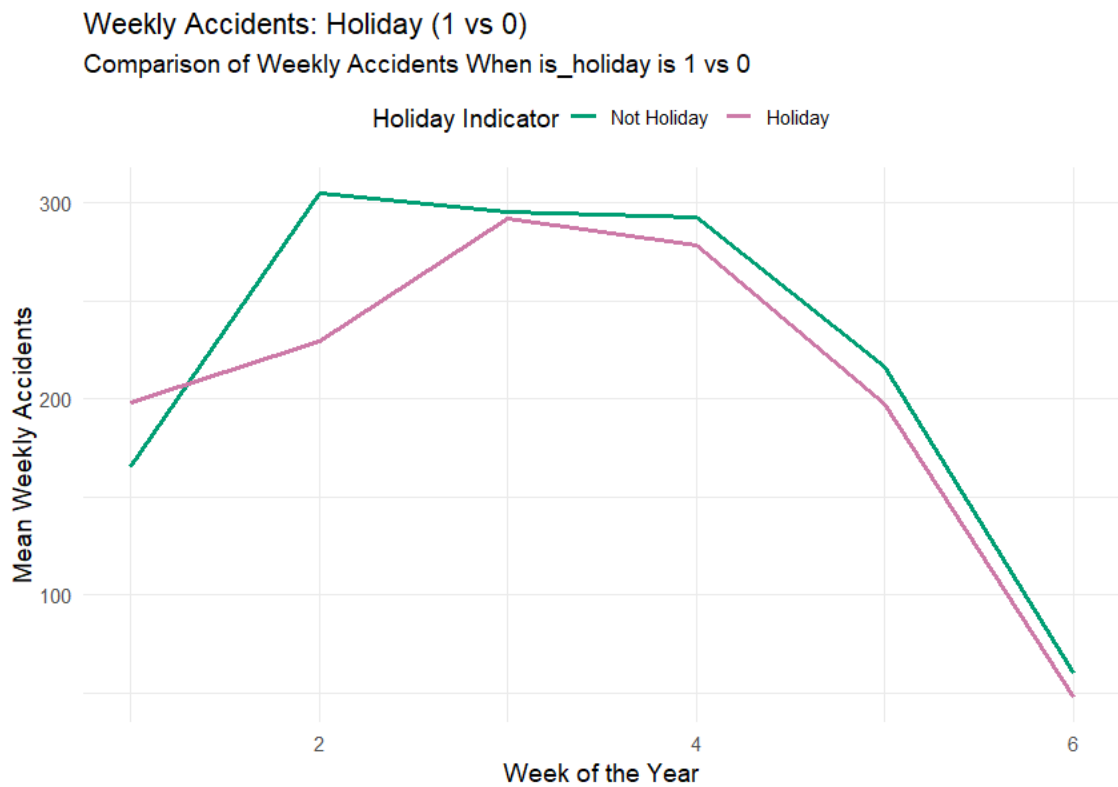
Σχήμα 5.3.3-1: Μέσος Όρος Τροχαίων Ατυχημάτων Ανα Εβδομάδα για Κάθε Εποχή



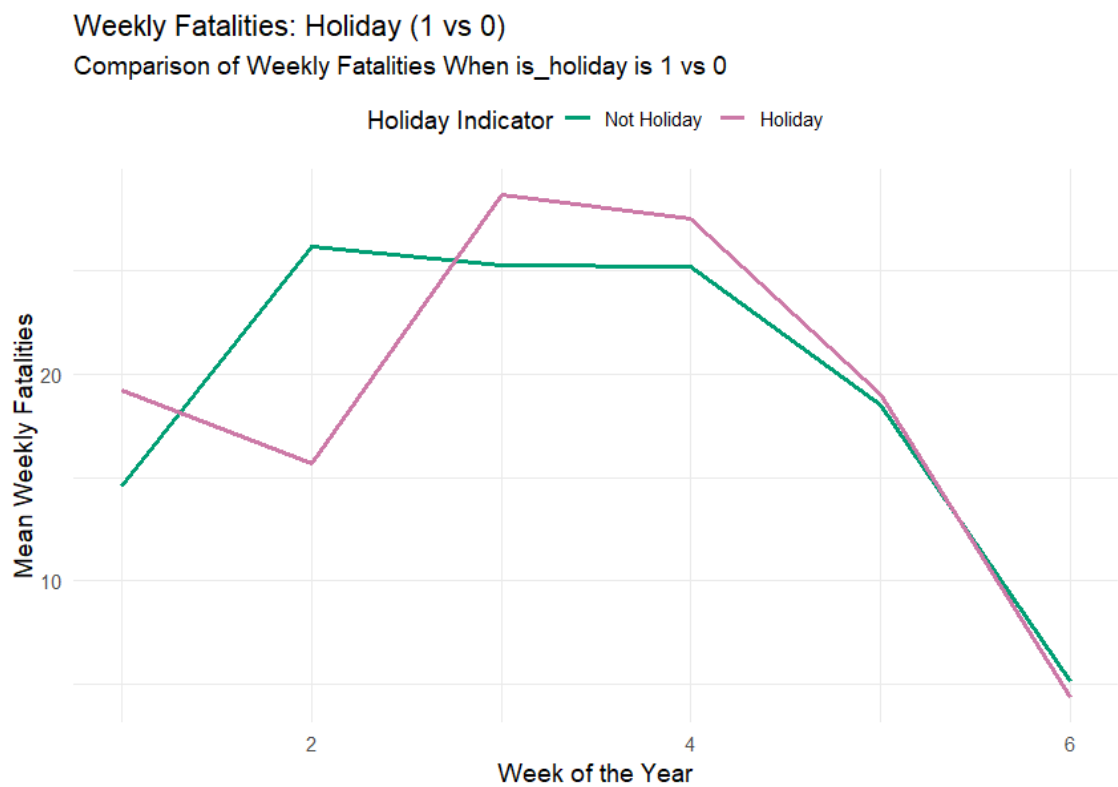
Σχήμα 5.3.3-2: Μέσος Όρος Θανάτων από Τροχαία Ατυχήματα Ανα Εβδομάδα για Κάθε Εποχή

Κατά την ανάλυση της επίδρασης των αργιών στη συχνότητα των τροχαίων ατυχημάτων, παρατηρείται ότι οι εβδομάδες που περιλαμβάνουν αργίες παρουσιάζουν οριακά μειωμένο αριθμό ατυχημάτων συγκριτικά με τις αντίστοιχες χωρίς αργίες, ιδίως στις πρώτες εβδομάδες του κάθε μήνα. Το μοτίβο αυτό μπορεί να αντανάκλα περιορισμένη μετακίνηση κατά τη διάρκεια των αργιών ή αλλαγές στις συνήθειες οδήγησης, όπως η αποφυγή μεγάλων διαδρομών ή η μείωση των μετακινήσεων εντός πόλης.

Ωστόσο, όσον αφορά τους θανάτους από τροχαία, το πρότυπο διαφέρει. Σε αρκετά σημεία, οι εβδομάδες με αργίες συνοδεύονται από αυξημένα επίπεδα θανατηφόρων περιστατικών σε σχέση με τις υπόλοιπες. Αυτό υποδηλώνει ότι, ενώ τα συνολικά ατυχήματα μπορεί να είναι λιγότερα, η σοβαρότητά τους κατά τις αργίες είναι μεγαλύτερη. Πιθανοί λόγοι για αυτό το φαινόμενο μπορεί να περιλαμβάνουν μεγαλύτερες ταχύτητες σε επαρχιακούς δρόμους και αυξημένη κατανάλωση αλκοόλ, που ενισχύουν τον κίνδυνο σοβαρών συνεπειών σε περίπτωση ατυχήματος.



Σχήμα 5.3.3-3: Τροχαία Ατυχήματα Όταν Υπάρχουν Αργίες / Διακοπές

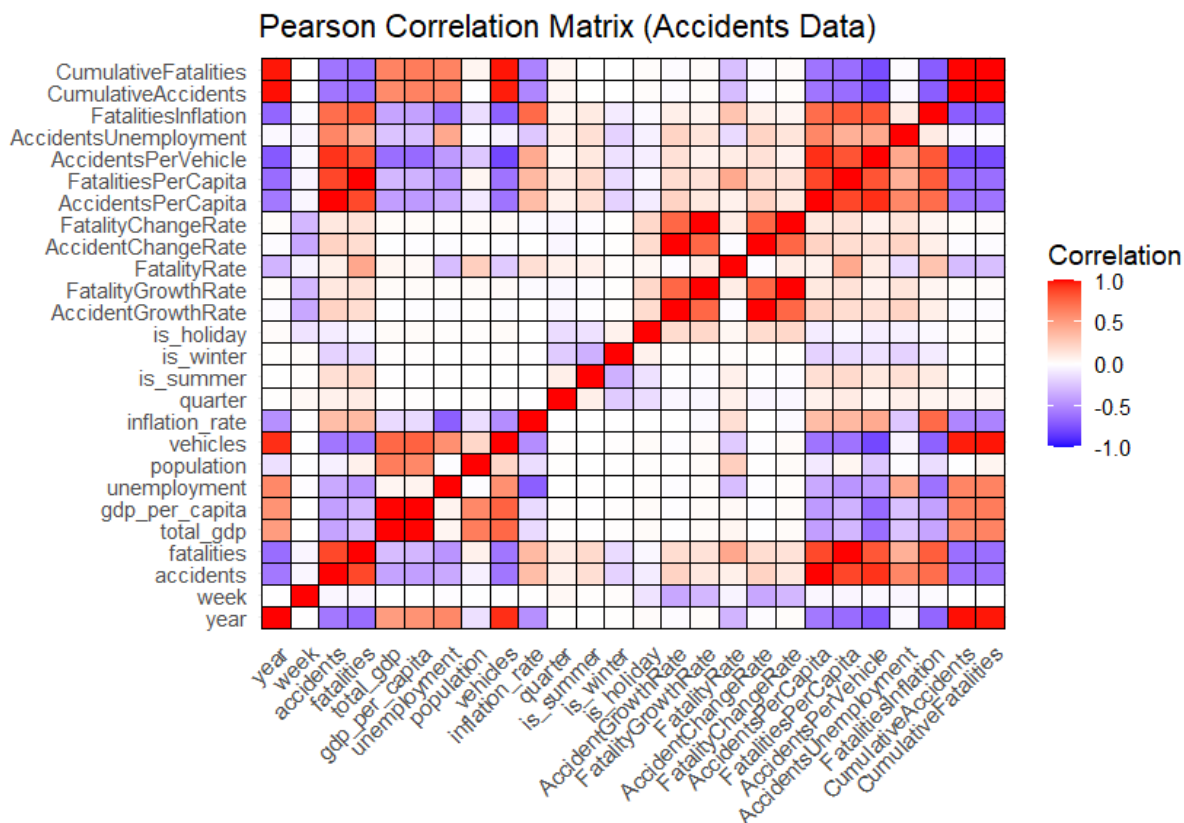


Σχήμα 5.3.3-4: Θάνατοι από Τροχαία Ατυχήματα Όταν Υπάρχουν Αργίες / Διακοπές

5.3.4 CORRELATION MATRICES

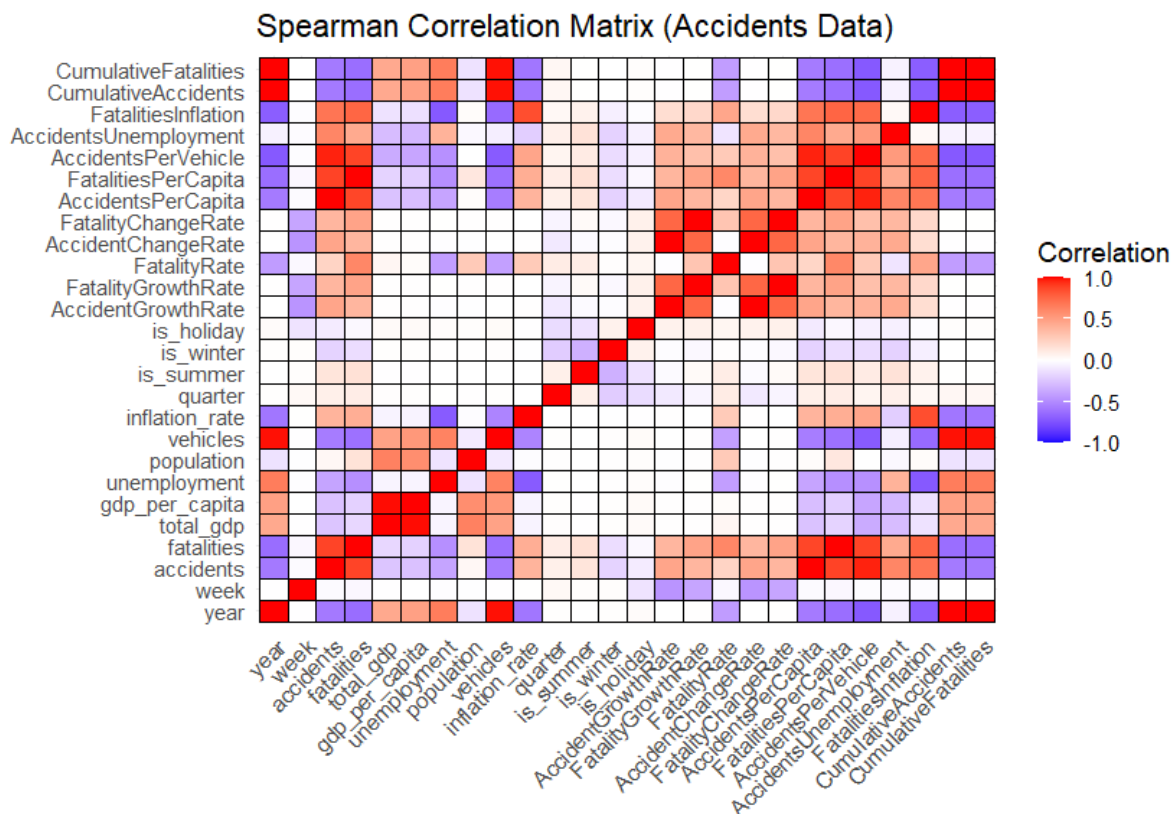
Στο πλαίσιο της πολυπαραγοντικής ανάλυσης των τροχαίων ατυχημάτων, η μελέτη των συσχετίσεων και της ομοδιακύμανσης μεταξύ μεταβλητών αποτελεί βασικό εργαλείο για την κατανόηση των εσωτερικών σχέσεων που διέπουν το φαινόμενο. Οι πίνακες συσχέτισης και ομοδιακύμανσης παρέχουν μια συνοπτική αλλά αποκαλυπτική αποτύπωση του τρόπου με τον οποίο εξελίσσονται από κοινού κρίσιμες μεταβλητές, όπως η ανεργία, ο αριθμός οχημάτων, το κατά κεφαλήν ΑΕΠ, αλλά και παράγωγοι δείκτες όπως τα ποσοστά και οι ρυθμοί μεταβολής ατυχημάτων και θανάτων. Μέσα από διαφορετικούς τύπους πινάκων, όπως ο Pearson για γραμμικές σχέσεις, ο Spearman και Kendall για μη γραμμικές και μονοτονικές τάσεις, καθώς και η ομοδιακύμανση για την απόλυτη μεταβολή, αποκτούμε πληρέστερη εικόνα για τη δυναμική των μεταβλητών και τις ενδεχόμενες συσχετίσεις μεταξύ τους.

Ο πίνακας συσχέτισης Pearson εξετάζει τις γραμμικές σχέσεις μεταξύ των μεταβλητών του συνόλου δεδομένων, αποτελώντας το πιο διαδεδομένο μέτρο συσχέτισης στη στατιστική ανάλυση. Παρατηρούνται ιδιαίτερα ισχυρές θετικές συσχετίσεις μεταξύ μεταβλητών που συνδέονται με την πάροδο του χρόνου, όπως το `year` και τα `CumulativeAccidents` και `CumulativeFatalities`, γεγονός που αντανακλά τη συσσώρευση συμβάντων με την πάροδο των ετών. Αντίστοιχα υψηλές τιμές παρατηρούνται και μεταξύ `total_gdp`, `vehicles`, και `population`, λόγω της παράλληλης αυξητικής τους πορείας. Αντίθετα, ο δείκτης `AccidentsPerVehicle` παρουσιάζει αρνητική συσχέτιση με το έτος, γεγονός που υποδηλώνει ότι, παρά την αύξηση των οχημάτων, η πιθανότητα ατυχήματος ανά όχημα μειώνεται. Τέλος, η μέτρια θετική συσχέτιση μεταξύ μεταβλητών όπως `AccidentRate` και `FatalityRate` αντανακλά τη σύνδεση του πλήθους ατυχημάτων με τις θανατηφόρες συνέπειές τους.



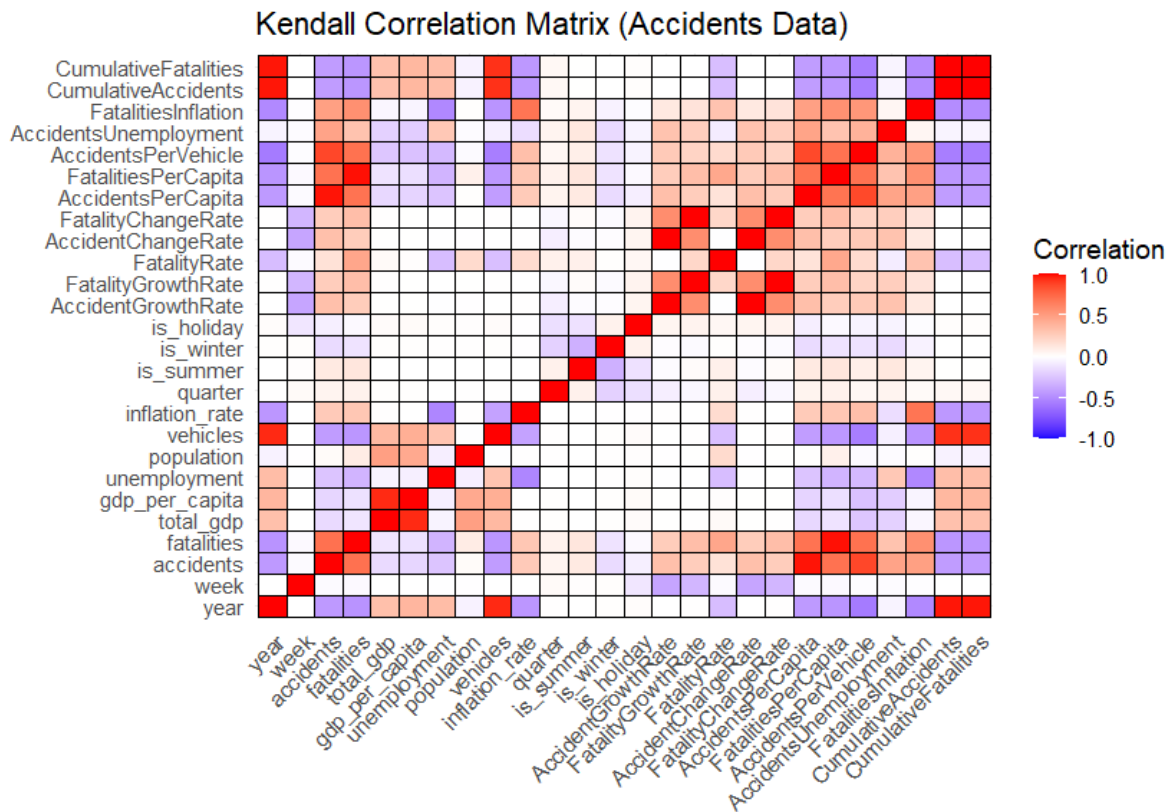
Σχήμα 5.3.4-1: Pearson Correlation Matrix για τα Δεδομένα Τροχαίων Ατυχημάτων

Ο πίνακας συσχέτισης Spearman, επιτρέπει τη μελέτη μονοτονικών σχέσεων, ανεξάρτητα από το αν αυτές είναι γραμμικές. Αυτό τον καθιστά ιδανικό για την ανάλυση δεδομένων χωρίς κανονικότητα ή μη γραμμικές τάσεις, όπως είναι συνήθως τα κοινωνικοοικονομικά ή καιρικά δεδομένα. Το πρότυπο των ισχυρών θετικών συσχετίσεων που παρατηρήθηκε στον Pearson επαναλαμβάνεται και εδώ, ιδιαίτερα για χρονικά εξελισσόμενες μεταβλητές. Ωστόσο, ο Spearman προσφέρει πιο σταθερές εκτιμήσεις σε ζεύγη όπως AccidentsUnemployment και FatalitiesPerCapita, όπου η σχέση δεν είναι αυστηρά γραμμική αλλά φαίνεται να διατηρεί μονοτονικό χαρακτήρα.



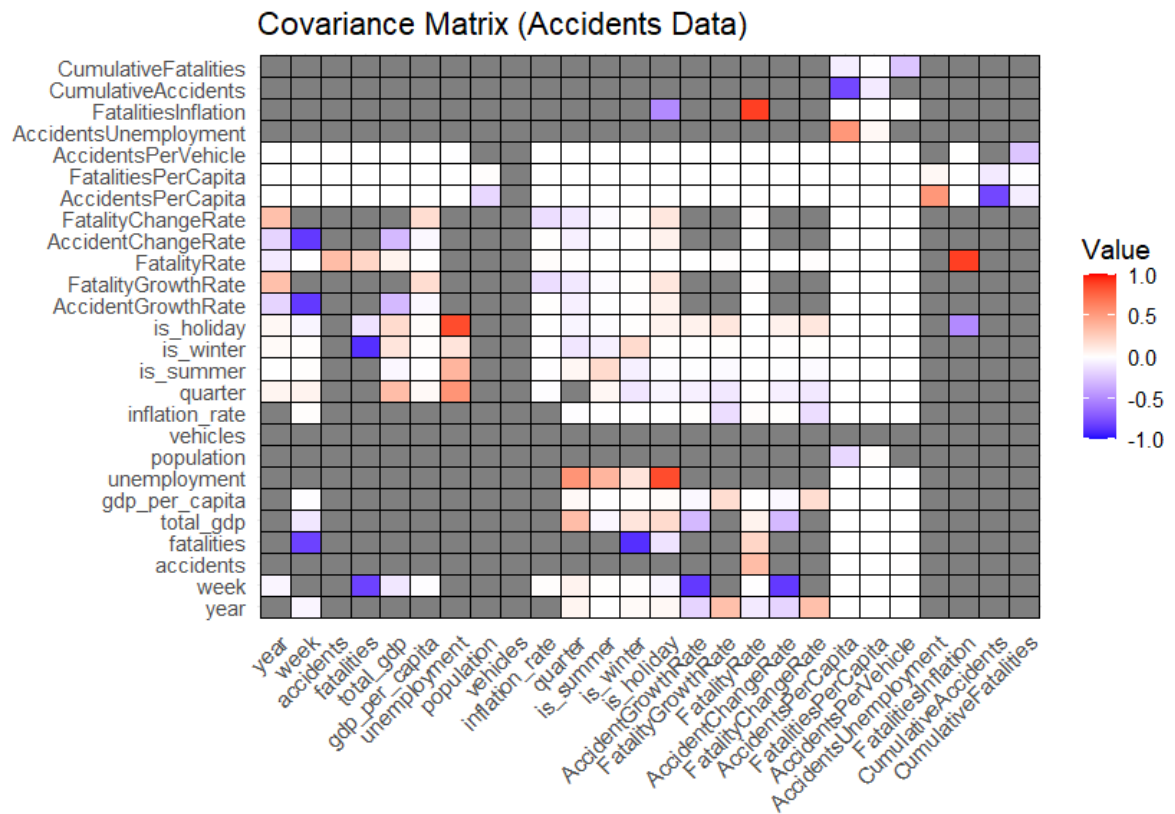
Σχήμα 5.3.4-2: Spearman Correlation Matrix για τα Δεδομένα Τροχαίων Ατυχημάτων

Ο πίνακας συσχέτισης Kendall καταγράφει με μεγαλύτερη ευαισθησία τις μονοτονικές σχέσεις σε σύγκριση με τον Spearman, καθώς βασίζεται σε ζεύγη παρατηρήσεων και συγκρίνει τις σχετικές ταξινομήσεις τους. Αν και τα μοτίβα που αποτυπώνονται είναι συμβατά με τα προηγούμενα, η κλίμακα τιμών είναι μικρότερη, κάτι που είναι χαρακτηριστικό της μεθόδου. Ο Kendall δίνει ιδιαίτερη έμφαση στην ισχύ των τάσεων σε μεταβλητές όπως *FatalityGrowthRate*, *AccidentsPerCapita* και *AccidentChangeRate*.



Σχήμα 5.3.4-3: Kendall Correlation Matrix για τα Δεδομένα Τροχαίων Ατυχημάτων

Τέλος, ο πίνακας ομοδιακύμανσης παρουσιάζει μια διαφορετική προσέγγιση, εστιάζοντας όχι στη σχετικότητα αλλά στη συνδιακύμανση των απόλυτων τιμών των μεταβλητών. Αν και επηρεάζεται έντονα από τις μονάδες μέτρησης και δεν κανονικοποιείται όπως η συσχέτιση, προσφέρει πολύτιμες ενδείξεις για την κοινή μεταβλητότητα ζευγών μεταβλητών. Εντοπίζονται ισχυρές θετικές ομοδιακυμάνσεις μεταξύ του *total_gdp* και των *vehicles* και *population*, γεγονός που είναι σύμφωνο με την οικονομική και πληθυσμιακή ανάπτυξη. Αντίθετα, μεταβλητές όπως το *is_winter*, που είναι δυαδικές, παρουσιάζουν μικρές ή μηδενικές τιμές ομοδιακύμανσης με πολλές συνεχείς μεταβλητές, γεγονός που επιβεβαιώνει την ασθενή ή ασυνεπή σύνδεση.



Σχήμα 5.3.4-4: Covariance Matrix για τα Δεδομένα Τροχαίων Ατυχημάτων

Κεφάλαιο 6 Στατιστική Μοντελοποίηση Χρονοσειρών

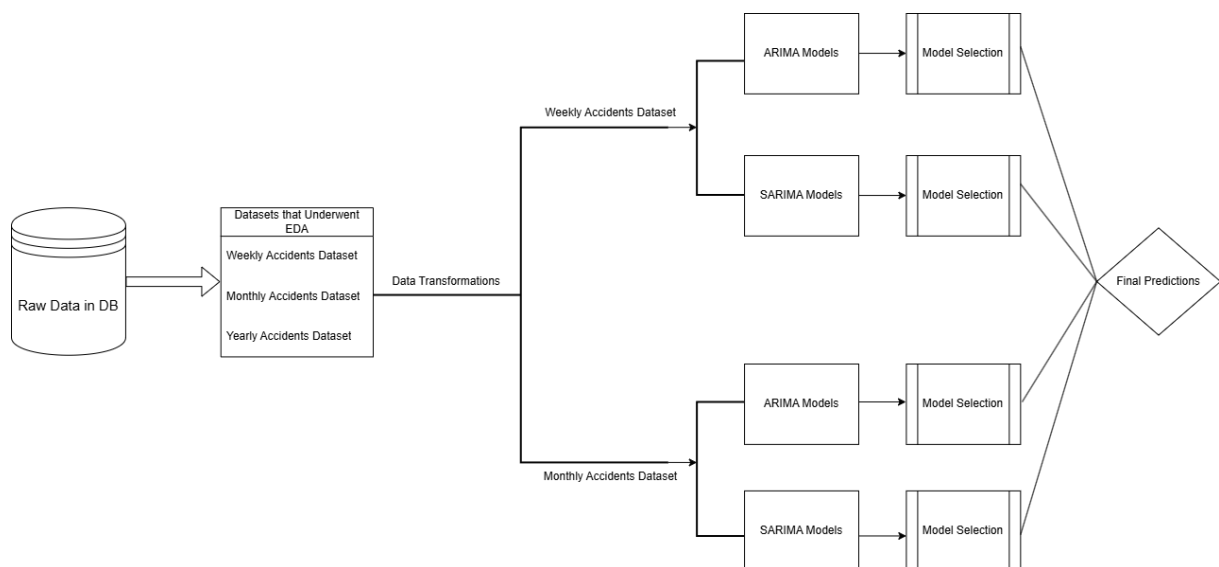
Η ανάγκη πρόβλεψης της μελλοντικής εξέλιξης των τροχαίων ατυχημάτων και θανάτων καθιστά τη μοντελοποίηση χρονοσειρών ιδιαίτερα χρήσιμη, καθώς επιτρέπει την αξιοποίηση της χρονικής πληροφορίας για την ανίχνευση υποκείμενων τάσεων, περιοδικότητας και συσχετίσεων. Στο πλαίσιο της παρούσας ανάλυσης, χρησιμοποιήθηκαν στοχαστικά μοντέλα όπως τα ARIMA και SARIMA, τα οποία συνδυάζουν στοιχεία αυτοπαλινδρόμησης και κινητού μέσου όρου με δυνατότητα ενσωμάτωσης εποχικών χαρακτηριστικών. Μέσω αυτής της προσέγγισης επιδιώχθηκε η αναπαράσταση της δομής της χρονοσειράς με τρόπο που να επιτρέπει αξιόπιστες προβλέψεις.

Η ανάπτυξη των μοντέλων βασίστηκε σε αυστηρά διαγνωστικά κριτήρια και στατιστικούς ελέγχους, με έμφαση σε έννοιες όπως η στασιμότητα, η αυτοσυσχέτιση και η επάρκεια των residuals. Χρησιμοποιήθηκαν κατάλληλοι μετασχηματισμοί για τη σταθεροποίηση της διακύμανσης, ενώ εφαρμόστηκαν διαφορισμοί για την απομάκρυνση των μακροπρόθεσμων τάσεων. Η επιλογή των βέλτιστων παραμέτρων πραγματοποιήθηκε με βάση δείκτες όπως το AIC και το BIC και MAPE, ενώ η επαλήθευση της καταλληλότητας κάθε μοντέλου βασίστηκε τόσο σε γραφικές όσο και σε αριθμητικές μεθόδους. Τα μοντέλα εφαρμόστηκαν μόνο σε εβδομαδιαία και μηνιαία δεδομένα καθώς προσφέρουν μεγαλύτερη χρονική ανάλυση και πιο πλούσια πληροφορία για την εκτίμηση των τάσεων και της εποχικότητας. Δεν συμπεριλήφθηκαν τα ετήσια δεδομένα για τη μοντελοποίηση, καθώς ο περιορισμένος αριθμός εγγραφών (μόλις 27 παρατηρήσεις για τα έτη 1996–2022) καθιστά δύσκολη την αξιόπιστη εφαρμογή στατιστικών μοντέλων και περιορίζει την ερμηνευτική και προβλεπτική τους ισχύ.

6.1 Μεθοδολογία και Workflow για την Στατιστική Μοντελοποίηση

Όσον αφορά την μεθοδολογία που ακολουθήθηκε, κάθε dataset (εβδομαδιαίο και μηνιαίο) υποβλήθηκε σε ξεχωριστή μοντελοποίηση, με την αξιολόγηση των μοντέλων να βασίζεται τόσο στην ακρίβεια της προσαρμογής στα ιστορικά δεδομένα όσο και στην προβλεπτική τους ικανότητα. Για κάθε χρονοσειρά δοκιμάστηκαν συνδυασμοί παραμέτρων τάξης (p, d, q) και εποχικών παραμέτρων (P, D, Q, s) με διάφορους τρόπους, με στόχο την επίτευξη ισορροπίας μεταξύ πολυπλοκότητας και απόδοσης. Η διαδικασία περιλάμβανε τόσο in-sample εκτίμηση όσο και εκτός δείγματος προβλέψεις (out-of-sample), για να διασφαλιστεί η γενίκευση των αποτελεσμάτων.

Η παραπάνω διαδικασία εφαρμόστηκε σε όλα τα διαθέσιμα datasets. Πιο συγκεκριμένα, πραγματοποιήθηκε μοντελοποίηση με ARIMA για τα εβδομαδιαία δεδομένα ατυχημάτων και θανάτων, καθώς και αντίστοιχη προσέγγιση με SARIMA, όπου υπήρχε ένδειξη εποχικότητας. Παράλληλα, εξετάστηκαν οι λογαριθμικά μετασχηματισμένες εκδοχές των ίδιων χρονοσειρών (log-transformed), τόσο για τα ARIMA όσο και για τα SARIMA μοντέλα, προκειμένου να αντιμετωπιστούν φαινόμενα ετεροσκεδαστικότητας και να βελτιωθεί η γραμμικότητα της σχέσης. Η ίδια αναλυτική στρατηγική ακολουθήθηκε εξίσου και για τα μηνιαία δεδομένα, καλύπτοντας πλήρως όλες τις παραλλαγές των διαθέσιμων χρονοσειρών (σε ακατέργαστη και log μορφή), τόσο για τα ατυχήματα όσο και για τους θανάτους.



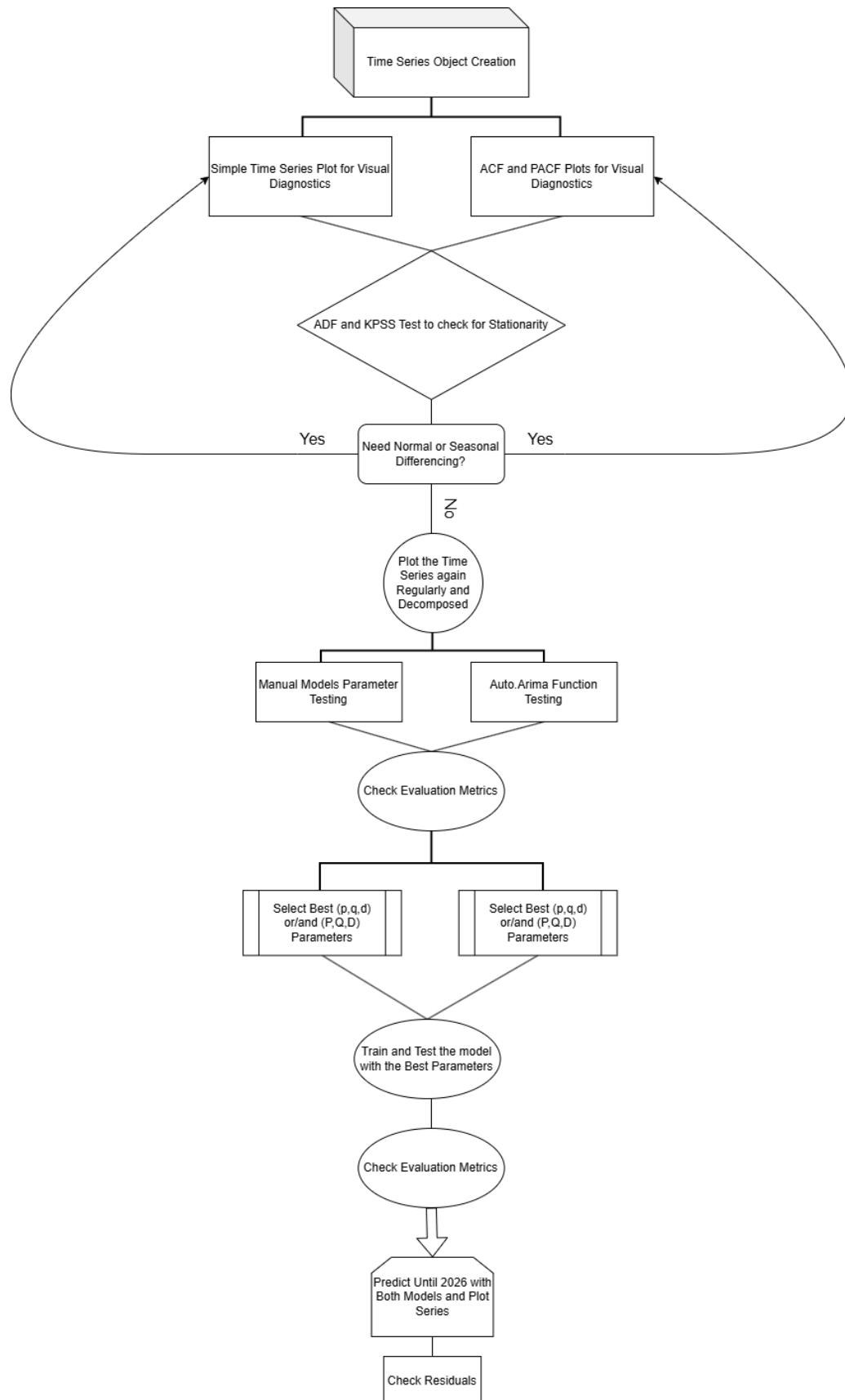
Σχήμα 5.3.4-1: Workflow Στατιστικής Μοντελοποίησης

6.2 Διαδικασία Επιλογής και Εξέτασης Μοντέλου Πρόβλεψης

Η επιλογή του κατάλληλου μοντέλου ARIMA ή SARIMA για την πρόβλεψη των εβδομαδιαίων και μηνιαίων ατυχημάτων και θανάτων βασίστηκε σε μια μεθοδική διαδικασία ελέγχου, εκπαίδευσης και αξιολόγησης. Αρχικά, οι χρονοσειρές μετασχηματίστηκαν σε time series αντικείμενα με τη σωστή συχνότητα και ελέγχθηκαν για στασιμότητα μέσω των ADF και KPSS τεστ. Σε περιπτώσεις όπου δεν υπήρχε συμφωνία μεταξύ των δύο, εφαρμόστηκε διαφορά πρώτης τάξης ($d = 1$). Στη συνέχεια εξετάστηκαν τα ACF και PACF διαγράμματα για την αρχική εκτίμηση των παραμέτρων AR και MA και δοκιμάστηκαν χειροκίνητα διάφοροι συνδυασμοί μοντέλων. Η απόδοση κάθε μοντέλου αξιολογήθηκε με βάση τα AIC/BIC, καθώς και με τις μετρικές σφάλματος RMSE, MAE και MAPE σε out-of-sample πρόβλεψη για την περίοδο 2018–2022.

Παράλληλα με τα χειροκίνητα μοντέλα, χρησιμοποιήθηκε και η συνάρτηση `auto.arima()` με πλήρη αναζήτηση (χωρίς stepwise), ώστε να εντοπιστούν αυτόματα βέλτιστα μοντέλα με βάση το AICc. Τα επικρατέστερα μοντέλα (χειροκίνητα και αυτόματα) επανεκπαιδεύτηκαν σε ολόκληρη τη χρονοσειρά και χρησιμοποιήθηκαν για πρόβλεψη μέχρι και το 2026. Η ίδια διαδικασία εφαρμόστηκε τόσο για ακατέργαστα όσο και για λογαριθμικά δεδομένα, ενώ αξιολογήθηκαν και εποχικές παραλλαγές SARIMA όταν υπήρχαν ενδείξεις εποχικότητας. Τα residuals ελέγχθηκαν με τα τεστ Ljung-Box, Box-Pierce και τη συνάρτηση `checkresiduals()` για την επιβεβαίωση της καταλληλότητας κάθε μοντέλου.

Συνολικά, η διαδικασία επιλογής, εκπαίδευσης και αξιολόγησης μοντέλων εφαρμόστηκε σε 16 ξεχωριστές περιπτώσεις, καλύπτοντας πλήρως τις διαστάσεις του προβλήματος. Πιο συγκεκριμένα, για κάθε συνδυασμό χρονικής ανάλυσης (εβδομαδιαία και μηνιαία) και μεταβλητής στόχου (ατυχήματα και θάνατοι), υλοποιήθηκαν τέσσερις εκδοχές μοντελοποίησης: ARIMA και SARIMA σε ακατέργαστα δεδομένα, καθώς και ARIMA και SARIMA σε λογαριθμικά μετασχηματισμένα δεδομένα.



Σχήμα 5.3.4-1: Μεθοδολογία του Model Selection and Testing

Μετά την εκτέλεση των 16 βασικών σεναρίων ARIMA και SARIMA για τα εβδομαδιαία και μηνιαία δεδομένα ατυχημάτων και θανάτων, προέκυψε ότι αρκετά από τα μοντέλα δεν πληρούσαν τα στατιστικά κριτήρια καταλληλότητας των residuals. Συγκεκριμένα, οι έλεγχοι στα residuals όπως το τεστ Ljung-Box ή η συνάρτηση `checkresiduals()` αποκάλυψαν σημαντική αυτοσυσχέτιση και μη λευκό θόρυβο σε αρκετές περιπτώσεις. Αυτό οδήγησε στην ανάγκη εκτέλεσης επιπλέον διαδικασιών επιλογής μοντέλου, μέσα από στοχευμένα grid searches για τις παραμέτρους των ARIMA/SARIMA. Η νέα διαδικασία αποσκοπούσε στη βελτιστοποίηση του MAPE (Mean Absolute Percentage Error), αλλά και στη διασφάλιση ότι το τελικό μοντέλο περνά επιτυχώς τους διαγνωστικούς ελέγχους. Έτσι, για κάθε χρονοσειρά επαναλήφθηκε η εκπαίδευση πολλαπλών μοντέλων με διαφορετικούς συνδυασμούς (p,d,q) και (P,D,Q) , ώστε να επιλεγούν εκείνα που ικανοποιούν ταυτόχρονα κριτήρια ακρίβειας και στατιστικής επάρκειας.

Εκτός αυτού, ιδιαίτερη πρόκληση παρουσίασε η μοντελοποίηση των εβδομαδιαίων δεδομένων, καθώς το αρχικό dataset περιλάμβανε ακανόνιστο αριθμό εβδομάδων ανά μήνα, εξαιτίας της ασυμβατότητας της ημερολογιακής εβδομάδας. Για να καταστεί δυνατή η εφαρμογή μοντέλων που βασίζονται σε σταθερή περιοδικότητα, εφαρμόστηκε αλγόριθμος μετασχηματισμού με στόχο την εξομάλυνση των εβδομάδων σε σταθερή βάση 4 ανά μήνα (δηλαδή 48 ανά έτος). Η μεθοδολογία αυτή περιλάμβανε την ανακατανομή των επιπλέον εβδομαδιαίων καταγραφών (όσων εβδομάδων υπερέβαιναν τις 4), με ισοκατανομή των αθροιστικών τιμών ατυχημάτων και θανάτων στις πρώτες τέσσερις εβδομάδες του μήνα. Η διαδικασία προέβλεπε επιμερισμό των επιπλέον περιστατικών με ακρίβεια, τόσο ισόποσα όσο και με υπόλοιπα σε επιλεγμένες εβδομάδες, ώστε να διατηρηθεί η συνολική ποσότητα και να επιτευχθεί συγχρονισμός με τη δομή χρονοσειράς που απαιτούν τα ARIMA/SARIMA μοντέλα. Μετά την προσαρμογή, δημιουργήθηκε το τελικό dataset `acc.w.ind.y_4wmod`, πάνω στο οποίο πραγματοποιήθηκε όλη η μετέπειτα διαδικασία.

6.2.1 ΔΗΜΙΟΥΡΓΙΑ ΑΝΤΙΚΕΙΜΕΝΟΥ TIME SERIES ΣΤΗΝ R

Η δημιουργία αντικειμένων χρονοσειράς στην R αποτελεί το πρώτο απαραίτητο βήμα για τη μοντελοποίηση με τεχνικές όπως ARIMA και SARIMA. Για τον σκοπό αυτό χρησιμοποιείται η συνάρτηση `ts()`, η οποία μετατρέπει μια ακολουθία αριθμητικών παρατηρήσεων σε αντικείμενο τύπου time series, καθορίζοντας τη χρονική έναρξη και τη συχνότητα

παρατήρησης. Η παράμετρος *start* ορίζει το έτος και τη μονάδα (π.χ. εβδομάδα ή μήνα) της πρώτης παρατήρησης, ενώ η *frequency* καθορίζει τον αριθμό των χρονικών περιόδων ανά έτος. Για παράδειγμα, για εβδομαδιαία δεδομένα η συχνότητα ορίζεται ως 48, λαμβάνοντας υπόψη το τροποποιημένο σύστημα τετραεβδομαδιαίων ετών (48 εβδομάδες), ενώ για μηνιαία δεδομένα η *frequency* είναι 12, αντιστοιχώντας στους 12 μήνες του έτους.

Η μετατροπή αυτή επιτρέπει στη R να αναγνωρίσει τη δομή της χρονοσειράς, γεγονός που είναι απαραίτητο για όλες τις επόμενες αναλυτικές διαδικασίες, όπως την οπτική απεικόνιση, τον έλεγχο στασιμότητας, την αποσύνθεση της τάσης και της εποχικότητας, καθώς και την εφαρμογή μοντέλων πρόβλεψης.

6.2.2 ΔΙΑΓΝΩΣΤΙΚΑ PLOTS ΚΑΙ TESTS ΚΑΙ SEASONAL / NORMAL DIFFERENCING

Στο στάδιο της ανάλυσης πριν την εφαρμογή μοντέλων ARIMA/SARIMA, όπως και προαναφέρθηκε, κρίθηκε αναγκαίο να διεξαχθούν μια σειρά από διαγνωστικοί έλεγχοι, ώστε να επιβεβαιωθεί η καταλληλότητα της χρονοσειράς για μοντελοποίηση. Αρχικά, εφαρμόζονται τα γραφήματα αυτοσυσχέτισης (ACF) και μερικής αυτοσυσχέτισης (PACF) στην αρχική (ακατέργαστη) χρονοσειρά ατυχημάτων. Αυτά τα διαγράμματα παρέχουν πληροφορίες για τη δομή εξάρτησης της σειράς με το παρελθόν της, καθώς και για την πιθανή παρουσία εποχικότητας ή άλλων μοτίβων. Το ACF δείχνει το βαθμό συσχέτισης κάθε παρατήρησης με τις προηγούμενες τιμές της, ενώ το PACF εστιάζει στην καθαρή συσχέτιση με συγκεκριμένα lags, αφού πρώτα αφαιρεθεί η επίδραση των ενδιάμεσων καθυστερήσεων. Έτσι, αυτά τα διαγράμματα είναι χρήσιμα για τον εντοπισμό της κατάλληλης τάξης (p , q) του μοντέλου ARIMA, καθώς αιχμές σε συγκεκριμένες υστερήσεις μπορούν να υποδεικνύουν την ανάγκη για αντίστοιχους όρους AR ή MA.

Στη συνέχεια, αξιολογείται η στασιμότητα της χρονοσειράς μέσω δύο διαφορετικών στατιστικών τεστ: του ADF (Augmented Dickey-Fuller) και του KPSS (Kwiatkowski–Phillips–Schmidt–Shin). Τα δύο αυτά τεστ ακολουθούν αντίθετες υποθέσεις: το ADF ελέγχει την ύπαρξη μοναδιαίας ρίζας (δηλαδή μη στασιμότητας) ως μηδενική υπόθεση, ενώ το KPSS εξετάζει τη στασιμότητα ως μηδενική υπόθεση. Η σύγκριση των αποτελεσμάτων τους προσφέρει μια πιο ισχυρή ένδειξη για την ανάγκη διαφοροποίησης. Στην περίπτωση όπου το ADF αποτυγχάνει να απορρίψει τη μη στασιμότητα ($p > 0.05$) και το KPSS απορρίπτει τη στασιμότητα ($p < 0.05$), τότε επιβεβαιώνεται ότι απαιτείται κανονική διαφοροποίηση ($d = 1$).

Η διαφοροποίηση πρώτης τάξης $\text{diff}()$ εφαρμόζεται για την αφαίρεση της τάσης και τη μετατροπή της σειράς σε στάσιμη, δηλαδή σε μια σειρά που έχει σταθερό μέσο και διακύμανση στο χρόνο. Αυτή η μετατροπή είναι απαραίτητη ώστε να ικανοποιούνται οι προϋποθέσεις των ARIMA / SARIMA μοντέλων. Το αποτέλεσμα της διαφοροποιημένης σειράς απεικονίζεται σε διάγραμμα, στο οποίο συνήθως παρατηρείται εξομάλυνση της τάσης, με πιο σταθερή διακύμανση γύρω από το μηδέν.

Μετά τη διαφοροποίηση, επαναλαμβάνονται τα τεστ ADF και KPSS για να επιβεβαιωθεί ότι η νέα σειρά είναι πλέον στατική. Στην ιδανική περίπτωση, το ADF απορρίπτει τη μη στασιμότητα και το KPSS αποτυγχάνει να απορρίψει τη στασιμότητα, γεγονός που επιβεβαιώνει ότι η διαφοροποίηση πρώτης τάξης είναι επαρκής. Αντίστοιχα, η ανάλυση ACF και PACF εφαρμόζεται ξανά και στη διαφοροποιημένη σειρά, καθώς τώρα παρέχει πιο αξιόπιστες πληροφορίες για τον προσδιορισμό των κατάλληλων παραμέτρων των μοντέλων ARIMA / SARIMA.

Τέλος, εφαρμόζεται και ο αποσυνθετικός αλγόριθμος `decompose()`, ο οποίος διαχωρίζει τη χρονοσειρά στα επιμέρους συστατικά της: τάση (trend), εποχικότητα (seasonal) και τυχαίο θόρυβο (residual). Η απεικόνιση αυτών των συστατικών προσφέρει καλύτερη κατανόηση της δομής της σειράς, ενώ βοηθά να διαπιστωθεί αν υπάρχει εμφανής περιοδικότητα που δεν εξαλείφεται πλήρως με τη διαφοροποίηση. Σε περιπτώσεις έντονης εποχικής συνιστώσας, μπορεί να κριθεί απαραίτητη η χρήση εποχικών διαφορών (seasonal differencing) και η εφαρμογή SARIMA μοντέλων με προσθήκη παραμέτρων (P, D, Q, s).

6.2.3 MANUAL FITTING KAI FITTING ME AUTO.ARIMA()

Η διαδικασία προσαρμογής μοντέλων ARIMA / SARIMA περιλάμβανε δύο προσεγγίσεις: τη χειροκίνητη επιλογή παραμέτρων και τη χρήση του αυτοματοποιημένου αλγορίθμου `auto.arima()` της βιβλιοθήκης `forecast`. Αρχικά, η χρονοσειρά χωρίστηκε σε δύο τμήματα, `training set` και `test set`, με το όριο τοποθετημένο χρονικά στο μέσο του 2017. Αυτό το χρονικό `split` επέτρεψε τη δοκιμή των μοντέλων σε `out-of-sample` δεδομένα, γεγονός που προσδίδει αξιοπιστία στη διαδικασία αξιολόγησης, καθώς εξετάζεται η ικανότητα του μοντέλου να γενικεύει πέρα από τα δεδομένα στα οποία εκπαιδεύτηκε. Στο πρώτο σκέλος, η προσέγγιση του `manual fitting` βασίστηκε σε συστηματική δοκιμή διαφόρων συνδυασμών τιμών για τις παραμέτρους p και q , διατηρώντας σταθερή την τιμή $d=1$, η οποία είχε ήδη καθοριστεί από τα διαγνωστικά tests (ADF και KPSS) ότι είναι απαραίτητη για τη σταθεροποίηση της χρονοσειράς. Οι δοκιμές περιορίστηκαν με τα p, q να λαμβάνουν τιμές $\{0, 1, 2, 3\}$ ώστε να αποφευχθεί η υπερβολική πολυπλοκότητα και να μην γίνει `overfitting` στο μοντέλο.

Για κάθε υποψήφιο μοντέλο, πραγματοποιήθηκε πρόβλεψη για το `test set` και υπολογίστηκαν τρεις μετρικές σφάλματος: RMSE (Root Mean Square Error), MAE (Mean Absolute Error) και MAPE (Mean Absolute Percentage Error). Οι μετρικές αυτές καταγράφονταν συστηματικά σε πίνακα Excel ώστε να διευκολυνθεί η σύγκριση των μοντέλων με ποσοτικούς όρους. Παράλληλα, σημειώνονταν και τα αντίστοιχα AIC (Akaike Information Criterion) και BIC (Bayesian Information Criterion) κάθε μοντέλου.

Στη συνέχεια, αξιοποιήθηκε η συνάρτηση `auto.arima()`, η οποία λειτουργεί ως εργαλείο αυτόματης επιλογής παραμέτρων βάσει εξαντλητικής αναζήτησης. Για να διασφαλιστεί η ποιότητα της αναζήτησης, οι επιλογές `stepwise = FALSE` και `approximation = FALSE` τέθηκαν σε ισχύ, οδηγώντας σε πλήρη δοκιμή όλων των πιθανών συνδυασμών, χωρίς να χρησιμοποιείται προσεγγιστική μέθοδος. Το `auto.arima()` επιλέγει το βέλτιστο μοντέλο σύμφωνα με το κριτήριο AICc (AIC με διόρθωση για μικρά δείγματα), το οποίο είναι πιο αυστηρό από το απλό AIC. Η διαδικασία αυτή λειτούργησε συμπληρωματικά προς το χειροκίνητο μοντέλο, είτε επιβεβαιώνοντας τη δομή του είτε προτείνοντας μια εναλλακτική με χαμηλότερα κριτήρια. Και εδώ, τα μοντέλα αξιολογήθηκαν με τις ίδιες μετρικές σφάλματος (RMSE, MAE, MAPE), διευκολύνοντας τη συγκριτική αξιολόγηση και την επιλογή της τελικής παραμετροποίησης.

Η σύγκριση μεταξύ της χειροκίνητης και της αυτόματης διαδικασίας αποτέλεσε σημαντικό κομμάτι της συνολικής στρατηγικής μοντελοποίησης, αφού συχνά παρατηρείται ότι

το `auto.arima()` οδηγεί σε διαφορετικά μοντέλα από αυτά που θα επέλεγε κάποιος χειροκίνητα. Η κριτική ανάλυση των αποτελεσμάτων περιλάμβανε την αξιολόγηση της συμπεριφοράς των residuals, την προβλεπτική ακρίβεια στο test set και την ερμηνευσιμότητα της δομής του μοντέλου. Αν και σε αρκετές περιπτώσεις τα αυτόματα μοντέλα παρουσίασαν χαμηλότερα AIC/BIC, προτιμήθηκε το χειροκίνητο μοντέλο όταν εμφάνιζε μικρότερη MAPE και καλύτερη διαγνωστική συμπεριφορά στα residuals. Σε άλλες περιπτώσεις, τα αποτελέσματα του `auto.arima()` αποτέλεσαν τη βάση για περαιτέρω παραμετροποίηση.

Τέλος, η επιλογή των επικρατέστερων μοντέλων πραγματοποιήθηκε όχι μόνο με βάση την απόδοσή τους στα δεδομένα εκπαίδευσης και ελέγχου, αλλά και με στόχο τη γενίκευση και τη χρήση τους στην τελική πρόβλεψη μέχρι το 2026. Τα επιλεγμένα μοντέλα εκπαιδεύτηκαν εκ νέου σε όλη τη διαθέσιμη χρονοσειρά και χρησιμοποιήθηκαν για προβλέψεις 192 εβδομάδων (ή αντίστοιχα 48 μηνών) στο μέλλον.

```

5628 #Trying some possible ARIMA Combinations limited to max(p,q) = 4#
5629
5630 train_ts_arima_fat_w <- window(fat_w_ts, end = c(2017, 29))
5631 test_ts_arima_fat_w <- window(fat_w_ts, start = c(2017, 30))
5632 model_arima_fat_w <- Arima(train_ts_arima_fat_w, order = c(1,1,2))
5633 fc_arima_fat_w <- forecast(model_arima_fat_w, h = length(test_ts_arima_fat_w))
5634 rmse_val_arima_fat_w <- rmse(test_ts_arima_fat_w, fc_arima_fat_w$mean)
5635 mae_val_arima_fat_w <- mae(test_ts_arima_fat_w, fc_arima_fat_w$mean)
5636 mape_val_arima_fat_w <- mape(test_ts_arima_fat_w, fc_arima_fat_w$mean) * 100
5637 summary(model_arima_fat_w)
5638 cat("AIC:", AIC(model_arima_fat_w), "\n")
5639 cat("BIC:", BIC(model_arima_fat_w), "\n")
5640 cat("RMSE:", rmse_val_arima_fat_w, "\n")
5641 cat("MAE:", mae_val_arima_fat_w, "\n")
5642 cat("MAPE:", mape_val_arima_fat_w, "%\n")
5643 AIC(model_arima_fat_w)
5644 BIC(model_arima_fat_w)
5645
5646 #Best for now is ARIMA(1,1,2), lets see with Auto.Arima#
5647
5648 model_auto_arima_fat_w <- auto.arima(
5649   train_ts_arima_fat_w,
5650   seasonal = FALSE,
5651   stepwise = FALSE,
5652   approximation = FALSE,
5653   trace = TRUE
5654 )
5655 fc_auto_arima_fat_w <- forecast(model_auto_arima_fat_w, h = length(test_ts_arima_fat_w))
5656 summary(fc_auto_arima_fat_w)
5657 summary(model_auto_arima_fat_w)
5658 rmse_val_auto_arima_fat_w <- rmse(test_ts_arima_fat_w, fc_auto_arima_fat_w$mean)
5659 mae_val_auto_arima_fat_w <- mae(test_ts_arima_fat_w, fc_auto_arima_fat_w$mean)
5660 mape_val_auto_arima_fat_w <- mape(test_ts_arima_fat_w, fc_auto_arima_fat_w$mean) * 100
5661 cat("AIC:", AIC(model_auto_arima_fat_w), "\n")
5662 cat("BIC:", BIC(model_auto_arima_fat_w), "\n")
5663 cat("RMSE:", rmse_val_auto_arima_fat_w, "\n")
5664 cat("MAE:", mae_val_auto_arima_fat_w, "\n")
5665 cat("MAPE:", mape_val_auto_arima_fat_w, "%\n")

```

Σχήμα 6.2.3-1: Ενδεικτικό Block Κώδικα για Manual Fitting και Fitting με Auto.Arima()

6.2.4 RESIDUALS CHECKS

Έπειτα από την εκπαίδευση των τελικών μοντέλων ARIMA, τόσο των χειροκίνητα παραμετροποιημένων όσο και αυτών που προτάθηκε από τον αλγόριθμο `auto.arima()`, κρίθηκε απαραίτητο να ελεγχθεί η συμπεριφορά των residuals, καθώς η επάρκεια ενός μοντέλου δεν εξαρτάται μόνο από τις προβλεπτικές του επιδόσεις, αλλά και από τη στατιστική του ορθότητα. Τα residuals ενός σωστά προσαρμοσμένου ARIMA μοντέλου θα πρέπει να συμπεριφέρονται ως λευκός θόρυβος, δηλαδή να μην περιέχουν υπολειπόμενη δομή, αυτοσυσχέτιση ή εποχικότητα. Για τον λόγο αυτό εφαρμόστηκαν τα τεστ Ljung–Box και Box–Pierce σε διαφορετικά lags (10 και 12), τα οποία εξετάζουν αν τα residuals είναι ανεξάρτητα.

Επιπλέον, χρησιμοποιήθηκε και η συνάρτηση `checkresiduals()` της βιβλιοθήκης `forecast`, η οποία προσφέρει έναν συνδυαστικό έλεγχο μέσα από γραφικές και στατιστικές μεθόδους όπως η απεικόνιση των residuals στον χρόνο, histogram για τον έλεγχο της κατανομής τους, και ACF plot για την ανίχνευση τυχόν υπολειπόμενης αυτοσυσχέτισης.

6.2.5 GRID SEARCHES

Σε αρκετές περιπτώσεις κατά τη μοντελοποίηση των χρονοσειρών, η διαδικασία fitting είτε μέσω χειροκίνητης παραμετροποίησης είτε με τη χρήση της συνάρτησης `auto.arima()` δεν απέφερε μοντέλα που να πληρούν επαρκώς τα κριτήρια για να περάσουν το residuals checks. Συγκεκριμένα, υπήρξαν περιπτώσεις όπου τα residuals παρουσίαζαν σημαντική αυτοσυσχέτιση ή συστηματικές αποκλίσεις, παραβιάζοντας έτσι την υπόθεση του λευκού θορύβου. Επειδή η επιλογή μοντέλου μόνο βάσει AIC ή BIC μπορεί να αγνοήσει ζητήματα στατιστικής επάρκειας, κρίθηκε αναγκαία η εφαρμογή μιας πιο συστηματικής προσέγγισης μέσω grid search.

Η τεχνική του grid search περιλάμβανε τον εξαντλητικό έλεγχο πλήθους διαφορετικών συνδυασμών παραμέτρων τάξης (p, d, q) και, όπου χρειαζόταν, εποχικών παραμέτρων (P, D, Q, s) σε ένα προκαθορισμένο εύρος τιμών. Για κάθε συνδυασμό μοντέλου, πραγματοποιούνταν εκπαίδευση, πρόβλεψη και υπολογισμός κρίσιμων μετρικών όπως RMSE, MAE, και MAPE, ενώ ταυτόχρονα εφαρμόζονταν διαγνωστικά τεστ (Ljung–Box, ACF/PACF, `checkresiduals`) για να διαπιστωθεί αν τα residuals ικανοποιούσαν τις απαραίτητες υποθέσεις. Μόνο τα μοντέλα

που παρουσίαζαν καλή προβλεπτική απόδοση και στατιστικά αποδεκτά residuals θεωρούνταν κατάλληλα για τελική επιλογή και προεκτάσεις πρόβλεψης.

```

7846 #Prepare time series#
7847
7848 acc_w_ts_gs_log <- loglp(ts(acc.w.ind.y_4wmod$accidents, start = c(1996, 1), frequency = 48))
7849 train_ts_arima_acc_w_gs_log <- window(acc_w_ts_gs_log, end = c(2017, 29))
7850 test_ts_arima_acc_w_gs_log <- window(acc_w_ts_gs_log, start = c(2017, 30))
7851 p_gs_acc_log <- 0:3; d_gs_acc_log <- 0:3; q_gs_acc_log <- 0:3
7852 results_gs_arima_acc_log <- list()
7853 tried_models_gs_arima_acc_log <- 0
7854 total_combinations_gs_arima_acc_log <- length(p_gs_acc_log) * length(d_gs_acc_log) * length(q_gs_acc_log)
7855 cat("Total combinations to try:", total_combinations_gs_arima_acc_log, "\n")
7856 for (i in p_gs_acc_log) {
7857   for (j in d_gs_acc_log) {
7858     for (k in q_gs_acc_log) {
7859       tried_models_gs_arima_acc_log <- tried_models_gs_arima_acc_log + 1
7860       order <- c(i, j, k)
7861       cat("\nTrying model", tried_models_gs_arima_acc_log, "of", total_combinations_gs_arima_acc_log,
7862         "\n-> ARIMA(", paste(order, collapse = ","), ")")
7863       tryCatch({
7864         fit <- Arima(train_ts_arima_acc_w_gs_log,
7865           order = order,
7866           method = "ML")
7867         h <- length(test_ts_arima_acc_w_gs_log)
7868         fc <- forecast(fit, h = h)
7869         fc_original <- expm1(fc$mean)
7870         test_original <- expm1(test_ts_arima_acc_w_gs_log)
7871         mape <- accuracy(fc_original, test_original)[ "Test set", "MAPE"]
7872         lb_test <- Box.test(residuals(fit), lag = 10, type = "Ljung-Box", fitdf = length(fit$coef))
7873         bp_test <- Box.test(residuals(fit), lag = 12, type = "Box-Pierce", fitdf = length(fit$coef))
7874         if (lb_test$p.value > 0.05 && bp_test$p.value > 0.05) {
7875           cat("Model passed residual tests. Saving ... \n")
7876           results_gs_arima_acc_log <- append(results_gs_arima_acc_log, list(list(
7877             order = paste0("(", paste(order, collapse = ","), ")"),
7878             AIC = AIC(fit),
7879             BIC = BIC(fit),
7880             MAPE = mape,
7881             LB_p = lb_test$p.value,
7882             BP_p = bp_test$p.value
7883           )))
7884         } else {
7885           cat("Residuals failed white noise tests.\n")
7886         }
7887       }, error = function(e) {
7888         message("Error for ARIMA(", paste(order, collapse = ","), "): ", e$message)
7889       })
7890     }
7891   }
7892 }

```

Σχήμα 6.2.5-1: Ενδεικτικό Block Κώδικα που πραγματοποιεί Grid Search για Εύρεση Μοντέλου που να Περνάει το Residual Check

6.2.6 LOG TRANSFORMATIONS

Στο πλαίσιο της συγκεκριμένης ανάλυσης, ήταν απαραίτητη η λογαριθμική μετατροπή (log transformation) των δεδομένων με σκοπό τη βελτίωση της σταθερότητας της διακύμανσης και την ενίσχυση της καταλληλότητας των μοντέλων ARIMA και SARIMA. Σε πολλές περιπτώσεις, οι αρχικές χρονοσειρές των τροχαίων ατυχημάτων και θανάτων παρουσίαζαν έντονη ετεροσκεδαστικότητα, δηλαδή, αυξανόμενη ή μη σταθερή διακύμανση σε συνάρτηση με το χρόνο, γεγονός που παραβιάζει βασικές υποθέσεις των γραμμικών μοντέλων και μπορεί να οδηγήσει σε λανθασμένες εκτιμήσεις των παραμέτρων ή ανακριβείς προβλέψεις. Η εφαρμογή του φυσικού λογαρίθμου (log) επέτρεψε τη συμπίεση των υψηλότερων τιμών

και την ομαλοποίηση της κατανομής, καθιστώντας τις χρονοσειρές πιο κατάλληλες για ανάλυση υπό συνθήκες στατιστικής ορθότητας.

Σε αυτό το πλαίσιο, η χρήση log μετασχηματισμένων χρονοσειρών προσέφερε μοντέλα που απέδωσαν πιο συνεπείς προβλέψεις σε ορισμένες περιπτώσεις, τόσο σε επίπεδο χρονικής συμπεριφοράς όσο και ως προς τις διαγνωστικές μετρικές (MAPE, AIC, residuals checks). Η ίδια διαδικασία fitting και αξιολόγησης επαναλήφθηκε πλήρως πάνω στα log δεδομένα, ώστε να εξεταστεί αν η νέα δομή παρήγαγε μοντέλα με καλύτερη προσαρμογή και μεγαλύτερη αξιοπιστία στις μακροπρόθεσμες προβλέψεις.

6.3 Αποτελέσματα από την Μεθοδολογία Επιλογής Μοντέλων ARIMA / SARIMA

Η διαδικασία επιλογής και αξιολόγησης μοντέλων ARIMA και SARIMA για τις χρονοσειρές τροχαίων ατυχημάτων και θανάτων, τόσο σε εβδομαδιαία όσο και σε μηνιαία βάση, οδήγησε σε ένα σύνολο υποψήφιων μοντέλων που κρίθηκαν κατάλληλα για την τελική πρόβλεψη. Μετά από διαδοχικά στάδια ανάλυσης, μοντελοποίησης, residual diagnostics και συγκριτικής αξιολόγησης ως προς τις προβλεπτικές τους επιδόσεις, δημιουργήθηκε ένας συγκεντρωτικός πίνακας αποτελεσμάτων που συνοψίζει τις βασικές μετρικές απόδοσης κάθε μοντέλου (όπως RMSE, MAE, MAPE) μαζί με τα στατιστικά τους χαρακτηριστικά (AIC, BIC) και τη δομή τους (π.χ. τάξη ARIMA/SARIMA).

Ο πίνακας αυτός λειτουργεί ως εργαλείο σύγκρισης και τεκμηρίωσης της καταλληλότητας των μοντέλων, καθιστώντας ορατό ποιοι συνδυασμοί παραμέτρων αποδίδουν καλύτερα ανά είδος δεδομένων και μετασχηματισμό (raw ή log). Μέσα από την παράθεση τόσο των αποτελεσμάτων του `auto.arima()` όσο και της χειροκίνητης προσέγγισης με `grid search`, επιτυγχάνεται πλήρης διαφάνεια στην επιλογή του τελικού μοντέλου, καθώς αναδεικνύονται περιπτώσεις όπου η αυτόματη προσέγγιση υπερτερεί και άλλες όπου η χειροκίνητη παρέμβαση και η ευρετική αναζήτηση οδηγούν σε ανώτερη προσαρμογή.

Πίνακας 6.2.6-1: Αποτελέσματα Καλύτερων Μοντέλων Ανάλογα με τον Τρόπο Fitting για το Εβδομαδιαίο Dataset Ατυχημάτων και Θανάτων

Model / Fitting	Best Model from Manual Fitting	Best Model from Auto.Arima()	Best Model from Grid Search
ARIMA on Regular Accidents	ARIMA(3,1,2)	ARIMA(4,1,1)	No GS Results
ARIMA on Regular Fatalities	ARIMA(1,1,2)	ARIMA(4,1,1) with Drift	No GS Results
ARIMA on Log Accidents	ARIMA(2,1,1)	ARIMA(4,1,1) with Drift	No GS Results
ARIMA on Log Fatalities	ARIMA(2,1,1)	ARIMA(4,1,1) with Drift	No GS Results
SARIMA on Regular Accidents	SARIMA(2,1,1)(0,1,1) [48]	SARIMA(1,0,0)(0,1,1) [48]	No GS Results
SARIMA on Regular Fatalities	SARIMA(3,1,1)(1,1,2) [48]	SARIMA(0,1,3)(1,0,0) [48] with Drift	SARIMA(3,0,3)(0,1,1) [48]
SARIMA on Log Accidents	SARIMA(2,1,1)(0,1,1) [48]	SARIMA(0,0,3)(2,1,0) [48] with Drift	No GS Results
SARIMA on Log Fatalities	SARIMA(1,1,2)(1,1,0) [48]	SARIMA(0,1,2)(1,0,0) [48] with Drift	SARIMA(3,0,3)(0,1,3) [48]

Πίνακας 6.2.6-2: Αποτελέσματα Καλύτερων Μοντέλων Ανάλογα με τον Τρόπο Fitting για το Μηνιαίο Dataset Ατυχημάτων και Θανάτων

Model / Fitting	Best Model from Manual Fitting	Best Model from Auto.Arima()	Best Model from Grid Search
ARIMA on Regular Accidents	ARIMA(3,1,2)	ARIMA(3,1,1) with Drift	No GS Results
ARIMA on Regular Fatalities	ARIMA(3,1,3)	ARIMA(3,1,3) with Drift	No GS Results
ARIMA on Log Accidents	ARIMA(2,1,1)	ARIMA(3,1,2) with Drift	No GS Results
ARIMA on Log Fatalities	ARIMA(3,1,1)	ARIMA(3,1,2) with Drift	No GS Results
SARIMA on Regular Accidents	SARIMA(1,0,2)(0,1,1) [12]	SARIMA(1,0,2)(0,1,1) [12]	SARIMA(3,0,3)(0,1,2) [12]
SARIMA on Regular Fatalities	SARIMA(1,0,1)(1,1,1) [12]	SARIMA(1,0,1)(1,1,1) [12] with Drift	No GS Results
SARIMA on Log Accidents	SARIMA(2,1,1)(0,1,1) [12]	SARIMA(1,0,2)(0,1,1) [12]	SARIMA(1,0,2)(1,0,3) [12]
SARIMA on Log Fatalities	SARIMA(2,1,1)(1,1,1) [12]	SARIMA(0,1,1)(1,1,1) [12]	SARIMA(1,0,2)(1,1,1) [12]

Πίνακας 6.2.6-3: Πίνακας Μετρικών Τελικών Μοντέλων ARIMA / SARIMA Τροχαίων Ατυχημάτων από το Εβδομαδιαίο Dataset

<u>Manual Model / Metrics</u>	<u>MAPE</u>	<u>Residual Check</u> <u>Pass</u>	<u>Grid Search</u> <u>Time</u>
ARIMA(3,1,2)	43.21%	No	-
Log ARIMA(2,1,1)	37.28%	No	-
SARIMA(2,1,1)(0,1,1) [48]	20.68%	No	-
Log SARIMA(2,1,1)(0,1,1) [48]	18.72%	No	-
<u>Auto.Arima() Model / Metrics</u>	<u>MAPE</u>	<u>Residual Check</u> <u>Pass</u>	-
ARIMA(4,1,1)	36.6%	No	-
Log ARIMA(4,1,1) with Drift	34.06%	No	-
SARIMA(1,0,0)(0,1,1) [48]	21.06%	No	-
Log SARIMA(0,0,3)(2,1,0) [48] with Drift	20.43%	No	-
<u>Grid Search Model / Metrics</u>	<u>MAPE</u>	<u>Residual Check</u> <u>Pass</u>	-
No GS Results for ARIMA Models	-	-	~ 1 min
No GS Results for ARIMA Models	-	-	~ 1 min
No GS Results for SARIMA Models	-	-	~ 220 hours
No GS Results for SARIMA Models	-	-	~ 220 hours

Ο Πίνακας 6.2.6-3 που παρουσιάζει τα αποτελέσματα των μοντέλων ARIMA και SARIMA για τα εβδομαδιαία δεδομένα τροχαίων ατυχημάτων αναδεικνύει τις προκλήσεις που πρέκυσαν κατά την προσπάθεια εξισορρόπησης μεταξύ καλής προβλεπτικής απόδοσης και στατιστικής επάρκειας. Παρότι αρκετά μοντέλα, ιδίως τα εποχικά SARIMA και οι λογαριθμικά μετασχηματισμένες εκδοχές τους παρουσίασαν σημαντικά βελτιωμένες τιμές MAPE, κανένα από αυτά δεν κατάφερε να περάσει επιτυχώς τα residuals checks, όπως το τεστ Ljung–Box και οπτική επιθεώρηση των ACF/PACF των residuals. Ενδεικτικά, το μοντέλο Log SARIMA(2,1,1)(0,1,1)[48] προσέφερε την καλύτερη τιμή MAPE (18.72%), επιβεβαιώνοντας τη χρησιμότητα του log μετασχηματισμού και της εποχικής συνιστώσας, ωστόσο η μη αποδοχή των residuals του, το καθιστά μη αποδεκτό για τελική χρήση. Αντίστοιχα, τα αποτελέσματα

του `auto.arima()` προσέγγισαν συγκρίσιμες επιδόσεις σε μοντέλα όπως το Log ARIMA(4,1,1) με MAPE 34.06%, αλλά και πάλι χωρίς να πληρούν τα κριτήρια των residuals checks.

Πίνακας 6.2.6-4: Πίνακας Μετρικών Τελικών Μοντέλων ARIMA / SARIMA Θανάτων από Τροχαία Ατυχήματα από το Εβδομαδιαίο Dataset

<u>Manual Model / Metrics</u>	<u>MAPE</u>	<u>Residual Check</u> <u>Pass</u>	<u>Grid Search</u> <u>Time</u>
ARIMA(1,1,2)	56.08%	No	-
Log ARIMA(2,1,1)	48.94%	No	-
SARIMA(3,1,1)(1,1,2) [48]	36.21%	No	-
Log SARIMA(1,1,2)(1,1,0) [48]	33.02%	No	-
<u>Auto.Arima() Model / Metrics</u>	<u>MAPE</u>	<u>Residual Check</u> <u>Pass</u>	-
ARIMA(4,1,1) with Drift	52.66%	No	-
Log ARIMA(4,1,1) with Drift	46.00%	No	-
SARIMA(0,1,3)(1,0,0) [48] with Drift	41.49%	No	-
Log SARIMA(0,1,2)(1,0,0) [48] with Drift	37.1%	No	-
<u>Grid Search Model / Metrics</u>	<u>MAPE</u>	<u>Residual Check</u> <u>Pass</u>	-
No GS Results for ARIMA Models	-	-	~ 1 min
No GS Results for ARIMA Models	-	-	~ 1 min
SARIMA(3,0,3)(0,1,1) [48]	32.35%	Yes	~ 220 hours
Log SARIMA(3,0,3)(0,1,3) [48]	29.92%	Yes	~ 220 hours

Ο Πίνακας 6.2.6-4 συγκεντρώνει τα αποτελέσματα των μοντέλων ARIMA και SARIMA που εφαρμόστηκαν στα εβδομαδιαία δεδομένα θανάτων από τροχαία ατυχήματα. Παρατηρείται και εδώ ότι τα χειροκίνητα μοντέλα (manual ARIMA/SARIMA) και τα αυτόματα (`auto.arima()`) παρήγαγαν τιμές MAPE που, αν και σε αρκετές περιπτώσεις χαμηλότερες σε σχέση με τα ατυχήματα, εξακολουθούσαν να συνοδεύονται από αποτυχία στα τεστ διαγνωστικού ελέγχου των residuals. Ιδιαίτερα οι εποχικές παραλλαγές SARIMA στα log-

transformed δεδομένα εμφάνισαν αισθητά βελτιωμένες επιδόσεις σε MAPE (π.χ. 33.02% για το Log SARIMA(1,1,2)(1,1,0)[48]), όμως δεν πληρούσαν τα κριτήρια των residual checks. Σε αντίθεση με τον πίνακα των ατυχημάτων, εδώ το grid search απέφερε σημαντικά αποτελέσματα. Τα μοντέλα SARIMA(3,0,3)(0,1,1)[48] και Log SARIMA(3,0,3)(0,1,3)[48] πέρασαν επιτυχώς το residual check και ταυτόχρονα σημείωσαν τις καλύτερες προβλεπτικές επιδόσεις με MAPE 32.70% και 29.92% αντίστοιχα.

Grid search complete! Top models by MAPE:

	order	seasonal	AIC	BIC	MAPE	LB_p	BP_p
1	(3,0,3)	(0,1,1)[48]	6997.986	7037.160	32.35594	0.07255321	0.07491951
2	(2,1,3)	(1,1,0)[48]	7135.821	7170.091	34.85299	0.30814249	0.31376808
3	(3,0,3)	(0,1,0)[48]	7393.087	7427.364	35.72839	0.12461201	0.12766383
4	(2,1,3)	(0,1,2)[48]	6994.847	7034.012	36.25517	0.06866831	0.07093844
5	(2,1,3)	(0,1,1)[48]	6992.961	7027.230	36.52624	0.11988562	0.12335914
6	(2,1,3)	(0,1,0)[48]	7388.941	7418.315	37.51760	0.23804891	0.24216454

Σχήμα 6.2.6-1: Αποτελέσματα Grid Search για Fitting Μοντέλων ARIMA / SARIMA με Πρόβλεψη τους Θανάτους από Τροχαία Ατυχήματα (Regular Fatalities Time Series)

Grid search complete! Top models by MAPE:

	order	seasonal	AIC	BIC	MAPE	LB_p	BP_p
1	(3,0,3)	(0,1,3)[48]	236.3874	285.3544	28.92729	0.06605620	0.06745427
2	(1,0,1)	(0,1,3)[48]	236.1287	265.5088	29.02345	0.21146267	0.21641875
3	(1,0,3)	(0,1,3)[48]	239.4797	278.6532	29.02408	0.12427378	0.12763345
4	(3,0,2)	(0,1,3)[48]	240.3169	284.3871	29.02416	0.05684274	0.05866491
5	(3,0,1)	(0,1,3)[48]	239.4676	278.6411	29.02468	0.12328224	0.12663040
6	(1,0,2)	(0,1,3)[48]	238.1179	272.3947	29.02574	0.14148403	0.14525279

Σχήμα 6.2.6-2: Αποτελέσματα Grid Search για Fitting Μοντέλων ARIMA / SARIMA με Πρόβλεψη τους Θανάτους από Τροχαία Ατυχήματα (Log Fatalities Time Series)

Πίνακας 6.2.6-5: Πίνακας Μετρικών Τελικών Μοντέλων ARIMA / SARIMA Τροχαίων Ατυχημάτων από το Μηνιαίο Dataset

<u>Manual Model / Metrics</u>	<u>MAPE</u>	<u>Residual Check</u> <u>Pass</u>	<u>Grid Search</u> <u>Time</u>
ARIMA(3,1,2)	14.55%	No	-
Log ARIMA(2,1,1)	19.03%	No	-
SARIMA(1,0,2)(0,1,1) [12]	12.40%	No	-
Log SARIMA(2,1,1)(0,1,1) [12]	14.33%	Yes	-
<u>Auto.Arima() Model / Metrics</u>	<u>MAPE</u>	<u>Residual Check</u> <u>Pass</u>	-
ARIMA(3,1,1) with Drift	18.69%	No	-
Log ARIMA(3,1,2) with Drift	17.71%	No	-
SARIMA(1,0,2)(0,1,1) [12]	12.49%	No	-
Log SARIMA(1,0,2)(0,1,1) [12]	12.56%	Yes	-
<u>Grid Search Model / Metrics</u>	<u>MAPE</u>	<u>Residual Check</u> <u>Pass</u>	-
No GS Results for ARIMA Models	-	-	~ 1 min
No GS Results for ARIMA Models	-	-	~ 1 min
SARIMA(3,0,3)(0,1,2) [12]	12.30%	Yes	~ 48 hours
Log SARIMA(1,0,2)(1,0,3) [12]	12.09%	Yes	~ 48 hours

Ο Πίνακας 6.2.6-5 που αφορά τα μηνιαία δεδομένα τροχαίων ατυχημάτων παρουσιάζει σημαντικά βελτιωμένες προβλεπτικές επιδόσεις σε σχέση με τα εβδομαδιαία, κάτι που αποδίδεται στην ομαλότερη φύση της χρονοσειράς και στη χαμηλότερη παρουσία θορύβου. Αν και αρκετά μοντέλα από τα manual και auto.arima απέτυχαν να περάσουν το residual check (παρά τις αποδεκτές τιμές MAPE), παρατηρείται ότι δύο SARIMA μοντέλα, το Log SARIMA(2,1,1)(0,1,1)[12] από τα χειροκίνητα και το αντίστοιχο auto.arima() μοντέλο, κατάφεραν να συνδυάσουν ικανοποιητικές μετρικές (14.33% και 12.56% αντίστοιχα) με στατιστικά αποδεκτά residuals. Το πιο εύρημα όμως που προέρχεται από το στάδιο του grid search, αποκάλυψε δύο μοντέλα με σαφώς καλύτερη απόδοση, το SARIMA(3,0,3)(0,1,2)[12] και το Log SARIMA(1,0,2)(1,0,3)[12], με MAPE 12.30% και 12.09% αντίστοιχα και πλήρη επιτυχία στον έλεγχο των residuals.

Πίνακας 6.2.6-6 Πίνακας Μετρικών Τελικών Μοντέλων ARIMA / SARIMA Θανάτων από Τροχαία Ατυχήματα από το Μηνιαίο Dataset

<u>Manual Model / Metrics</u>	<u>MAPE</u>	<u>Residual Check</u> <u>Pass</u>	<u>Grid Search</u> <u>Time</u>
ARIMA(3,1,3)	24.71%	No	-
Log ARIMA(3,1,1)	31.48%	No	-
SARIMA(1,0,1)(1,1,1) [12]	16.59%	No	-
Log SARIMA(2,1,1)(1,1,1) [12]	17.33%	Yes	-
<u>Auto.Arima() Model / Metrics</u>	<u>MAPE</u>	<u>Residual Check</u> <u>Pass</u>	-
ARIMA(3,1,1) with Drift	28.02%	No	-
Log ARIMA(3,1,2) with Drift	21.28%	No	-
SARIMA(1,0,1)(1,1,1) [12] with Drift	21.62%	No	-
Log SARIMA(0,1,1)(1,1,1) [12]	17.34%	Yes	-
<u>Grid Search Model / Metrics</u>	<u>MAPE</u>	<u>Residual Check</u> <u>Pass</u>	-
No GS Results for ARIMA Models	-	-	~ 1 min
No GS Results for ARIMA Models	-	-	~ 1 min
No GS Results for SARIMA Models	-	-	~ 48 hours
Log SARIMA(1,0,2)(1,1,1) [48]	17.14%	Yes	~ 48 hours

Ο Πίνακας 6.2.6-6, που αφορά τη μοντελοποίηση των μηνιαίων θανάτων από τροχαία ατυχήματα, αναδεικνύει μια πιο ισορροπημένη εικόνα σε σχέση με τα υπόλοιπα datasets. Παρότι αρκετά μοντέλα, είτε χειροκίνητα είτε αυτόματα, παρουσίασαν σχετικά αποδεκτές τιμές MAPE, οι περισσότεροι συνδυασμοί απέτυχαν στον έλεγχο των residuals. Ωστόσο, αξιοσημείωτη είναι η απόδοση του Log SARIMA(2,1,1)(1,1,1)[12], το οποίο αποτελεί ένα από τα λίγα μοντέλα που συνδυάζει ικανοποιητική προβλεπτική ακρίβεια (17.33%) με στατιστικά αποδεκτά residuals, υποδηλώνοντας καλή συνολική προσαρμογή. Το καλύτερο αποτέλεσμα προήλθε από το grid search, όπου το Log SARIMA(1,0,2)(1,1,1)[48] πέτυχε τον χαμηλότερο MAPE (17.14%) διατηρώντας παράλληλα στατιστική εγκυρότητα.

6.4 Τελική Εκτίμηση των Στατιστικών Μοντέλων και Προβλέψεις

Η τελική φάση της ανάλυσης περιλάμβανε την επιλογή των βέλτιστων μοντέλων ARIMA και SARIMA για κάθε χρονοσειρά, με βάση τη συνολική τους απόδοση στις προηγούμενες αξιολογήσεις. Τα μοντέλα που επικράτησαν ήταν εκείνα που παρουσίασαν τον καλύτερο συνδυασμό μεταξύ χαμηλών σφαλμάτων πρόβλεψης (MAPE, RMSE, MAE) και αποδεκτών διαγνωστικών ελέγχων για τα residuals. Για κάθε χρονικό βάθος (εβδομαδιαίο ή μηνιαίο) και κάθε στόχο (ατυχήματα ή θάνατοι), επιλέχθηκαν τελικά 2 (ARIMA και SARIMA) μοντέλα που πληρούσαν τόσο κριτήρια στατιστικής επάρκειας όσο και προβλεπτικής αξιοπιστίας.

Με τα επιλεγμένα μοντέλα εκπαιδευμένα στο σύνολο των διαθέσιμων ιστορικών δεδομένων (1996–2022), πραγματοποιήθηκαν επεκταμένες προβλέψεις για την περίοδο 2023–2026. Οι προβλέψεις αυτές, σε εβδομαδιαία και μηνιαία βάση, αποτυπώνουν την αναμενόμενη πορεία των ατυχημάτων και των θανάτων στο μέλλον υπό την προϋπόθεση της συνέχισης των ίδιων τάσεων και μοτίβων.

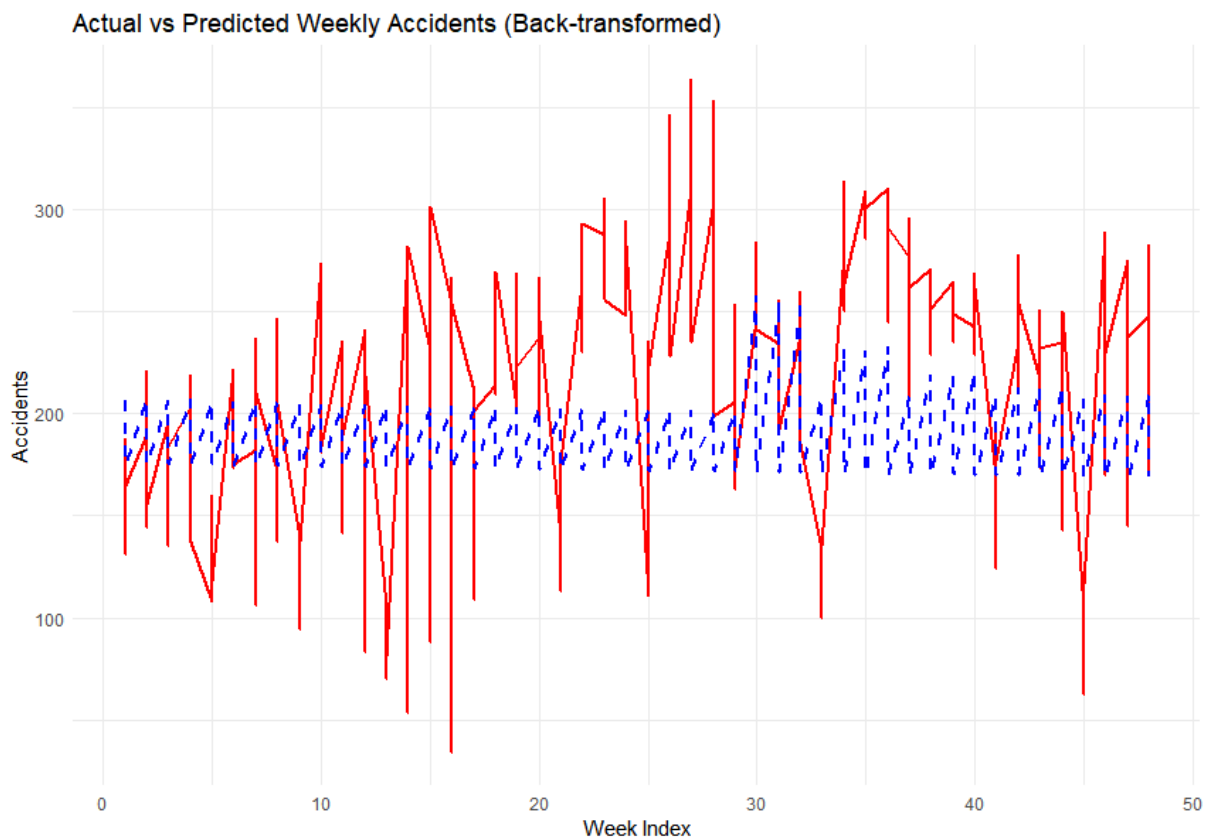
6.4.1 ARIMA ΜΟΝΤΕΛΟ ΓΙΑ ΠΡΟΒΛΕΨΗ ΕΒΔΟΜΑΔΙΑΙΩΝ ΤΡΟΧΑΙΩΝ ΑΤΥΧΗΜΑΤΩΝ

Το ARIMA μοντέλο που εφαρμόστηκε στα εβδομαδιαία δεδομένα τροχαίων ατυχημάτων ακολουθεί μια δομημένη διαδικασία εκπαίδευσης και αξιολόγησης. Η χρονοσειρά αρχικά λογαριθμοποιήθηκε (log transformation) ώστε να εξομαλυνθεί η διακύμανση και να ενισχυθεί η σταθερότητά της. Χωρίστηκε σε training (1996–2017) και test set (2017–2022), πάνω στα οποία εφαρμόστηκε ARIMA(4,1,1) με drift. Η πρόβλεψη που προέκυψε για το test set απομετασχηματίστηκε μέσω της συνάρτησης `expm1()` ώστε να συγκριθεί με τις πραγματικές παρατηρήσεις. Η προβλεπτική ακρίβεια του μοντέλου αξιολογήθηκε με RMSE, MAE και MAPE, προσφέροντας μια ολοκληρωμένη εικόνα απόδοσης τόσο σε απόλυτο όσο και σχετικό επίπεδο. Παράλληλα, πραγματοποιήθηκε fitting του ίδιου μοντέλου πάνω σε ολόκληρο το dataset και προβλέφθηκαν οι τιμές έως το τέλος του 2026.

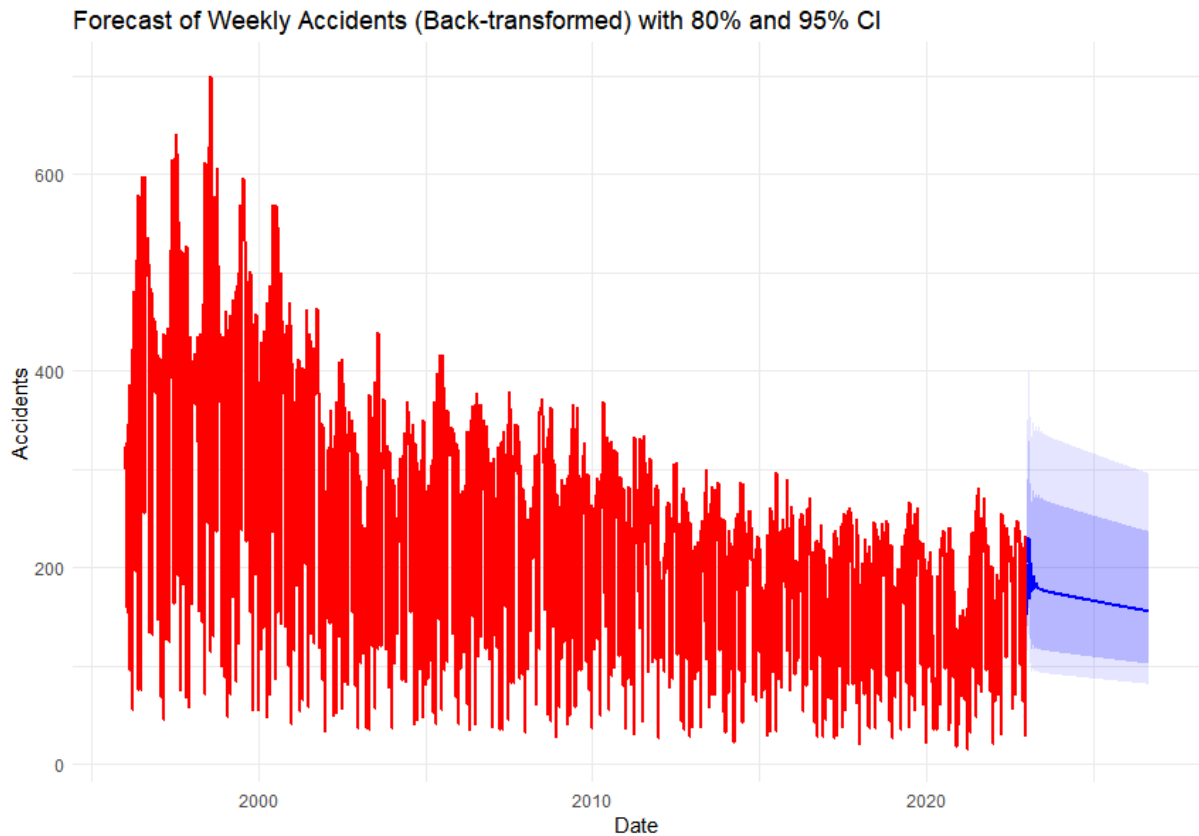
Τα αποτελέσματα των προβλέψεων απεικονίζονται γραφικά με τις πραγματικές τιμές να εμφανίζονται με κόκκινο και τις προβλεπόμενες με μπλε διακεκομμένη γραμμή. Επιπλέον, η πλήρης πρόβλεψη μέχρι το 2026 συνοδεύεται από εύρη εμπιστοσύνης 80% και 95%, τα οποία υποδεικνύουν τη στατιστική αβεβαιότητα γύρω από τις μελλοντικές τιμές. Τέλος, πραγματοποιήθηκε διαγνωστικός έλεγχος στα residuals του τελικού μοντέλου μέσω του τεστ

Ljung–Box, του Box-Pierce και της συνάρτησης `checkresiduals()`, όπου εντοπίζεται ότι, παρά τη λογική κατανομή, ορισμένα επίπεδα αυτοσυσχέτισης ενδέχεται να επιμένουν, περιορίζοντας την πλήρη στατιστική επάρκεια του μοντέλου. Παρ' όλα αυτά, η συνολική απόδοση κρίνεται ικανοποιητική για μεσοπρόθεσμες προβλέψεις.

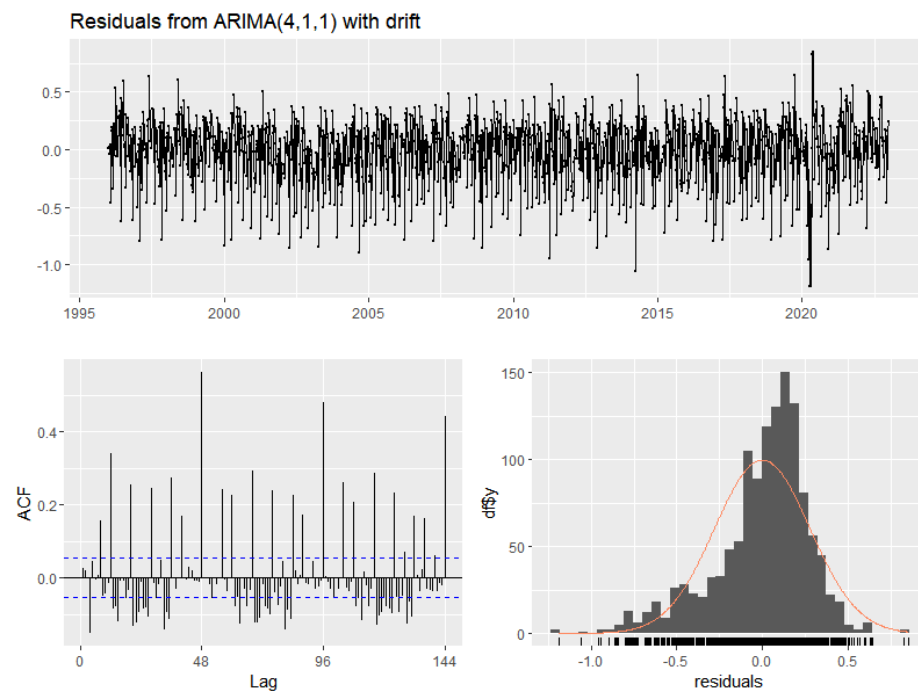
Όμως, το μοντέλο $ARIMA(4,1,1)$, εμφάνισε χαμηλή επίδοση στις υπολογιζόμενες μετρικές. Οι τιμές σφάλματος $RMSE = 69.03$ και $MAE = 58.49$ δείχνουν σημαντικές αποκλίσεις των προβλέψεων από τις πραγματικές τιμές, ενώ ο δείκτης $MAPE$ έφτασε το 32.23%, καθιστώντας το μοντέλο συγκριτικά λιγότερο ακριβές. Παράλληλα, οι τιμές $AIC = 249.70$ και $BIC = 284.30$ είναι αρκετά υψηλές, υποδηλώνοντας λιγότερο αποδοτική προσαρμογή και αυξημένη πολυπλοκότητα χωρίς ανάλογη βελτίωση στην ακρίβεια. Τα αποτελέσματα αυτά ενισχύουν την ανάγκη ενσωμάτωσης της εποχικότητας σε εβδομαδιαίες χρονοσειρές, καθώς η απουσία εποχικής παραμετροποίησης φαίνεται να περιορίζει αισθητά τη δυνατότητα του μοντέλου να αποδώσει τις διακυμάνσεις των δεδομένων και να προσφέρει αξιόπιστες προβλέψεις.



Σχήμα 6.4.1-1: Actual vs Predicted Τιμές Εβδομαδιαίων Τροχαίων Ατυχημάτων του Μοντέλου $ARIMA(4,1,1)$ with Drift



Σχήμα 6.4.1-2: Προβλέψεις του Μοντέλου ARIMA(4,1,1) για Εβδομαδιαία Τροχαία Ατυχήματα έως το 2026



Σχήμα 6.4.1-3: Διαγνωστικά των Residuals του Μοντέλου ARIMA(4,1,1) σε Εβδομαδιαία Τροχαία Ατυχήματα

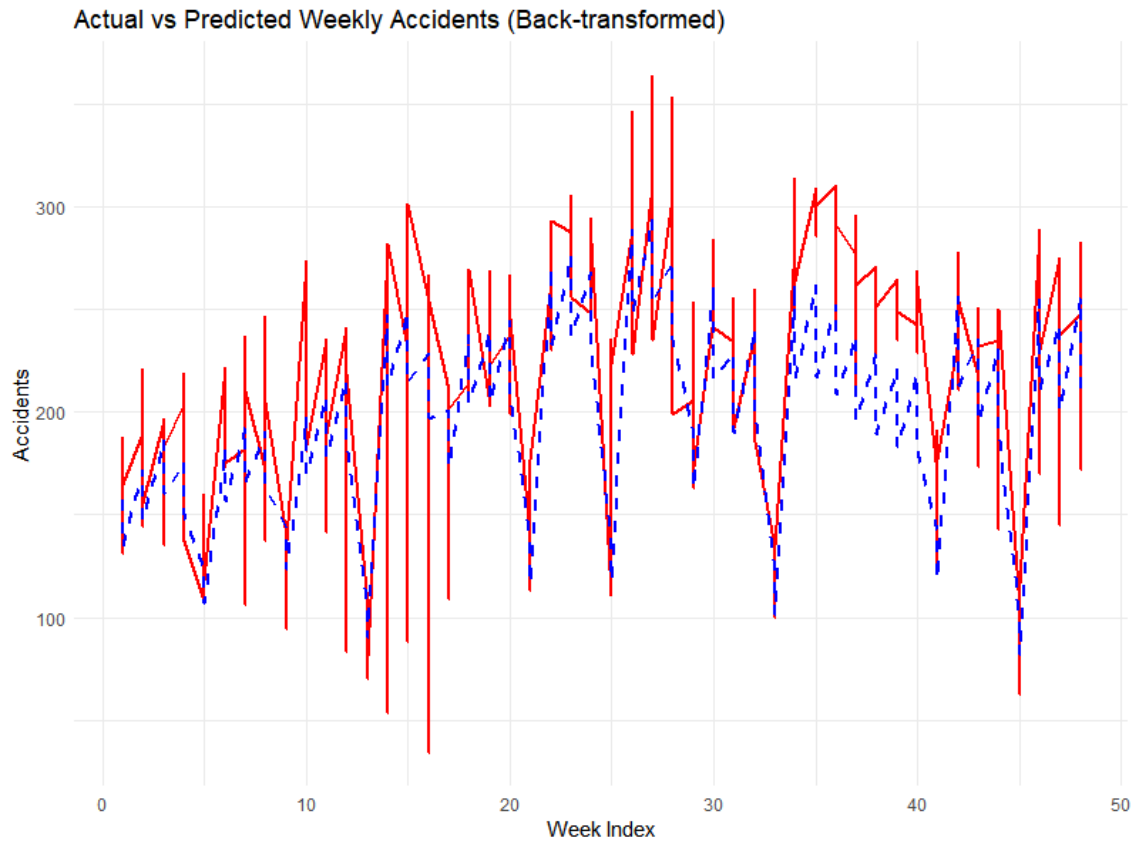
6.4.2 SARIMA ΜΟΝΤΕΛΟ ΓΙΑ ΠΡΟΒΛΕΨΗ ΕΒΔΟΜΑΔΙΑΙΩΝ ΤΡΟΧΑΙΩΝ ΑΤΥΧΗΜΑΤΩΝ

Η εφαρμογή του εποχικού μοντέλου SARIMA για την πρόβλεψη εβδομαδιαίων τροχαίων ατυχημάτων ανέδειξε σημαντικά χαρακτηριστικά της χρονοσειράς που σχετίζονται με την περιοδικότητα των δεδομένων. Το επιλεγμένο μοντέλο SARIMA(2,1,1)(0,1,1)[48], ενσωματώνει τόσο μη εποχικές όσο και εποχικές διαφορές, ανταποκρινόμενο έτσι στην ύπαρξη ετήσιων κύκλων. Η απόδοσή του αξιολογήθηκε πρώτα στο test set, στο οποίο κατέγραψε ικανοποιητικές επιδόσεις με χαμηλές τιμές σφαλμάτων όπως RMSE και MAPE, ενώ ταυτόχρονα εμφάνισε αξιοσημείωτη σταθερότητα. Όπως καταγράφεται στο διάγραμμα Actual vs Predicted, η πρόβλεψη ευθυγραμμίζεται ικανοποιητικά με την παρατηρηθείσα διακύμανση της σειράς, επιδεικνύοντας επαρκή ικανότητα παρακολούθησης της βραχυπρόθεσμης δυναμικής. Το σφάλμα πρόβλεψης παρουσιάζεται ομαλό και χωρίς ισχυρές αποκλίσεις, υποδεικνύοντας ότι το μοντέλο έχει καταφέρει να αποτυπώσει τόσο τη δομή όσο και τον εποχικό ρυθμό της σειράς.

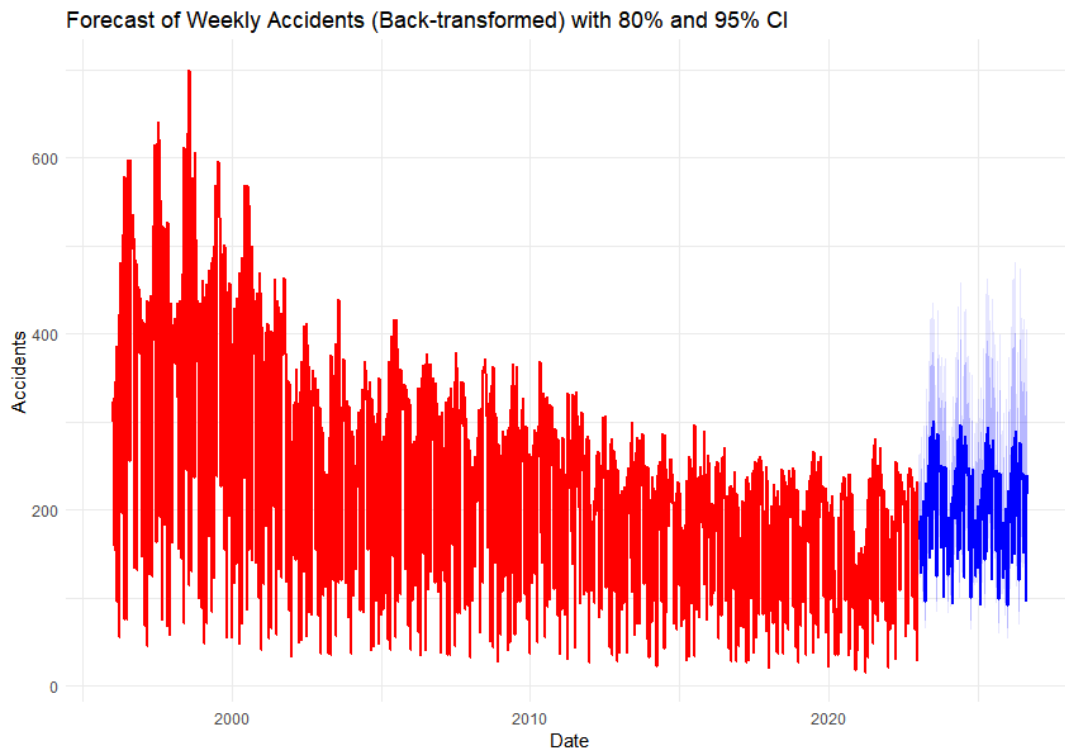
Με την εφαρμογή του τελικού μοντέλου στο πλήρες σύνολο δεδομένων και την πρόβλεψη των τιμών μέχρι το τέλος του 2026, προκύπτει μια συνολική εικόνα της εκτιμώμενης μελλοντικής συμπεριφοράς. Η προβλεπτική καμπύλη υποδεικνύει διατήρηση της ετήσιας εποχικότητας, με κυκλικές μεταβολές στην ένταση των ατυχημάτων και εύρη αβεβαιότητας που αποτυπώνονται στις λωρίδες εμπιστοσύνης 80% και 95%. Η περιοδικότητα της πρόβλεψης είναι ιδιαίτερα ευδιάκριτη και ενισχύει την αξιοπιστία του μοντέλου για μεσοπρόθεσμες προβλέψεις. Παράλληλα, ο διαγνωστικός έλεγχος των residuals μέσω των τεστ Ljung–Box και Box–Pierce, καθώς και της συνάρτησης checkresiduals(), επιβεβαίωσε την κανονικότητα, την απουσία συστηματικών residuals και την ανεξαρτησία των σφαλμάτων.

Επίσης, τα αποτελέσματα αξιολόγησης του μοντέλου SARIMA(2,1,1)(0,1,1)[48] για τα εβδομαδιαία δεδομένα τροχαίων ατυχημάτων επιβεβαιώνουν την εξαιρετική του απόδοση σε επίπεδο ακρίβειας πρόβλεψης. Συγκεκριμένα, το μοντέλο παρουσίασε πολύ χαμηλό σφάλμα με RMSE ίσο με 43.72 και MAE ίσο με 33.31, τιμές που καταδεικνύουν μικρές αποκλίσεις μεταξύ των πραγματικών και των προβλεπόμενων τιμών. Το ποσοστιαίο σφάλμα MAPE ανήλθε σε μόλις 18.73%, ποσοστό ικανοποιητικό για δεδομένα υψηλής μεταβλητότητας, όπως αυτά των εβδομαδιαίων ατυχημάτων. Παράλληλα, οι τιμές AIC = -580.52 και BIC = -556.04 αποτυπώνουν την προσαρμοστική ικανότητα του μοντέλου χωρίς υπερβολική πολυπλοκότητα. Συνολικά, οι μετρικές επιβεβαιώνουν ότι το συγκεκριμένο μοντέλο όχι μόνο καταφέρνει να

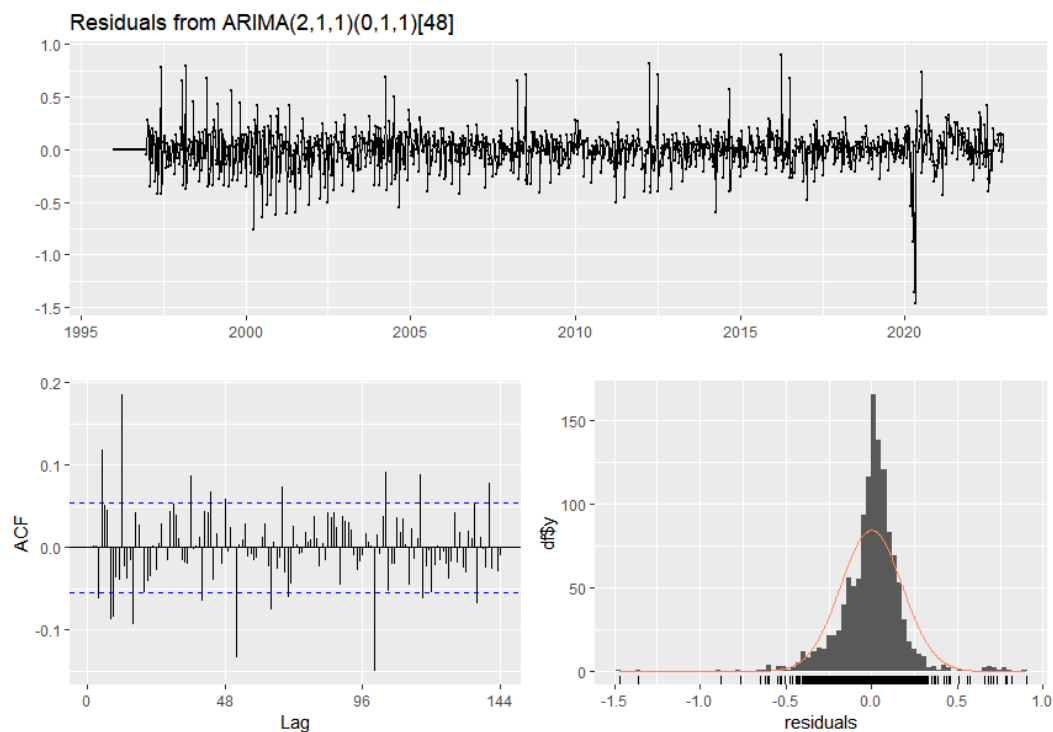
αποδώσει ικανοποιητικά τη δομή της χρονοσειράς, αλλά το κάνει και με τρόπο στατιστικά ορθό και σταθερό.



Σχήμα 6.4.2-1: Actual vs Predicted Τιμές Εβδομαδιαίων Τροχαίων Ατυχημάτων του Μοντέλου SARIMA(2,1,1)(0,1,1)[48]



Σχήμα 6.4.2-2: Προβλέψεις του Μοντέλου SARIMA(2,1,1)(0,1,1)[48] για Εβδομαδιαία Τροχαία Ατυχήματα έως το 2026



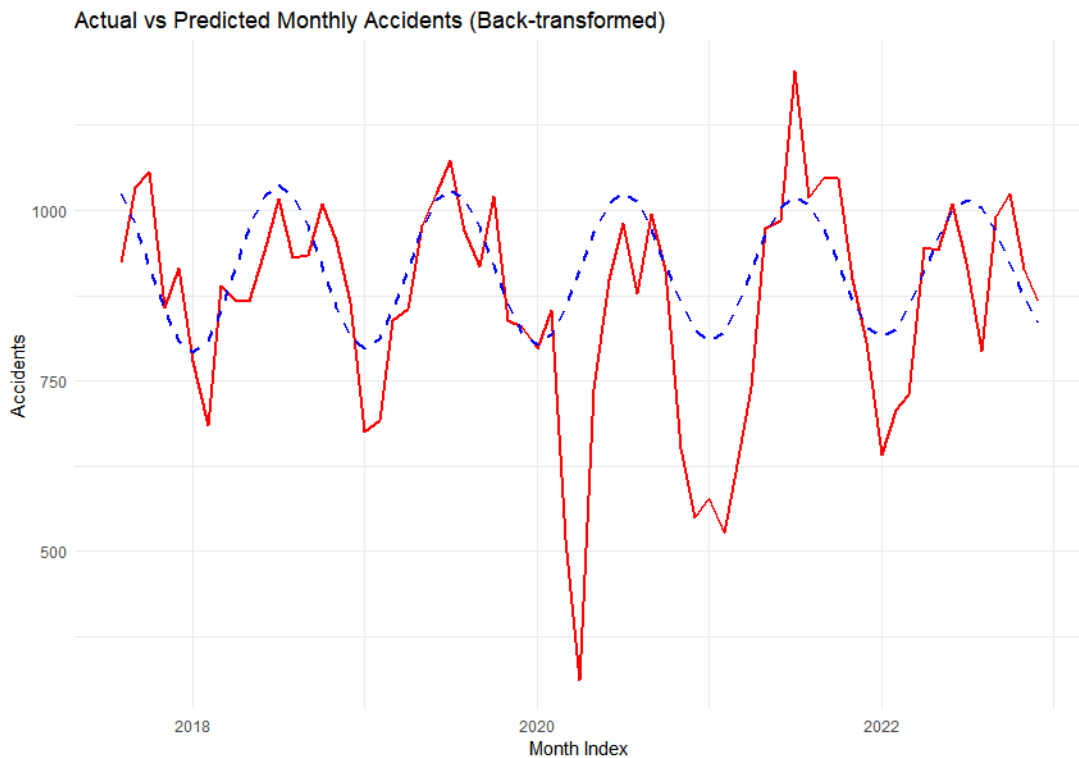
Σχήμα 6.4.2-3: Διαγνωστικά των Residuals του Μοντέλου SARIMA(2,1,1)(0,1,1)[48] σε Εβδομαδιαία Τροχαία Ατυχήματα

6.4.3 ARIMA ΜΟΝΤΕΛΟ ΓΙΑ ΠΡΟΒΛΕΨΗ ΜΗΝΙΑΙΩΝ ΤΡΟΧΑΙΩΝ ΑΤΥΧΗΜΑΤΩΝ

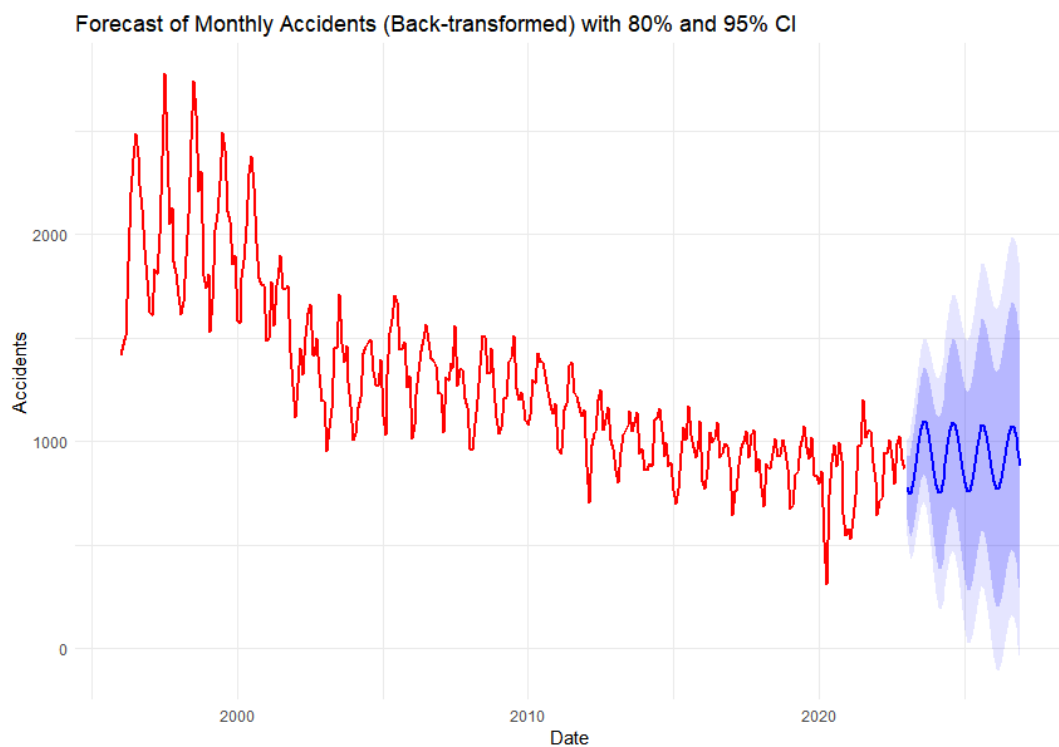
Οι μηνιαίες προβλέψεις των τροχαίων ατυχημάτων μέσω του μοντέλου ARIMA(3,1,2) παρουσιάζουν ιδιαίτερο ενδιαφέρον τόσο ως προς την ικανότητα προσαρμογής στα ιστορικά δεδομένα όσο και ως προς την εποχική δυναμική που αποτυπώνεται στις προγνώσεις για την περίοδο 2023–2026. Όπως φαίνεται στο πρώτο διάγραμμα, η προβλεπόμενη πορεία διατηρεί τα κυρίαρχα περιοδικά χαρακτηριστικά των παλαιότερων ετών, επιδεικνύοντας σαφείς κυκλικές διακυμάνσεις, οι οποίες ενδέχεται να αντανakλούν εποχικούς παράγοντες, όπως αυξημένη κινητικότητα το καλοκαίρι ή περιόδους διακοπών. Η προσαρμογή της γραμμής πρόβλεψης στο ιστορικό σήμα θεωρείται επαρκής, παρά του ότι τα διαστήματα εμπιστοσύνης διαπλατώνονται καθώς απομακρυνόμαστε χρονικά.

Αναλύοντας τα διαγνωστικά των residuals του μοντέλου ARIMA(3,1,2), παρατηρείται ότι η κατανομή τους προσεγγίζει την κανονική και παρουσιάζει σχετικά περιορισμένη αυτοσυσχέτιση, ιδιαίτερα στα πρώτα lags. Η κατανομή είναι συμμετρική, χωρίς έντονη ασυμμετρία ή παρουσία outliers. Αν και κάποια σημεία υπερβαίνουν τα όρια εμπιστοσύνης στο ACF plot, δεν παρατηρούνται επίμονες αυτοσυσχετίσεις. Τέλος, στο διάγραμμα προβλεπόμενων έναντι πραγματικών τιμών, η συμπεριφορά του μοντέλου είναι σταθερά κοντά στις πραγματικές τιμές, διατηρώντας ικανοποιητική προβλεπτική συνέπεια ακόμη και σε περιόδους έντονης μεταβλητότητας, όπως η απότομη πτώση στις αρχές του 2020.

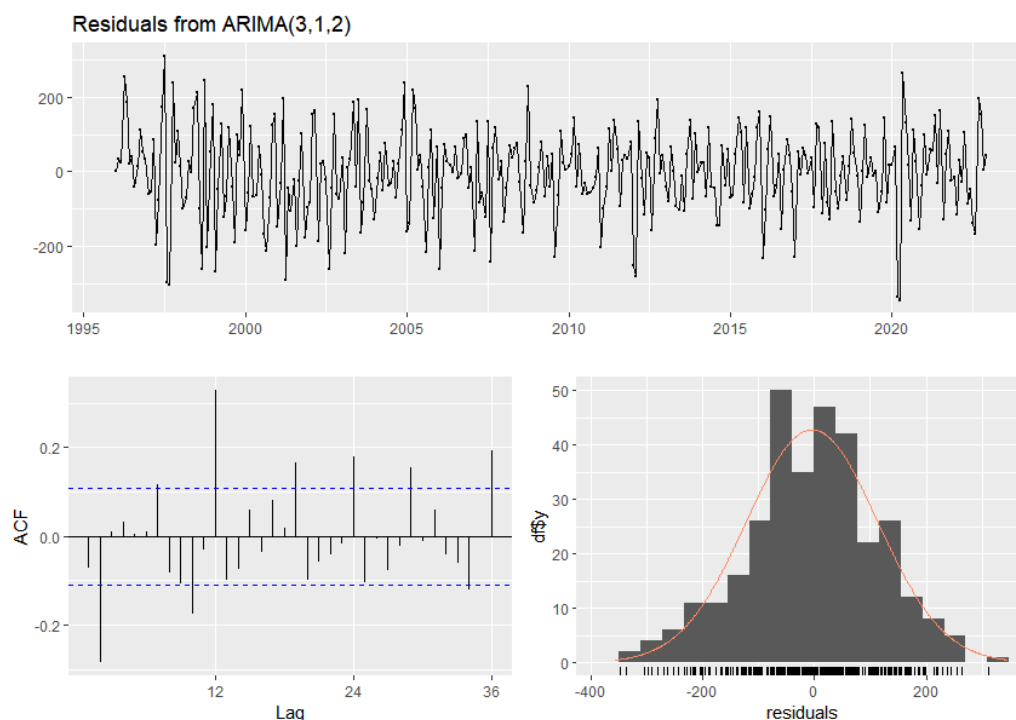
Η αξιολόγηση του μοντέλου επιβεβαιώνει ποσοτικά την επάρκειά του. Ο συντελεστής MAPE έφτασε το 14.55%, αντανakλώντας σχετικά χαμηλό σφάλμα πρόβλεψης ως ποσοστό των πραγματικών τιμών, ενώ οι τιμές RMSE και MAE ανήλθαν σε 139.03 και 95.90 αντίστοιχα, επιβεβαιώνοντας την αριθμητική συνέπεια των προβλέψεων. Αν και οι τιμές AIC (3210.96) και BIC (3232.28) δεν είναι οι χαμηλότερες μεταξύ όλων των μοντέλων που δοκιμάστηκαν, η συνολική απόδοση του ARIMA(3,1,2) σε επίπεδο μηνιαίας πρόβλεψης κρίνεται ικανοποιητική τόσο από στατιστική όσο και από πρακτική πλευρά, ιδιαίτερα εφόσον διατηρεί τη δυναμική της εποχικότητας και παρέχει σταθερές εκτιμήσεις σε βάθος χρόνου.



Σχήμα 6.4.3-1: Actual vs Predicted Τιμές Μηνιαίων Τροχαίων Ατυχημάτων του Μοντέλου $ARIMA(3,1,2)$



Σχήμα 6.4.3-2: Προβλέψεις του Μοντέλου $ARIMA(3,1,2)$ για Μηνιαία Τροχαία Ατυχήματα έως το 2026



Σχήμα 6.4.3-3: Διαγνωστικά των Residuals του Μοντέλου ARIMA(3,1,2) σε Μηνιαία Τροχαία Ατυχήματα

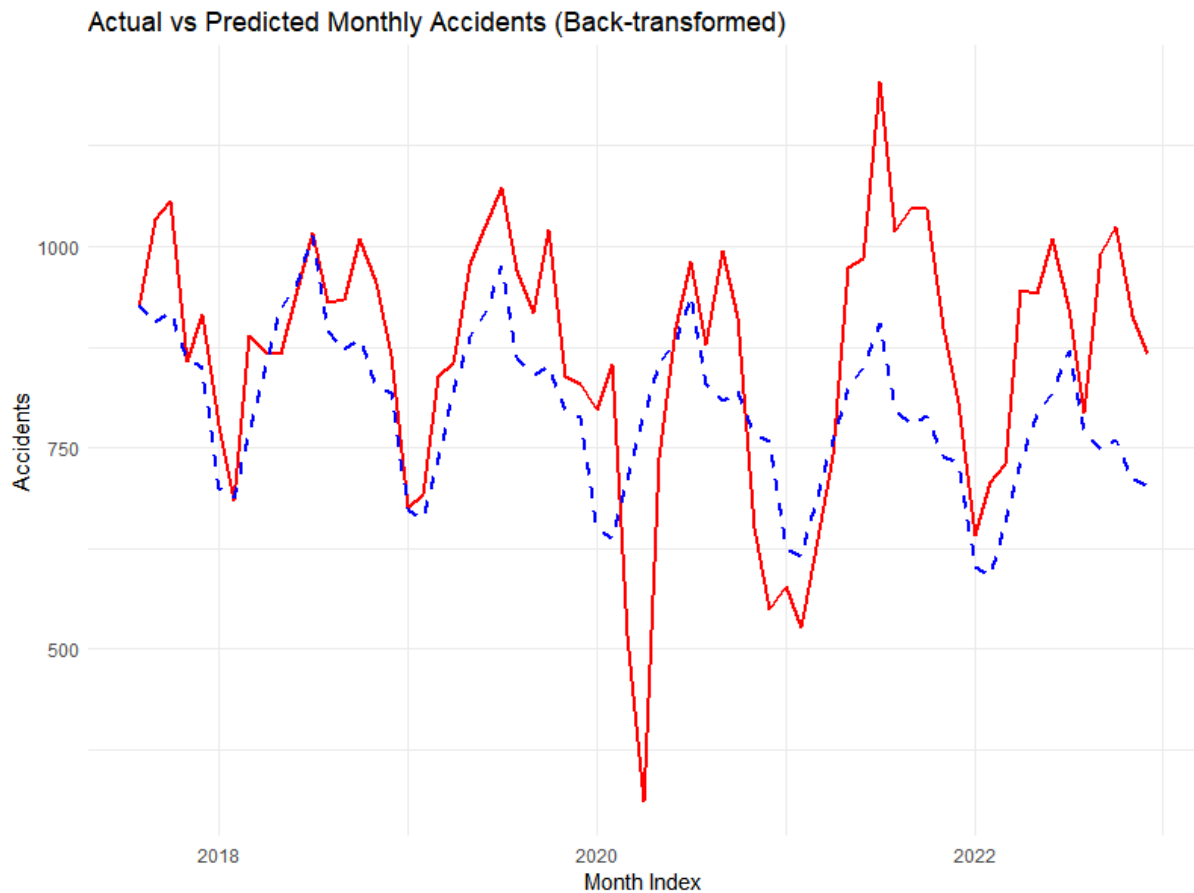
6.4.4 SARIMA ΜΟΝΤΕΛΟ ΓΙΑ ΠΡΟΒΛΕΨΗ ΜΗΝΙΑΙΩΝ ΤΡΟΧΑΙΩΝ ΑΤΥΧΗΜΑΤΩΝ

Ακολουθώντας τη διαδικασία μοντελοποίησης SARIMA για τα μηνιαία δεδομένα τροχαίων ατυχημάτων, χρησιμοποιήθηκε το μοντέλο SARIMA(2,1,1)(0,1,1)[12], το οποίο συνδυάζει μια ολοκληρωμένη μη εποχική συνιστώσα με μία εποχική διαφορά και MA περιοδικότητα 12 μηνών. Η εκπαίδευση του μοντέλου βασίστηκε σε λογαριθμικά μετασχηματισμένα δεδομένα για την καλύτερη σταθεροποίηση της διακύμανσης, ενώ η διαδικασία πρόβλεψης περιλάμβανε επαναμετασχηματισμό στις αρχικές μονάδες με την συνάρτηση `expm1()`, προκειμένου τα αποτελέσματα να ερμηνευτούν απευθείας. Το διάγραμμα των προβλέψεων μέχρι το 2026 καταδεικνύει υψηλό βαθμό σταθερότητας στις κυκλικές διακυμάνσεις, με εμφανή διατήρηση της εποχικότητας και συγκράτηση του επιπέδου των ατυχημάτων γύρω από τον μέσο όρο της πρόσφατης δεκαετίας. Επίσης, τα διαστήματα εμπιστοσύνης (80% και 95%) παραμένουν αρκετά συγκρατημένα.

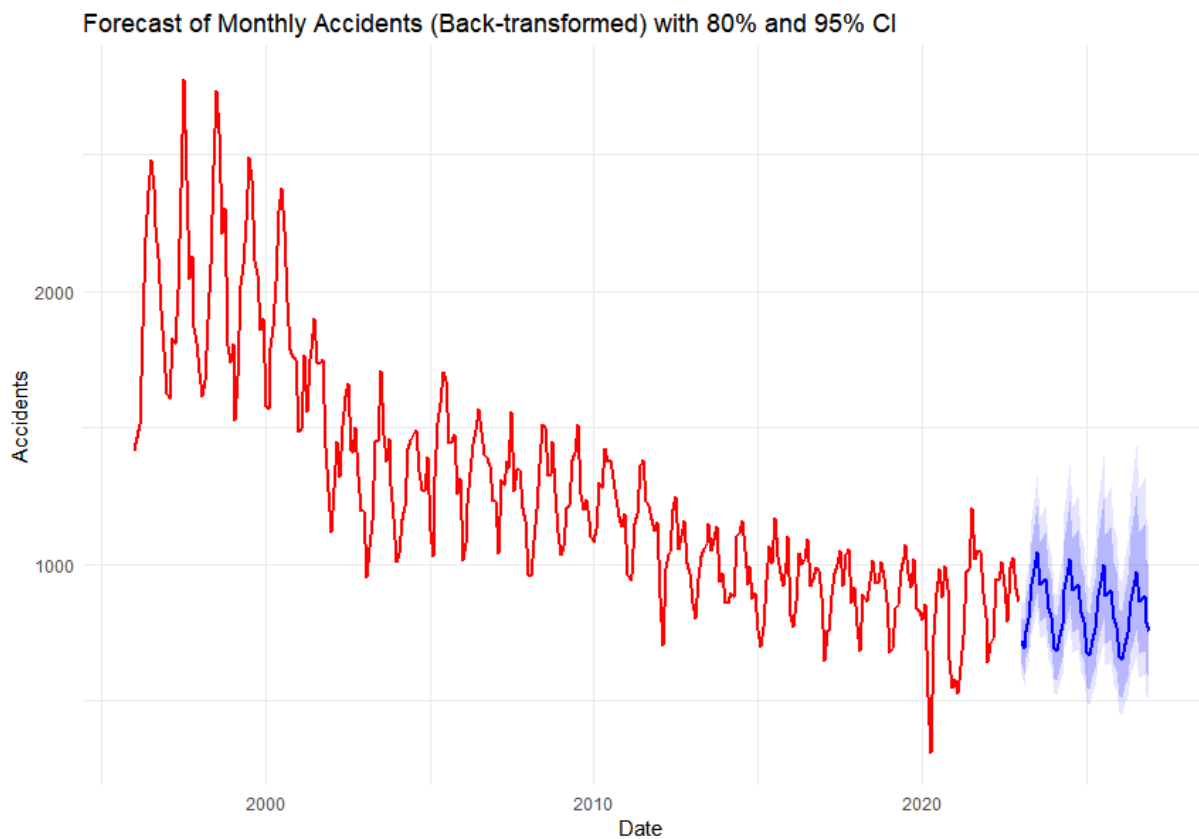
Η οπτικοποίηση των residuals του μοντέλου ενισχύει την εντύπωση της επάρκειας. Η κατανομή τους προσεγγίζει την κανονικότητα, δεν υπάρχουν σημαντικά μοτίβα με βάση το

ACF Plot και η χρονική κατανομή των σφαλμάτων δεν δείχνει συστηματικές αποκλίσεις, επιτρέποντας έτσι τη θεώρηση των residuals ως λευκός θόρυβος. Όμως, κάποια σημεία υπερβαίνουν τα όρια εμπιστοσύνης στις πρώτες καθυστερήσεις, υποδηλώνοντας ενδεχομένως τη δυνατότητα περαιτέρω βελτίωσης. Το διάγραμμα Actual vs Predicted αναδεικνύει την ικανότητα του μοντέλου να παρακολουθεί πιστά τις διακυμάνσεις των πραγματικών τιμών, ακόμα και σε περιόδους έντονης μεταβλητότητας.

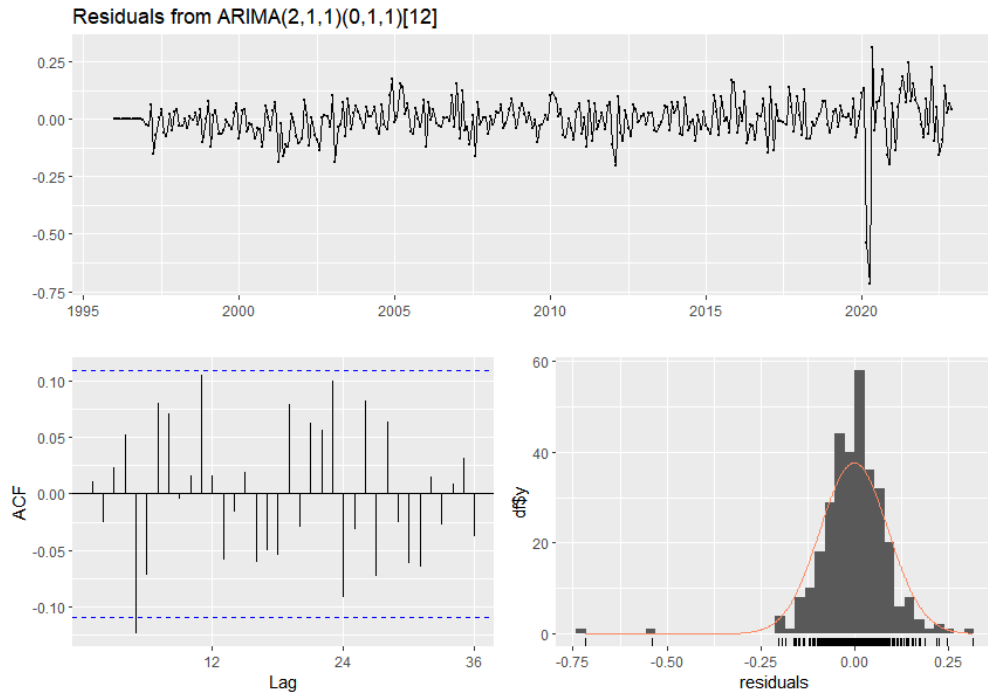
Ο δείκτης MAPE ανέρχεται στο 14.33%, ελαφρώς χαμηλότερος από τον αντίστοιχο του ARIMA, υποδεικνύοντας καλύτερη σχετική ακρίβεια πρόβλεψης. Η τιμή RMSE διαμορφώθηκε στα 143.30, ενώ το MAE στα 112.39, και ταυτόχρονα το μοντέλο παρουσίασε τις καλύτερες επιδόσεις στα AIC (-596.33) και BIC (-578.80).



Σχήμα 6.4.4-1: Actual vs Predicted Τιμες Μηνιαίων Τροχαίων Ατυχημάτων του Μοντέλου SARIMA(2,1,1)(0,1,1)[12]



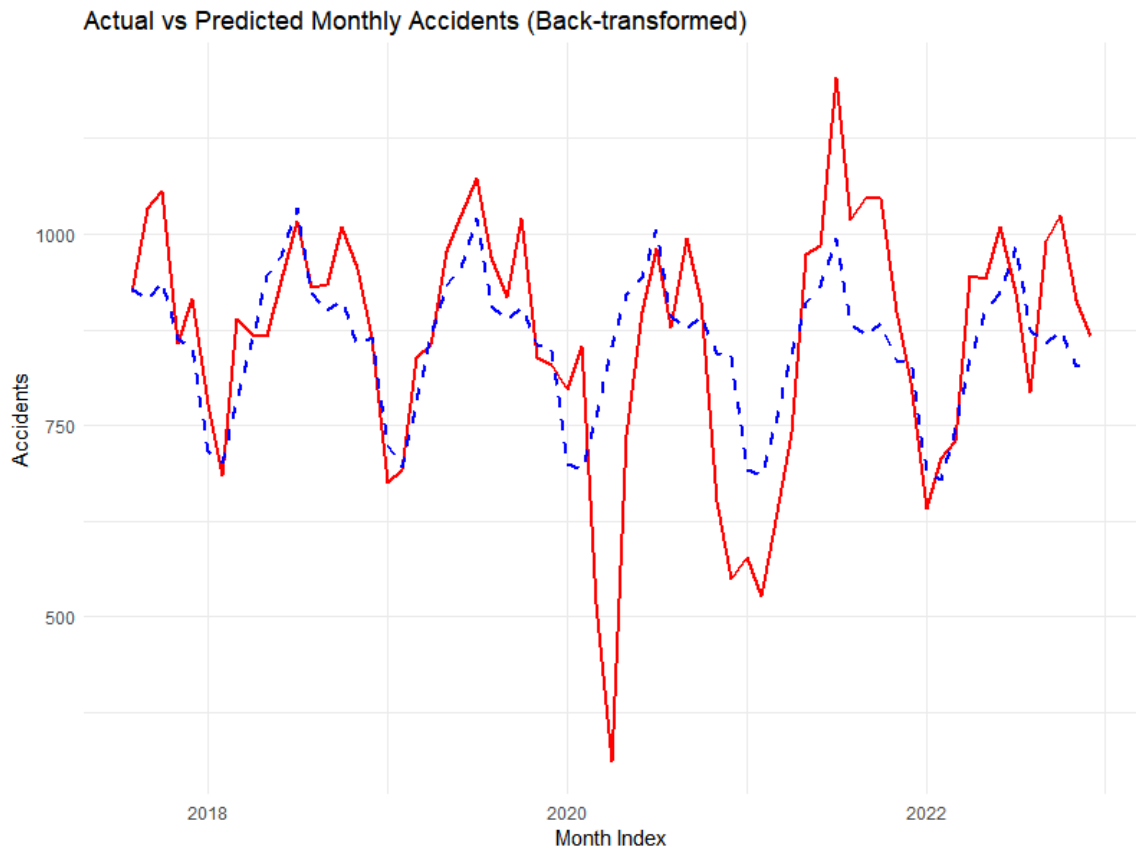
Σχήμα 6.4.4-2: Προβλέψεις του Μοντέλου SARIMA(2,1,1)(0,1,1)[12] για Μηνιαία Τροχαία Ατυχήματα έως το 2026



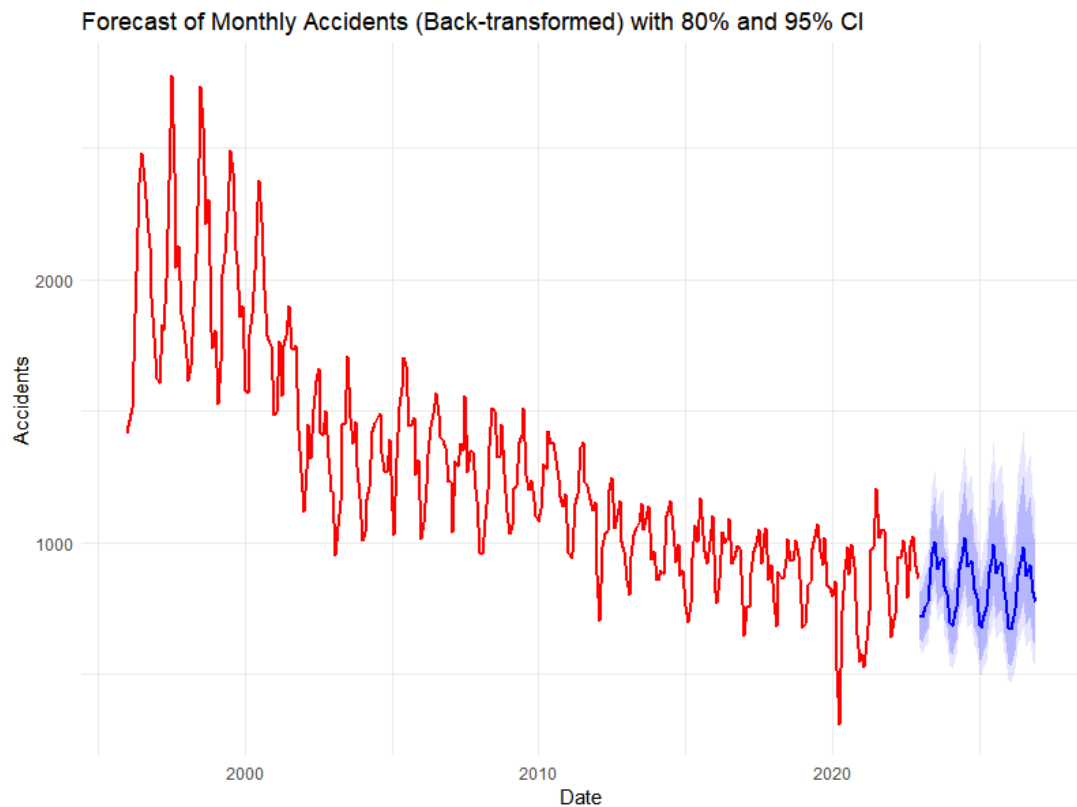
Σχήμα 6.4.4-3: Διαγνωστικά των Residuals του Μοντέλου SARIMA(2,1,1)(0,1,1)[12] σε Μηνιαία Τροχαία Ατυχήματα

6.4.5 GRID SEARCH ΑΠΟΤΕΛΕΣΜΑ ΓΙΑ ΤΑ ΜΗΝΙΑΙΑ ΔΕΔΟΜΕΝΑ ΤΡΟΧΑΙΩΝ ΑΤΥΧΗΜΑΤΩΝ

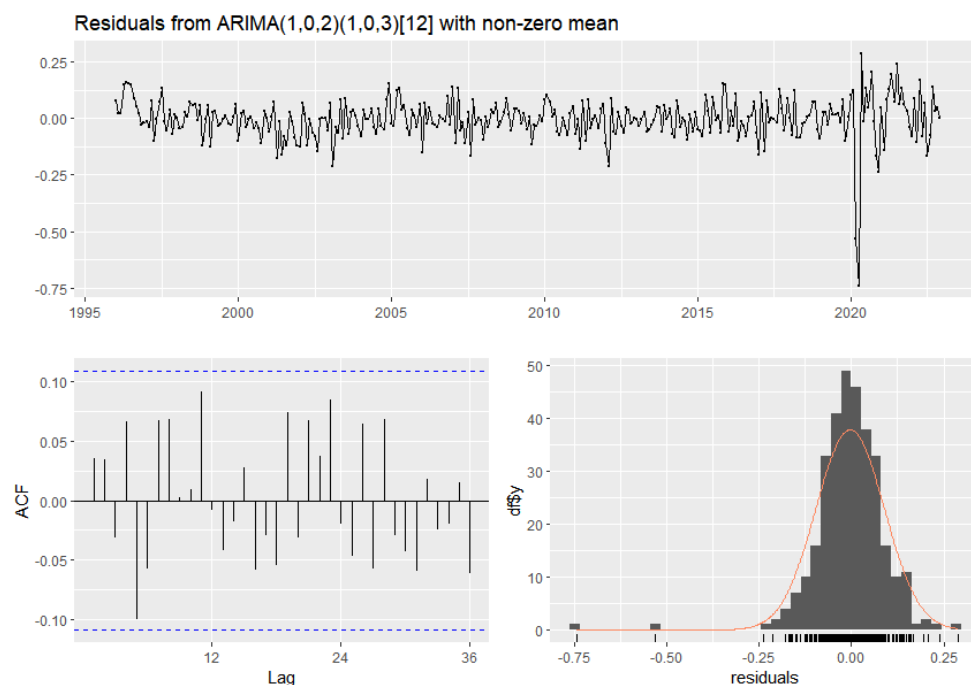
Τέλος, το μοντέλο SARIMA(1,0,2)(1,0,3)[12], το οποίο επιλέχθηκε από το grid search, παρουσιάζει τις καλύτερες επιδόσεις σε σχέση με όλα τα υπόλοιπα μοντέλα που δοκιμάστηκαν για τη μηνιαία χρονοσειρά ατυχημάτων. Η απόδοσή του χαρακτηρίζεται από χαμηλό σφάλμα πρόβλεψης (RMSE = 121.38, MAE = 86.62) και ιδιαίτερα ικανοποιητική σχετική ακρίβεια (MAPE = 12.33%). Ταυτόχρονα, τα residuals του μοντέλου περνούν επιτυχώς τους στατιστικούς ελέγχους αυτοσυσχέτισης και κανονικότητας, επιβεβαιώνοντας την καταλληλότητά του ως τελικό μοντέλο πρόβλεψης.



Σχήμα 6.4.5-1: Actual vs Predicted Τιμές Μηνιαίων Τροχαίων Ατυχημάτων του Μοντέλου $SARIMA(1,0,2)(1,0,3)[12]$



Σχήμα 6.4.5-2: Προβλέψεις του Μοντέλου SARIMA(1,0,2)(1,0,3)[12] για Μηνιαία Τροχαία Ατυχήματα έως το 2026



Σχήμα 6.4.5-3: Διαγνωστικά των Residuals του Μοντέλου SARIMA(1,0,2)(1,0,3)[12] σε Μηνιαία Τροχαία Ατυχήματα

6.4.6 ΕΠΙΠΛΕΟΝ ΠΙΝΑΚΕΣ ΚΑΙ ΜΕΤΡΙΚΕΣ ΓΙΑ ΤΑ ΜΟΝΤΕΛΑ ΤΡΟΧΑΙΩΝ ΑΤΥΧΗΜΑΤΩΝ

Πίνακας 6.4.6-1: Residual Checks στα Μοντέλα ARIMA / SARIMA Τροχαίων Ατυχημάτων

Τελικό Μοντέλο Τροχαίων Ατυχημάτων	Residual Check Pass
Εβδομαδιαίο ARIMA(4,1,1) with Drift	✗
Εβδομαδιαίο SARIMA(2,1,1)(0,1,1)[48]	✗
Μηνιαίο ARIMA(3,1,2)	✗
Μηνιαίο SARIMA(2,1,1)(0,1,1)[12]	✓
Μηνιαίο SARIMA(1,0,2)(1,0,3)[12] (GS)	✓

Πίνακας 6.4.6-2: Ενδεικτικές Τιμές Προβλέψεων Εβδομαδιαίους Μοντέλου ARIMA(4,1,1) with Drift

Date	Accidents
07/01/2023	151
14/1/2023	202
21/01/2023	204
28/01/2023	230
04/02/2023	169
11/02/2023	190
18/02/2023	190
25/02/2023	204
04/03/2023	176
11/03/2023	184
18/03/2023	183
25/03/2023	191

Πίνακας 6.4.6-3: Ενδεικτικές Τιμές Προβλέψεων Εβδομαδιαίου Μοντέλου
SARIMA(2,1,1)(0,1,1)[48]

Date	Accidents
07/01/2023	165
14/1/2023	181
21/01/2023	187
28/01/2023	181
04/02/2023	128
11/02/2023	193
18/02/2023	189
25/02/2023	192
04/03/2023	137
11/03/2023	210
18/03/2023	202
25/03/2023	196

Πίνακας 6.4.6-4: Ενδεικτικές Τιμές Προβλέψεων Μηνιαίου Μοντέλου ARIMA(3,1,2)

Date	Accidents
01/01/2023	778
01/02/2023	744
01/03/2023	753
01/04/2023	808
01/05/2023	894
01/06/2023	986
01/07/2023	1061
01/08/2023	1099
01/09/2023	1090
01/10/2023	1036
01/11/2023	954
01/12/2023	863

Πίνακας 6.4.6-5: Ενδεικτικές Τιμές Προβλέψεων Μηνιαίου Μοντέλου
SARIMA(2,1,1)(0,1,1)[12]

Date	Accidents
01/01/2023	717
01/02/2023	693
01/03/2023	756
01/04/2023	801
01/05/2023	928
01/06/2023	983
01/07/2023	1045
01/08/2023	927
01/09/2023	935
01/10/2023	950
01/11/2023	841
01/12/2023	808

Πίνακας 6.4.6-6: Ενδεικτικές Τιμές Προβλέψεων Μηνιαίου Μοντέλου
SARIMA(1,0,2)(1,0,3)[12] (GS)

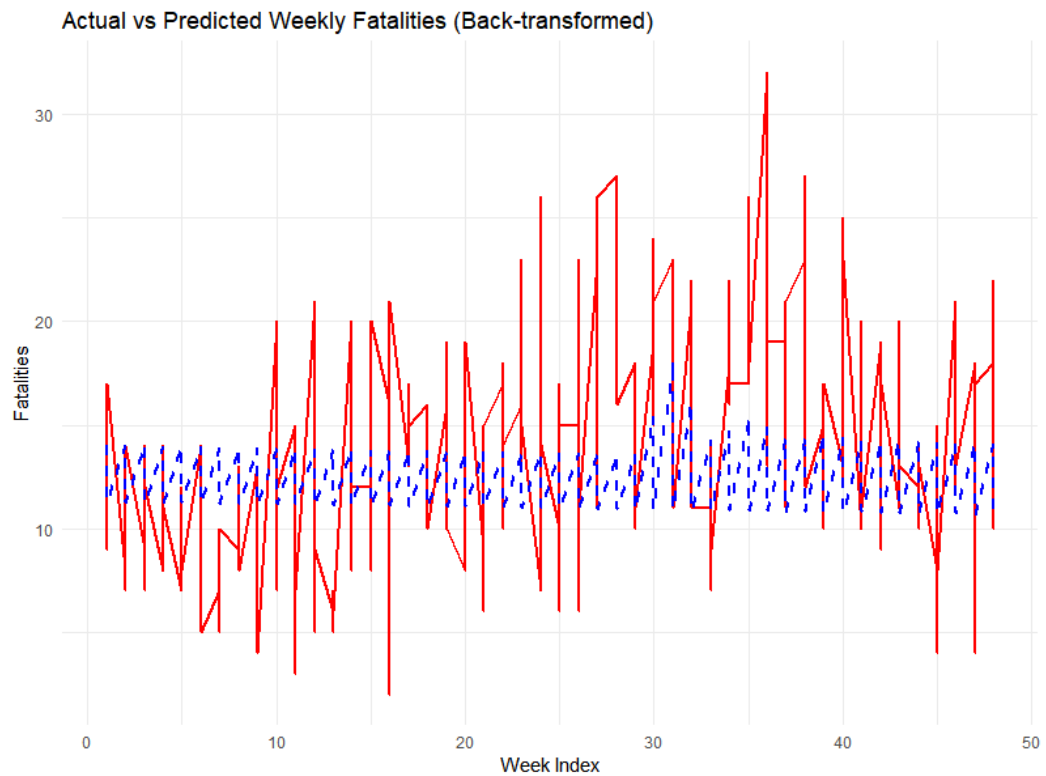
Date	Accidents
01/01/2023	720
01/02/2023	717
01/03/2023	756
01/04/2023	778
01/05/2023	905
01/06/2023	962
01/07/2023	1006
01/08/2023	898
01/09/2023	924
01/10/2023	942
01/11/2023	834
01/12/2023	805

6.4.7 ARIMA ΜΟΝΤΕΛΟ ΓΙΑ ΠΡΟΒΛΕΨΗ ΕΒΔΟΜΑΔΙΑΙΩΝ ΘΑΝΑΤΩΝ ΑΠΟ ΤΡΟΧΑΙΑ ΑΤΥΧΗΜΑΤΑ

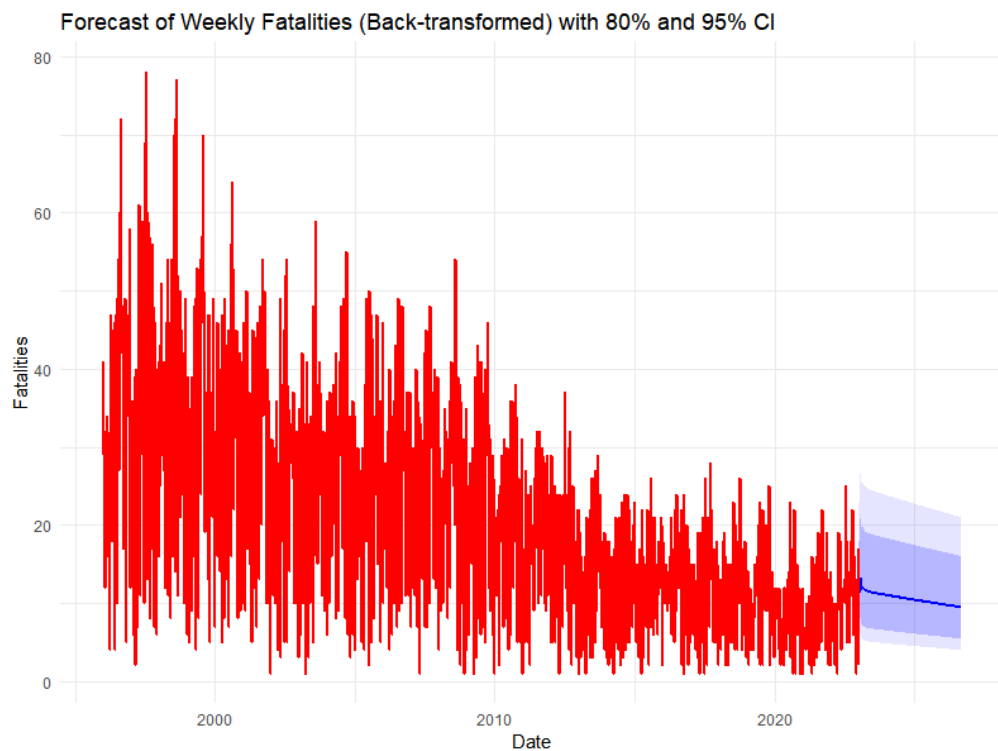
Η εφαρμογή του μοντέλου ARIMA(4,1,1) με drift στους εβδομαδιαίους θανάτους από τροχαία ατυχήματα επιδιώκει να αξιοποιήσει τη χρονική δυναμική της σειράς για την πρόβλεψη μελλοντικών τιμών. Η εκπαίδευση πραγματοποιήθηκε, όπως και πριν, σε ιστορικά δεδομένα μέχρι και τα μέσα του 2017, ενώ οι προβλέψεις συγκρίθηκαν με τα πραγματικά δεδομένα της περιόδου επαλήθευσης. Όπως φαίνεται στο γράφημα Actual vs Predicted, το μοντέλο καταφέρνει να προσεγγίσει τη γενική πορεία της χρονοσειράς, αλλά αδυνατεί να αποτυπώσει την απότομη διακύμανση των εβδομαδιαίων θανάτων, ιδιαίτερα σε σημεία με αιχμές ή βύθιση.

Η πρόβλεψη μέχρι το 2026 παρουσιάζεται στο δεύτερο γράφημα με τις ζώνες εμπιστοσύνης 80% και 95%. Η προβολή της σειράς υποδεικνύει μια ήπια καθοδική τάση στους θανάτους, εντός ορίων που δείχνουν σχετική σταθερότητα και απουσία αυξημένων κινδύνων. Παρά την εμπιστοσύνη που προκύπτει από την εξομάλυνση και τη σταθερότητα της πρόβλεψης, η ευρεία διακύμανση των πραγματικών δεδομένων στο παρελθόν υποδεικνύει ότι υπάρχει πιθανότητα για σφάλματα, ιδίως όταν εμφανίζονται ακραία επεισόδια.

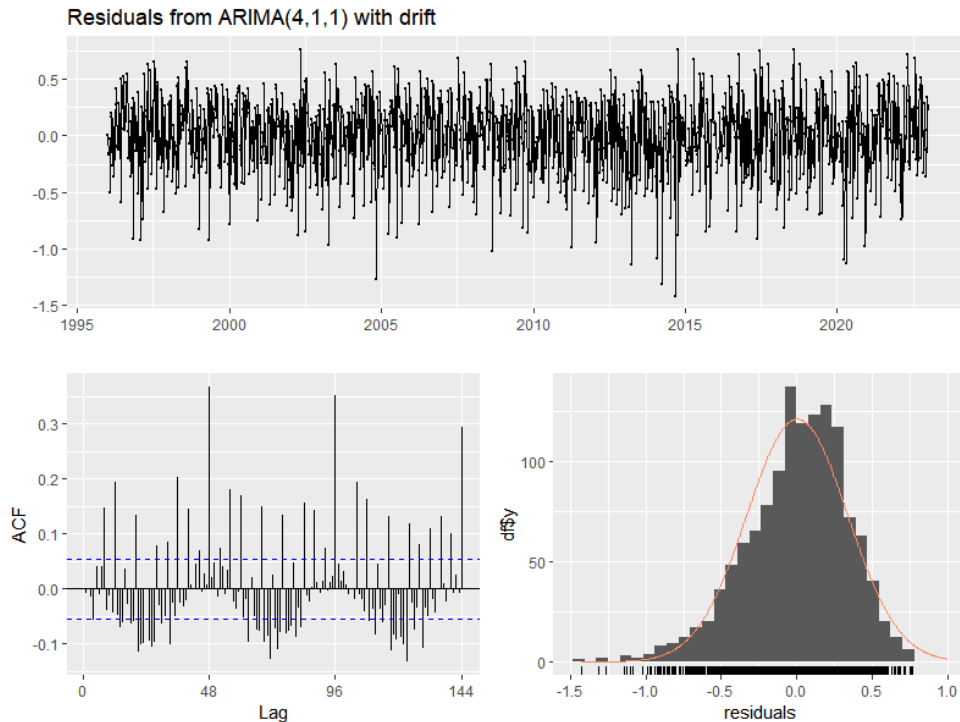
Οι διαγνωστικοί έλεγχοι υποδεικνύουν ικανοποιητικά residuals όσον αφορά τη σταθερότητα και την απουσία συστηματικής αυτοσυσχέτισης. Το γράφημα των residuals δείχνει καλή κατανομή γύρω από το μηδέν, ενώ το ACF δεν εμφανίζει στατιστικά σημαντικές κορυφώσεις. Ωστόσο, παρά την επάρκεια των διαγνωστικών, η απόδοση του μοντέλου στις μετρικές είναι περιορισμένη με RMSE = 5.42, MAE = 4.30 και MAPE = 37.39%. Η υψηλή τιμή του MAPE αποκαλύπτει ότι το μοντέλο παρουσιάζει μεγαλύτερο σχετικό σφάλμα όταν οι πραγματικές τιμές είναι χαμηλές.



Σχήμα 6.4.7-1: Actual vs Predicted Τιμές Εβδομαδιαίων Θανάτων από Τροχαία Ατυχήματα του Μοντέλου ARIMA(4,1,1) with Drift



Σχήμα 6.4.7-2 Προβλέψεις του Μοντέλου ARIMA(4,1,1) with Drift για Εβδομαδιαίους Θανάτους από Τροχαία Ατυχήματα έως το 2026



Σχήμα 6.4.7-3: Διαγνωστικά των Residuals του Μοντέλου ARIMA(4,1,1) with Drift σε Εβδομαδιαίους Θανάτους από Τροχαία Ατυχήματα

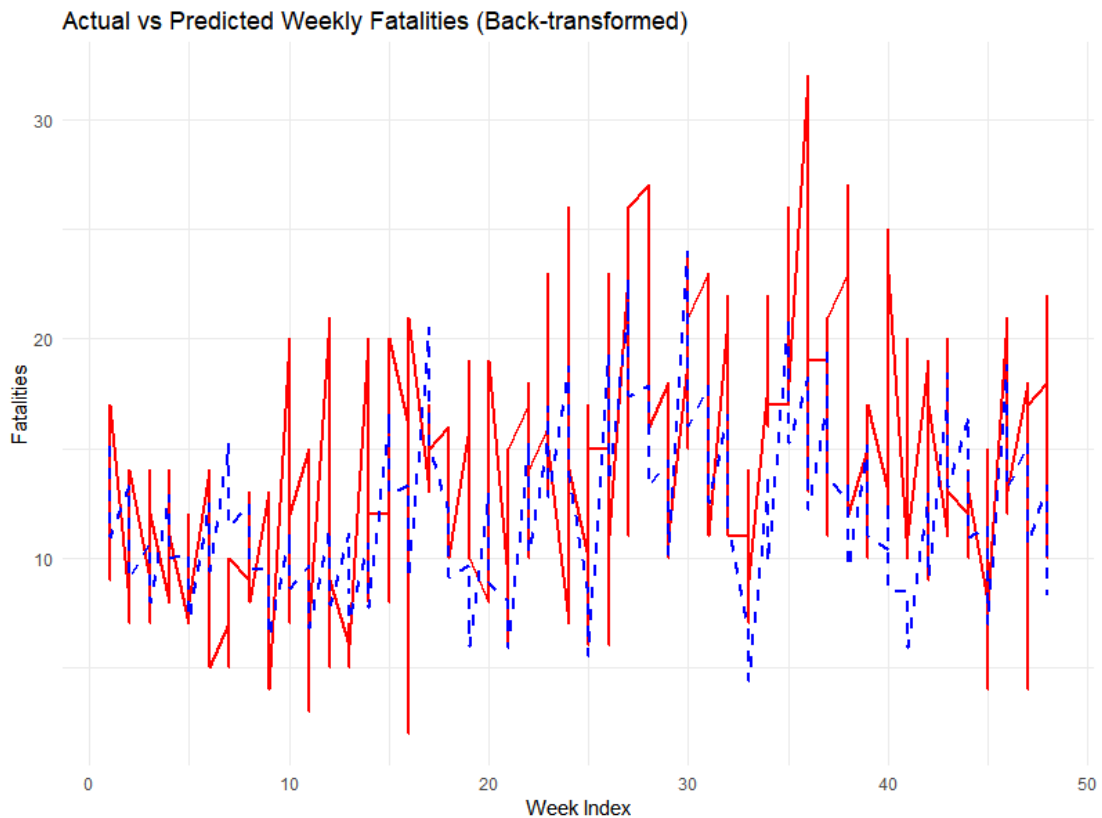
6.4.8 SARIMA ΜΟΝΤΕΛΟ ΓΙΑ ΠΡΟΒΛΕΨΗ ΕΒΔΟΜΑΔΙΑΙΩΝ ΘΑΝΑΤΩΝ ΑΠΟ ΤΡΟΧΑΙΑ ΑΤΥΧΗΜΑΤΑ

Το μοντέλο SARIMA(1,1,2)(1,1,0)[48] εφαρμόστηκε στα εβδομαδιαία δεδομένα θανάτων, αξιοποιώντας τον λογαριθμικό μετασχηματισμό για εξομάλυνση της διακύμανσης. Η σύγκριση των πραγματικών και των προβλεπόμενων τιμών στο διάστημα ελέγχου (test set) υποδεικνύει υψηλή επικάλυψη μεταξύ των δύο σειρών, με τη διακεκομμένη μπλε γραμμή (πρόβλεψη) να ακολουθεί σε μεγάλο βαθμό την κόκκινη (πραγματικά δεδομένα). Ωστόσο, καταγράφεται κάπως μειωμένη ικανότητα του μοντέλου να προβλέψει ακραίες κορυφές.

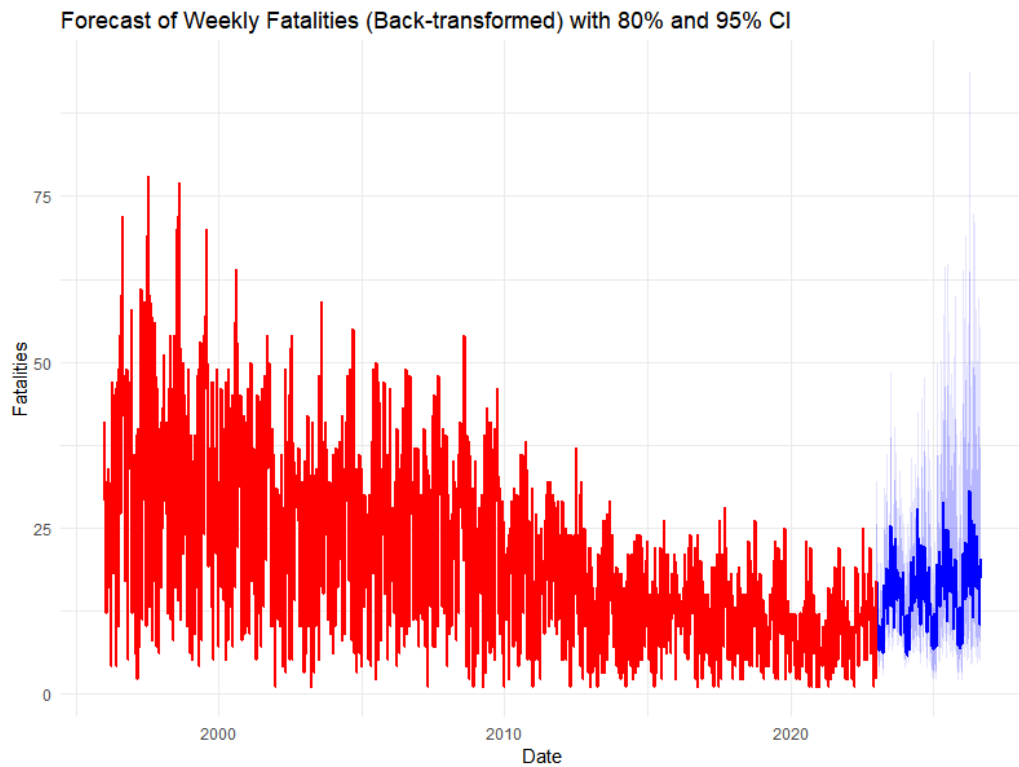
Η προβολή των προβλέψεων μέχρι το 2026, συνοδευόμενη από τα διαστήματα εμπιστοσύνης 80% και 95%, αναδεικνύει μια σταθερή τάση με μικρές εποχικές διακυμάνσεις. Η γενική εικόνα υποδηλώνει σταθεροποίηση των θανάτων σε χαμηλότερα επίπεδα συγκριτικά με τις προηγούμενες δεκαετίες, χωρίς ενδείξεις άμεσης εκτροπής προς ανοδική πορεία. Το εύρος των διαστημάτων εμπιστοσύνης παραμένει διαχειρίσιμο, γεγονός που αντανακλά ικανοποιητική προβλεπτική αξιοπιστία. Η κατανομή των residuals στο αντίστοιχο γράφημα

παρουσιάζει συμμετρία και φυσιολογική κατανομή, ενώ το ACF δείχνει μη σημαντική αυτοσυσχέτιση.

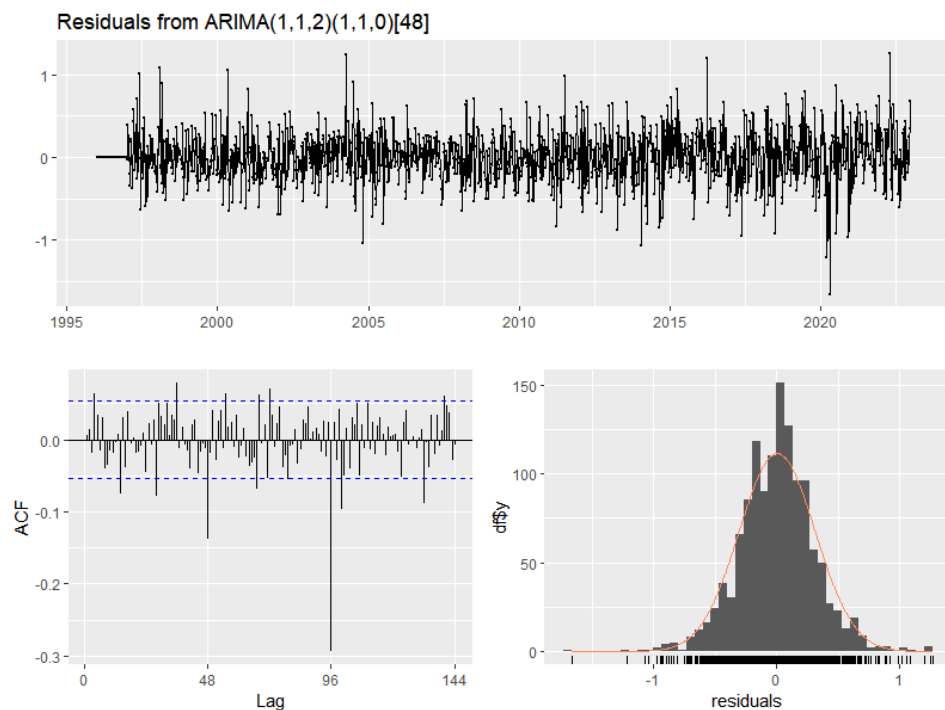
Οι μετρικές αξιολόγησης αποτυπώνουν την καλή απόδοση του μοντέλου SARIMA σε αυτό το πλαίσιο. Η τιμή του AIC (-436.45) και του BIC (-460.93) είναι ενδεικτικές καλής προσαρμογής, ενώ οι τιμές σφάλματος RMSE (5.15), MAE (4.08) και MAPE (33.03%) δείχνουν ένα σχετικά όχι και τόσο ικανοποιητικό επίπεδο ακρίβειας.



Σχήμα 6.4.8-1: Actual vs Predicted Τιμές Εβδομαδιαίων Θανάτων από Τροχαία Ατυχήματα του Μοντέλου SARIMA(1,1,2)(1,1,0)[48]



Σχήμα 6.4.8-2: Προβλέψεις του Μοντέλου SARIMA(1,1,2)(1,1,0)[48] για Εβδομαδιαίους Θανάτους από Τροχαία Ατυχήματα έως το 2026



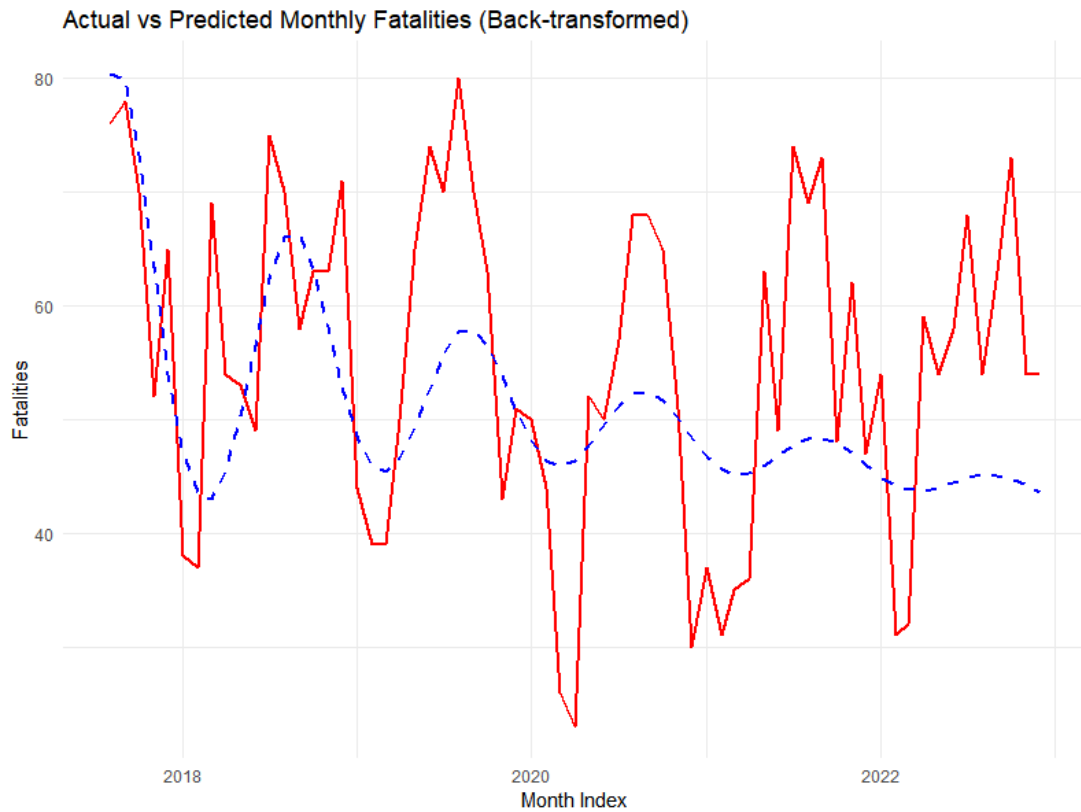
Σχήμα 6.4.8-3: Διαγνωστικά των Residuals του Μοντέλου SARIMA(1,1,2)(1,1,0)[48] σε Εβδομαδιαίους Θανάτους από Τροχαία Ατυχήματα

6.4.9 ARIMA ΜΟΝΤΕΛΟ ΓΙΑ ΠΡΟΒΛΕΨΗ ΜΗΝΙΑΙΩΝ ΘΑΝΑΤΩΝ ΑΠΟ ΤΡΟΧΑΙΑ ΑΤΥΧΗΜΑΤΑ

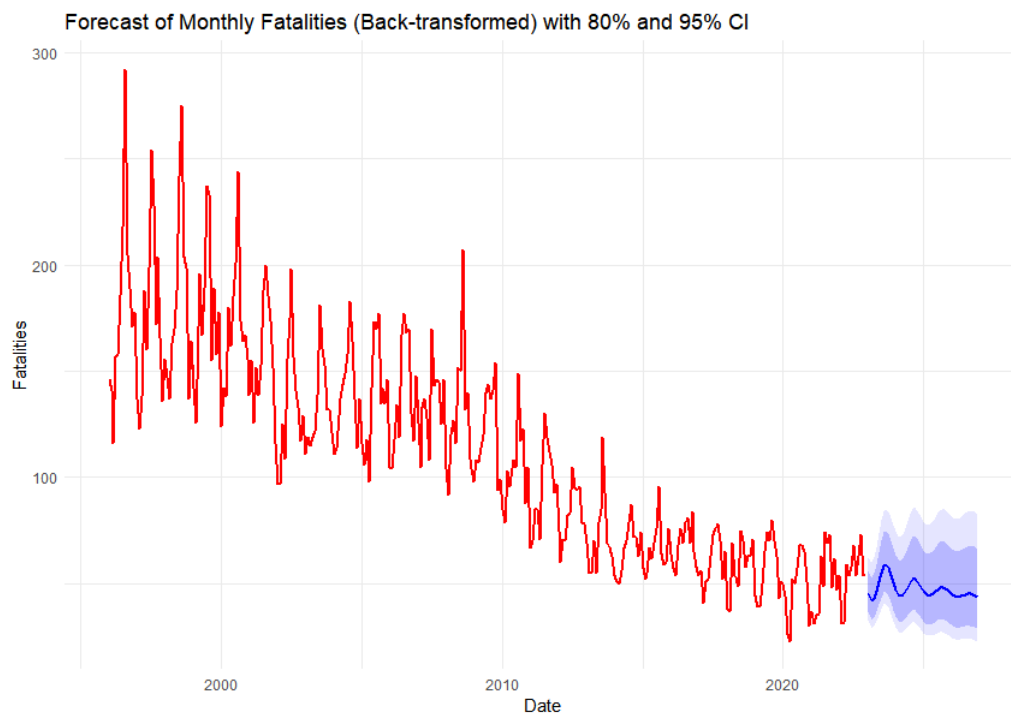
Το μοντέλο ARIMA(3,1,2) με drift εφαρμόστηκε στα μηνιαία δεδομένα θανάτων και τα αποτελέσματά του παρουσιάζονται μέσα από την παρακάτω σειρά διαγραμμάτων. Στο γράφημα προβλέψεων με διαστήματα εμπιστοσύνης 80% και 95%, φαίνεται ότι η σειρά διατηρεί τη φθίνουσα πορεία της και μετά το 2022, προβλέποντας σχετικά σταθερά επίπεδα θανάτων με ήπιες διακυμάνσεις. Οι εκτιμώμενες τιμές κινούνται γύρω από τις 60 απώλειες ανά μήνα, με αβεβαιότητα προς τα άκρα. Τα διαστήματα εμπιστοσύνης υποδεικνύουν αυξανόμενη αβεβαιότητα όσο ο χρονικός ορίζοντα της πρόβλεψης απομακρύνεται.

Η σύγκριση μεταξύ πραγματικών και προβλεπόμενων τιμών στο test set επιβεβαιώνει ότι το μοντέλο ακολουθεί ικανοποιητικά τις διακυμάνσεις της σειράς, αν και εμφανίζει κάποιες αποκλίσεις στις κορυφώσεις και τα κατώτερα σημεία, ιδίως σε περιόδους όπου παρατηρούνται ακραίες τιμές. Το διάγραμμα των residuals επιβεβαιώνει τον λευκό θόρυβο των residuals του μοντέλου, γεγονός που αποτυπώνεται τόσο στον έλεγχο αυτοσυσχέτισης (ACF) όσο και στην κατανομή τους, η οποία πλησιάζει την κανονική.

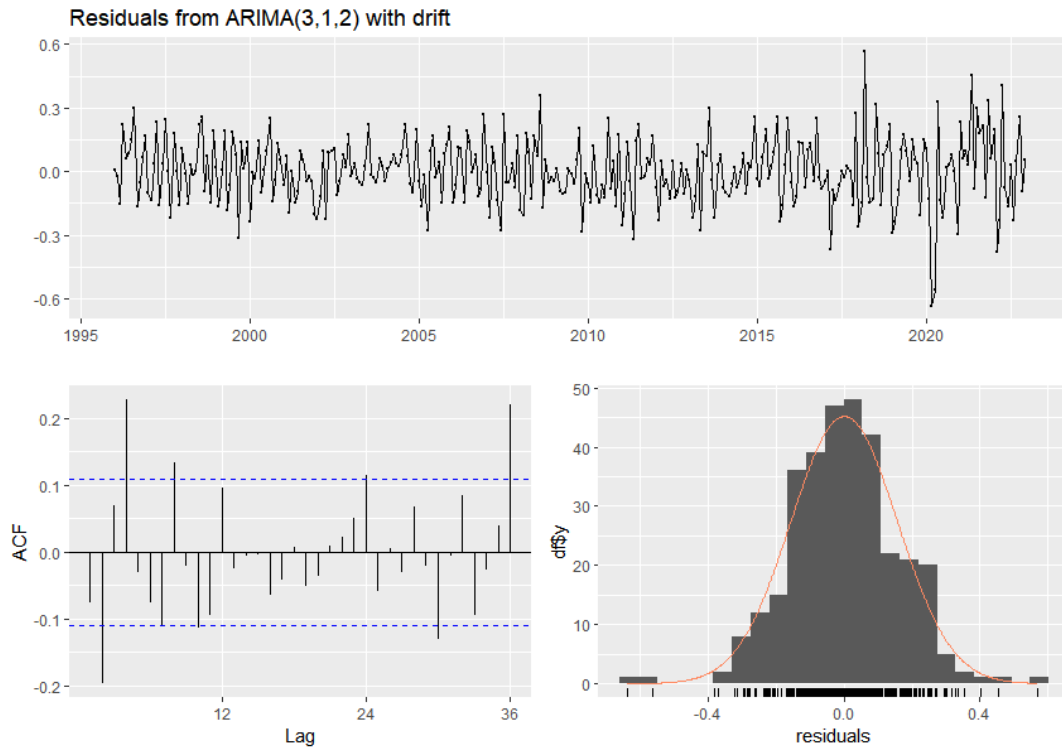
Όσον αφορά τις μετρικές απόδοσης του μοντέλου, η τιμή AIC είναι -284.04 και η BIC -259.17, υποδηλώνοντας καλή ισορροπία μεταξύ πολυπλοκότητας και προσαρμογής. Η RMSE 13.15 και η MAE 10.90 καταδεικνύουν σχετικά μικρά σφάλματα πρόβλεψης σε απόλυτες τιμές. Η τιμή του MAPE ανέρχεται σε 21.28%, υποδεικνύοντας ικανοποιητική ακρίβεια, λαμβάνοντας υπόψη τη φύση των δεδομένων και την ύπαρξη μεταβλητότητας.



Σχήμα 6.4.9-1: Actual vs Predicted Τιμες Μηνιαίων Θανάτων από Τροχαία Ατυχήματα του Μοντέλου ARIMA(3,1,2) with Drift



Σχήμα 6.4.9-2: Προβλέψεις του Μοντέλου ARIMA(3,1,2) with Drift για Μηνιαίους Θανάτους από Τροχαία Ατυχήματα έως το 2026



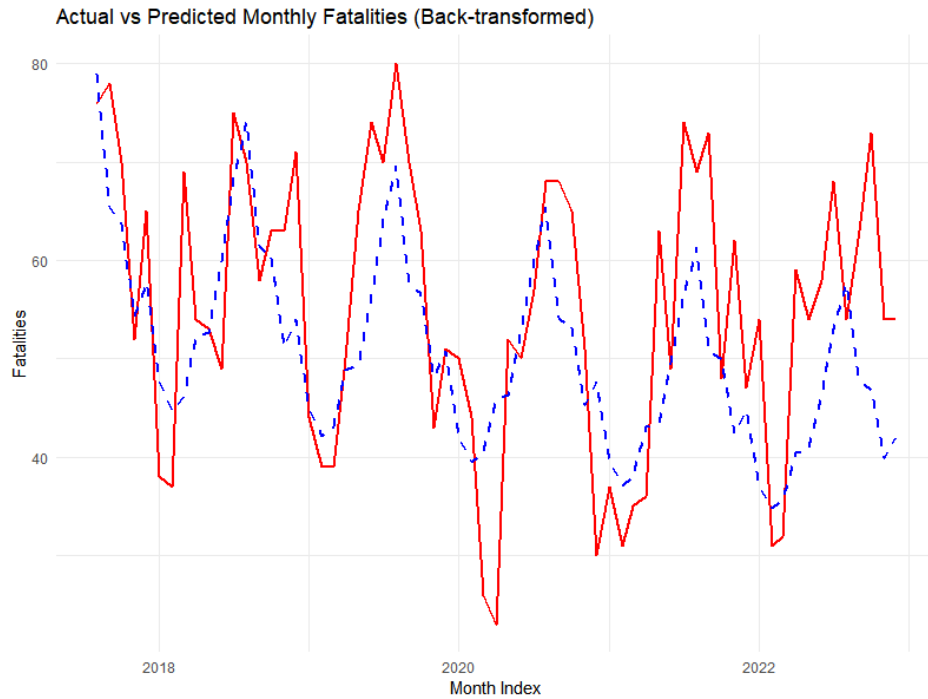
Σχήμα 6.4.9-3: Διαγνωστικά των Residuals του Μοντέλου ARIMA(3,1,2) with Drift σε Μηνιαίους Θανάτους από Τροχαία Ατυχήματα

6.4.10 SARIMA ΜΟΝΤΕΛΟ ΓΙΑ ΠΡΟΒΛΕΨΗ ΜΗΝΙΑΙΩΝ ΘΑΝΑΤΩΝ ΑΠΟ ΤΡΟΧΑΙΑ ΑΤΥΧΗΜΑΤΑ

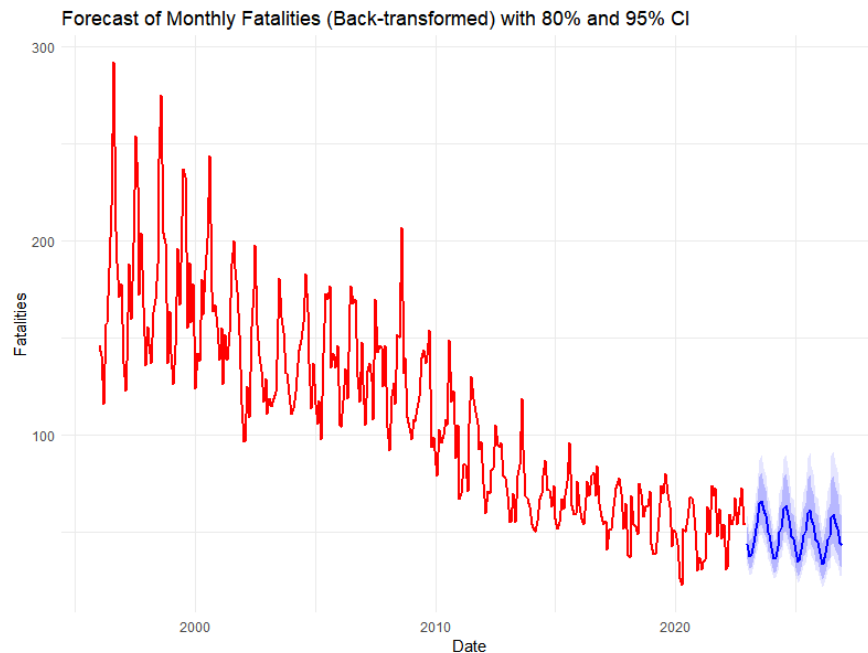
Το μοντέλο SARIMA(2,1,1)(1,1,1)[12], το οποίο αντίστοιχα εφαρμόστηκε για την πρόβλεψη των μηνιαίων θανάτων από τροχαία ατυχήματα, ανέδειξε αρκετά ενθαρρυντικά αποτελέσματα τόσο σε επίπεδο προσαρμογής όσο και προβλεπτικής ακρίβειας. Η γραφική απεικόνιση της πρόβλεψης έως το 2026 δείχνει μια σταθεροποιημένη πορεία στους μελλοντικούς θανάτους, με περιοδικότητα και ένταση εντός των διαστημάτων εμπιστοσύνης 80% και 95%. Η σαφής εποχική δυναμική που ενσωματώνει το μοντέλο, επιτρέπει πιο ρεαλιστικές εκτιμήσεις, λαμβάνοντας υπόψη τις επαναλαμβανόμενες διακυμάνσεις ανά έτος.

Τα διαγνωστικά των residuals αποτυπώνουν επίσης θετικά ευρήματα. Το γράφημα ACF δείχνει ότι δεν υφίσταται σημαντική αυτοσυσχέτιση στα residuals, ενώ το ιστόγραμμα υποδεικνύει συμμετρία και προσέγγιση της κανονικότητας, ενισχύοντας τη στατιστική αξιοπιστία του μοντέλου. Οι μετρικές αξιολόγησης παρουσιάζονται ιδιαίτερα ικανοποιητικές, με RMSE ίσο με 11.28, MAE περίπου 9.08 και MAPE της τάξης του 17.33%. Ταυτόχρονα, οι

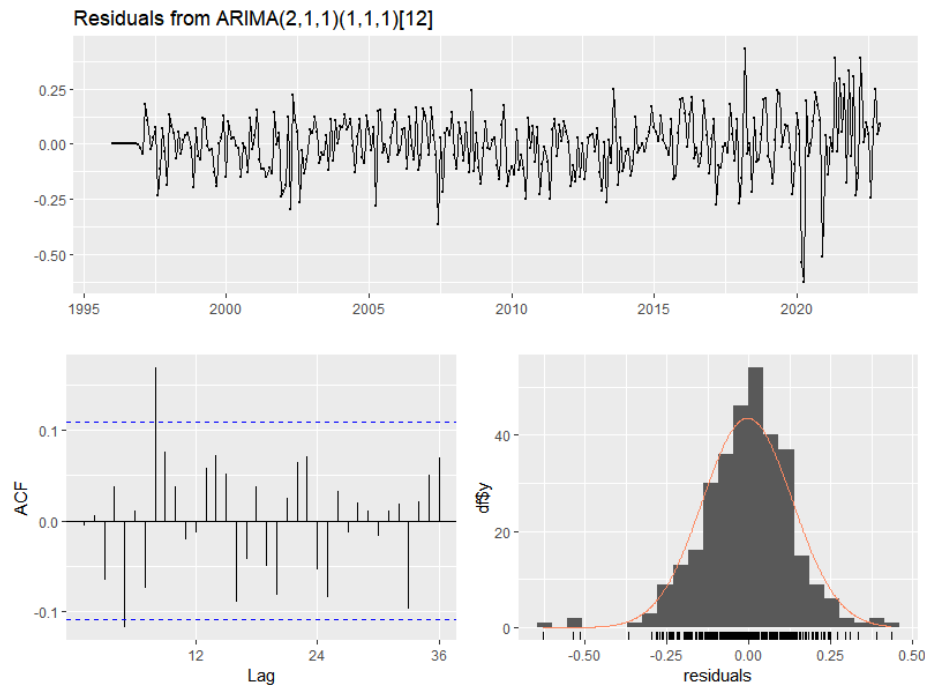
τιμές των AIC (-334.68) και BIC (-313.65) δηλώνουν καλή προσαρμογή, συγκριτικά με άλλα υποψήφια μοντέλα.



Σχήμα 6.4.10-1: Actual vs Predicted Τιμες Μηνιαίων Θανάτων από Τροχαία Ατυχήματα του Μοντέλου SARIMA(2,1,1)(1,1,1)[12]



Σχήμα 6.4.10-2: Προβλέψεις του Μοντέλου SARIMA(2,1,1)(1,1,1)[12] για Μηνιαίους Θανάτους από Τροχαία Ατυχήματα έως το 2026

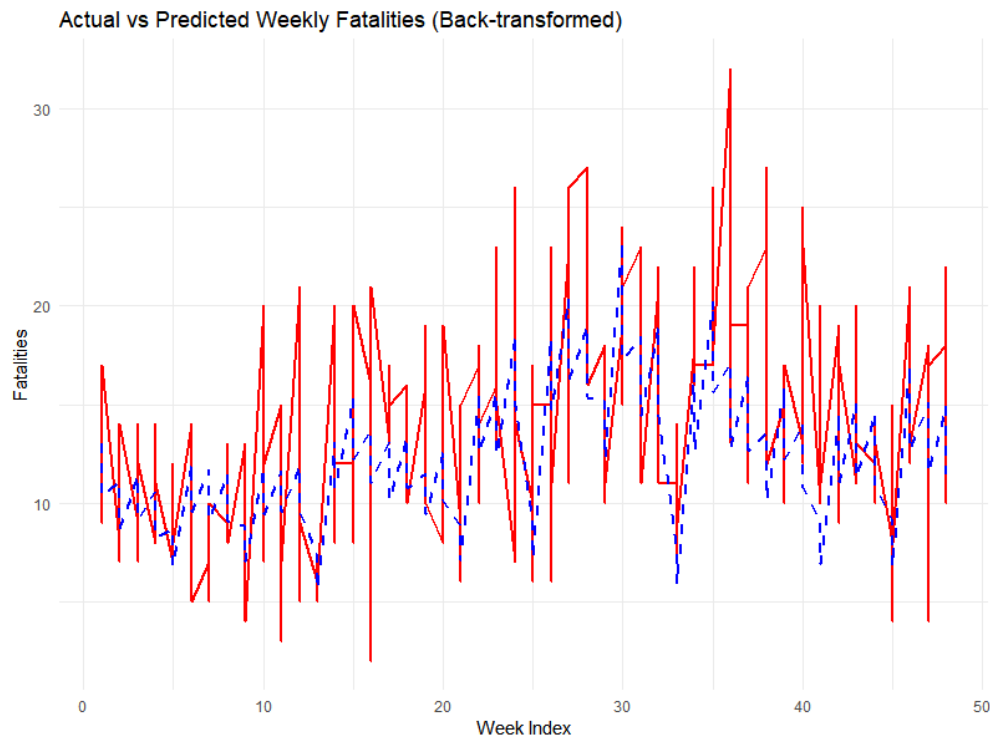


Σχήμα 6.4.10-3: Διαγνωστικά των Residuals του Μοντέλου SARIMA(2,1,1)(1,1,1)[12] σε Μηνιαίους Θανάτους από Τροχαία Ατυχήματα

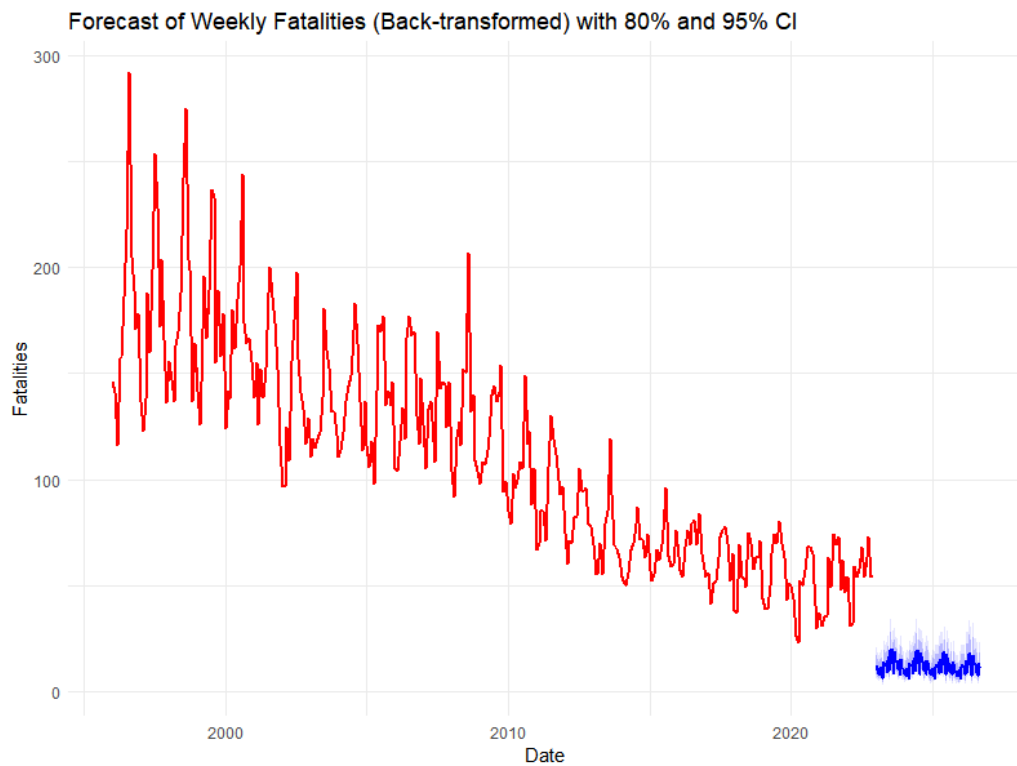
6.4.11 GRID SEARCH ΑΠΟΤΕΛΕΣΜΑΤΑ ΓΙΑ ΕΒΔΟΜΑΔΙΑΙΑ ΚΑΙ ΜΗΝΙΑΙΑ ΔΕΔΟΜΕΝΑ ΘΑΝΑΤΩΝ ΑΠΟ ΤΡΟΧΑΙΑ ΑΤΥΧΗΜΑΤΑ

Το μοντέλο SARIMA(3,0,3)(0,1,3)[48] προέκυψε από το Grid Search ως το πιο αποδοτικό για την πρόβλεψη εβδομαδιαίων θανάτων από τροχαία ατυχήματα. Όπως απεικονίζεται στα αντίστοιχα διαγράμματα, η πρόβλεψη των πραγματικών τιμών στο test set κινείται πολύ κοντά στις παρατηρήσεις, με την εποχική συμπεριφορά να αναπαρίσταται μέσω των περιοδικών διακυμάνσεων. Η πρόβλεψη έως το 2026 διατηρεί μια σταθερή αλλά διακριτή εποχική δομή, ενώ τα διαστήματα εμπιστοσύνης αποτυπώνουν την προβλεπτική αβεβαιότητα, ιδιαίτερα σε μεγαλύτερο χρονικό ορίζοντα.

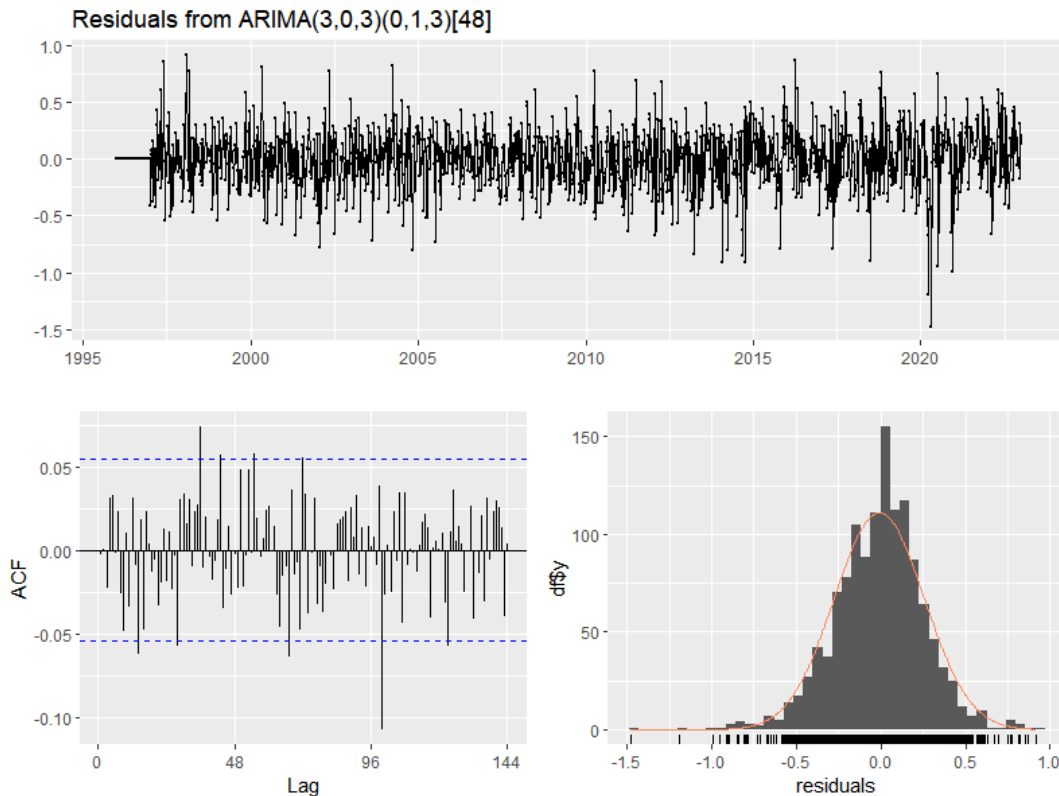
Αναφορικά με τις υπολογισθείσες μετρικές, το μοντέλο πέτυχε πολύ χαμηλές τιμές σφάλματος με RMSE = 4.50 και MAE = 3.50, ενώ το MAPE ήταν στο 29.04%, καθιστώντας το μοντέλο πιο ακριβές σε σύγκριση με άλλες εκδοχές που εξετάστηκαν.



Σχήμα 6.4.11-1: Actual vs Predicted Τιμες Εβδομαδιαίων Θανάτων από Τροχαία Ατυχήματα του Μοντέλου SARIMA(3,0,3)(0,1,3)[48]



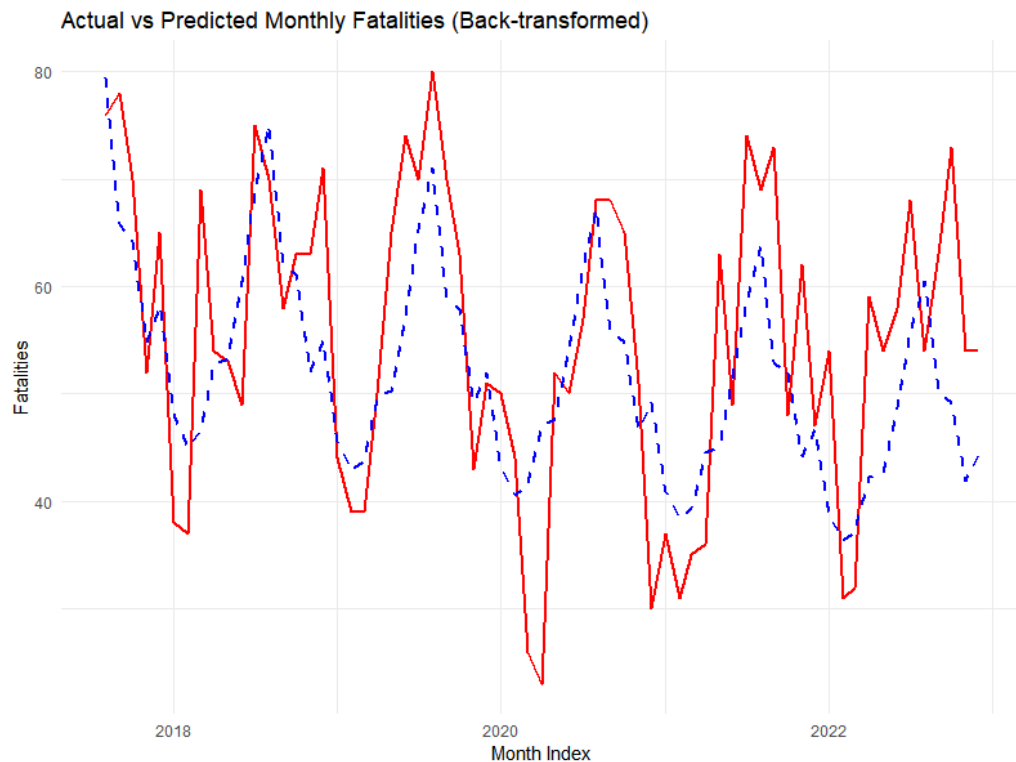
Σχήμα 6.4.11-2: Προβλέψεις του Μοντέλου SARIMA(3,0,3)(0,1,3)[48] για Εβδομαδιαίους Θανάτους από Τροχαία Ατυχήματα έως το 2026



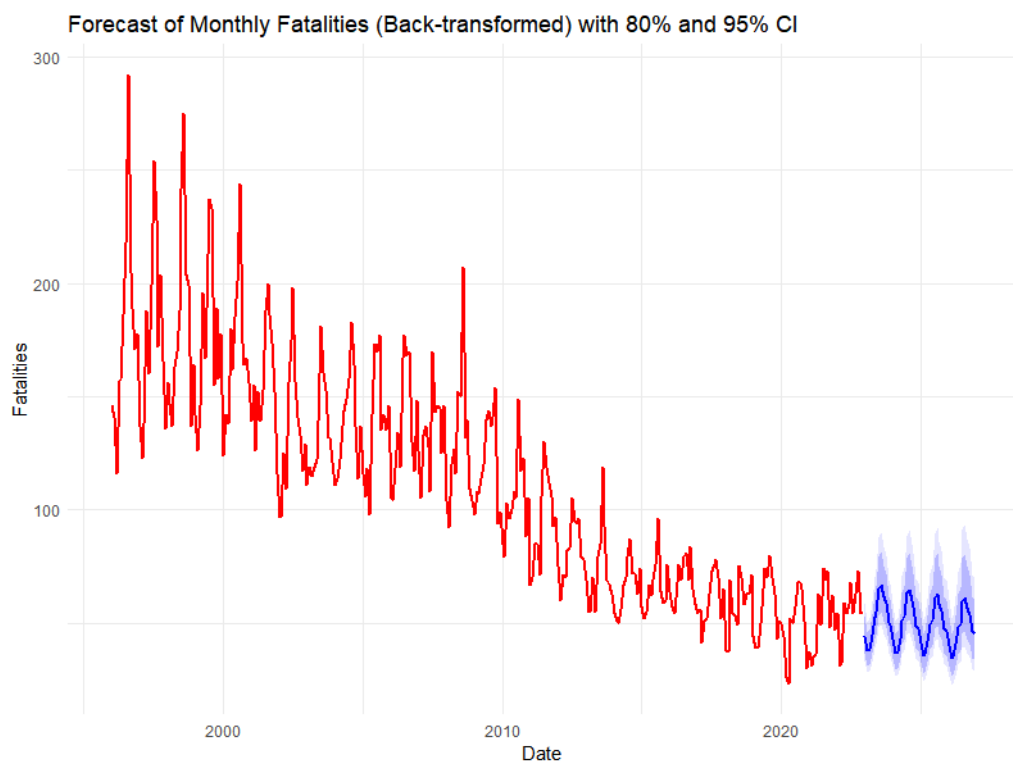
Σχήμα 6.4.11-3: Διαγνωστικά των Residuals του Μοντέλου SARIMA(3,0,3)(0,1,3)[48] σε Εβδομαδιαίους Θανάτους από Τροχαία Ατυχήματα

Η διαδικασία μοντελοποίησης των μηνιαίων θανάτων με χρήση του τελικού SARIMA μοντέλου προερχόμενου από grid search (SARIMA(1,0,2)(1,1,1)[12]) αποδείχθηκε ιδιαίτερα επιτυχημένη, συνδυάζοντας υψηλή προβλεπτική ακρίβεια με ικανοποιητική συμπεριφορά των residuals. Όπως φαίνεται στο διάγραμμα της πρόβλεψης μέχρι το 2026, το μοντέλο αποδίδει καμπύλες πρόβλεψης που ακολουθούν τη μακροχρόνια φθίνουσα τάση των θανάτων, ενώ ταυτόχρονα αποτυπώνουν την εποχικότητα με εύρος αβεβαιότητας 80% και 95%.

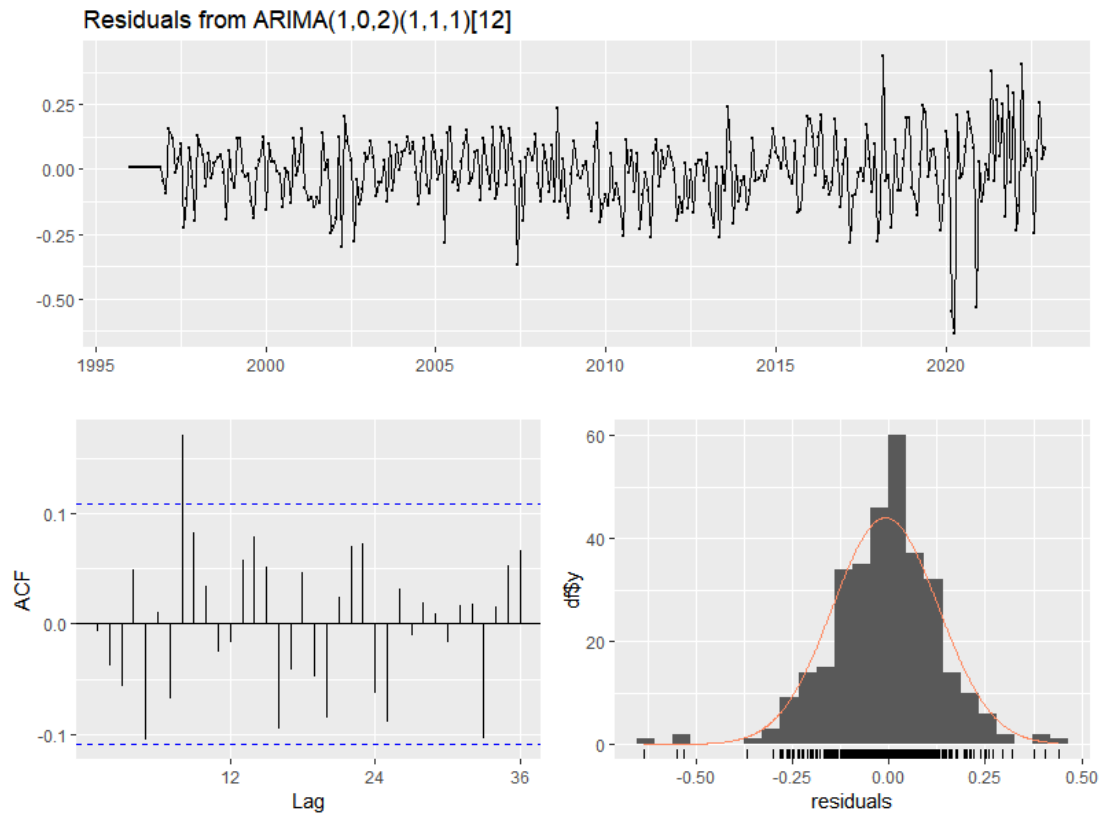
Η εκτίμηση της απόδοσης σε επίπεδο test set επιβεβαιώνει τη βελτιωμένη ακρίβεια του grid searched μοντέλου σε σύγκριση με προηγούμενες εκδοχές. Οι τιμές RMSE (10.65), MAE (8.71) και κυρίως το MAPE (17.14%) είναι οι χαμηλότερες που επιτεύχθηκαν στην συγκεκριμένη περίπτωση, ενώ συνοδεύονται από AIC και BIC τιμές ίδιες ή καλύτερες από εναλλακτικά μοντέλα.



Σχήμα 6.4.11-4: Actual vs Predicted Τιμες Μηνιαίων Θανάτων από Τροχαία Ατυχήματα του Μοντέλου SARIMA(1,0,2)(1,1,1)[12]



Σχήμα 6.4.11-5: Προβλέψεις του Μοντέλου SARIMA(1,0,2)(1,1,1)[12] για Μηνιαίους Θανάτους από Τροχαία Ατυχήματα έως το 2026



Σχήμα 6.4.11-6: Διαγνωστικά των Residuals του Μοντέλου SARIMA(1,0,2)(1,1,1)[12] σε Μηνιαίους Θανάτους από Τροχαία Ατυχήματα

6.4.12 ΕΠΙΠΛΕΟΝ ΠΙΝΑΚΕΣ ΚΑΙ ΜΕΤΡΙΚΕΣ ΓΙΑ ΤΑ ΜΟΝΤΕΛΑ ΘΑΝΑΤΩΝ ΑΠΟ ΤΡΟΧΑΙΑ ΑΤΥΧΗΜΑΤΑ

Πίνακας 6.4.12-1: Residual Checks στα Μοντέλα ARIMA / SARIMA Θανάτων από Τροχαία Ατυχημάτων

Τελικό Μοντέλο Θανάτων από Τροχαία Ατυχήματα	Residual Check Pass
Εβδομαδιαίο ARIMA(4,1,1) with Drift	✗
Εβδομαδιαίο SARIMA(1,1,2)(1,1,0)[48]	✗
Μηνιαίο ARIMA(3,1,2) with Drift	✗
Μηνιαίο SARIMA(2,1,1)(1,1,1)[12]	✓
Εβδομαδιαίο SARIMA(3,0,3)(0,1,3)[48]	✓
Μηνιαίο SARIMA(1,0,2)(1,1,1)[12]	✓

Πίνακας 6.4.12-2: Ενδεικτικές Τιμές Προβλέψεων Εβδομαδιαίους Μοντέλου ARIMA(4,1,1) with Drift

Date	Fatalities
07/01/2023	11
14/1/2023	12
21/01/2023	13
28/01/2023	13
04/02/2023	12
11/02/2023	12
18/02/2023	12
25/02/2023	12
04/03/2023	12
11/03/2023	12
18/03/2023	12
25/03/2023	12

Πίνακας 6.4.12-3: Ενδεικτικές Τιμές Προβλέψεων Εβδομαδιαίου Μοντέλου
SARIMA(1,1,2)(1,1,0)[48]

Date	Fatalities
07/01/2023	17
14/1/2023	12
21/01/2023	10
28/01/2023	10
04/02/2023	9
11/02/2023	6
18/02/2023	8
25/02/2023	10
04/03/2023	7
11/03/2023	10
18/03/2023	8
25/03/2023	11

Πίνακας 6.4.12-4: Ενδεικτικές Τιμές Προβλέψεων Μηνιαίου Μοντέλου ARIMA(3,1,2) with Drift

Date	Fatalities
01/01/2023	46
01/02/2023	43
01/03/2023	42
01/04/2023	43
01/05/2023	47
01/06/2023	51
01/07/2023	55
01/08/2023	58
01/09/2023	59
01/10/2023	57
01/11/2023	54
01/12/2023	50

Πίνακας 6.4.12-5: Ενδεικτικές Τιμές Προβλέψεων Μηνιαίου Μοντέλου
SARIMA(2,1,1)(1,1,1)[12]

Date	Fatalities
01/01/2023	44
01/02/2023	38
01/03/2023	39
01/04/2023	44
01/05/2023	52
01/06/2023	54
01/07/2023	65
01/08/2023	66
01/09/2023	61
01/10/2023	58
01/11/2023	50
01/12/2023	49

Πίνακας 6.4.12-6: Ενδεικτικές Τιμές Προβλέψεων Εβδομαδιαίου Μοντέλου
SARIMA(3,0,3)(0,1,3)[48]

Date	Fatalities
07/01/2023	12
14/1/2023	10
21/01/2023	11
28/01/2023	10
04/02/2023	8
11/02/2023	10
18/02/2023	9
25/02/2023	11
04/03/2023	8
11/03/2023	11
18/03/2023	10
25/03/2023	11

Πίνακας 6.4.12-7: Ενδεικτικές Τιμές Προβλέψεων Μηνιαίου Μοντέλου
SARIMA(1,0,2)(1,1,1)[12]

Date	Fatalities
01/01/2023	44
01/02/2023	38
01/03/2023	39
01/04/2023	44
01/05/2023	52
01/06/2023	55
01/07/2023	65
01/08/2023	67
01/09/2023	62
01/10/2023	59
01/11/2023	50
01/12/2023	49

Κεφάλαιο 7 Υλοποίηση NN LSTM

Η ανάγκη υλοποίησης ενός Νευρωνικού Δικτύου τύπου LSTM στην παρούσα εργασία προέκυψε από τις πολυδιάστατες και δυναμικές ιδιότητες του φαινομένου που μελετάται, δηλαδή την εξέλιξη των τροχαίων ατυχημάτων και θανάτων στο χρόνο. Οι χρονοσειρές που σχετίζονται με την οδική ασφάλεια δεν εμφανίζουν απλώς γραμμικές τάσεις ή περιοδικότητες και συχνά υπόκεινται σε έντονες εποχικές διακυμάνσεις, δομικές αλλαγές, μακροχρόνιες εξαρτήσεις και σύνθετες σχέσεις με εξωτερικούς κοινωνικοοικονομικούς και θεσμικούς παράγοντες. Τα παραδοσιακά στατιστικά μοντέλα (ARIMA/SARIMA), αν και αποτελεσματικά σε πολλές περιπτώσεις, έχουν περιορισμένη ικανότητα να συλλάβουν τέτοιες πολύπλοκες και μη γραμμικές αλληλεπιδράσεις, ειδικά όταν το υπόδειγμα εμπλουτίζεται με μεγάλο αριθμό χαρακτηριστικών (features) και προηγούμενων χρονικών τιμών (lags). Σε αυτό το πλαίσιο, η μετάβαση σε πιο ισχυρές και ευέλικτες μεθόδους μηχανικής μάθησης, όπως τα LSTM νευρωνικά δίκτυα, κρίνεται απαραίτητη.

Εκτός αυτού, η υλοποίηση LSTM επέτρεψε την αξιοποίηση του συνόλου χαρακτηριστικών που δημιουργήθηκαν μέσω του feature engineering. Το μοντέλο μπορούσε να ενσωματώσει πληροφορία που προέρχεται τόσο από χρονικές μεταβλητές (όπως εβδομάδα, εποχή, εορταστικά διαστήματα), όσο και από μακροοικονομικούς δείκτες, πολιτικές οδικής ασφάλειας, και ιστορικές επιδόσεις (lags και rolling μεταβλητές). Αυτή η πολυδιάστατη εισροή, σε συνδυασμό με τη μαθησιακή ικανότητα του LSTM, προσέφερε μια ποιοτικά αναβαθμισμένη προσέγγιση σε σχέση με τα παραδοσιακά μοντέλα, τόσο ως προς την προβλεπτική ακρίβεια όσο και ως προς την προσαρμοστικότητα στις στατιστικές ιδιομορφίες της χρονοσειράς.

7.1 Μεθοδολογία Υλοποίησης NN LSTM

Η μεθοδολογική προσέγγιση που ακολουθήθηκε για την υλοποίηση των μοντέλων LSTM στη μελέτη της πρόβλεψης τροχαίων ατυχημάτων και θανάτων βασίστηκε σε ένα πλήρες και πολυεπίπεδο σύνολο διαδικασιών, το οποίο σχεδιάστηκε ώστε να ανταποκρίνεται στις απαιτήσεις τόσο της χρονοσειριακής φύσης των δεδομένων όσο και της πολυπλοκότητας των χαρακτηριστικών. Πρώτο στάδιο αποτέλεσε η συγκέντρωση, ενοποίηση και προεπεξεργασία των δεδομένων σε εβδομαδιαία και μηνιαία βάση για όλη την περίοδο 1996–2022. Στο πλαίσιο αυτό, χρησιμοποιήθηκαν τα συνδυαστικά datasets που προέκυψαν από την μεθοδολογία προεπεξεργασίας και feature engineering των δεδομένων στο RStudio, τόσο για εβδομαδιαία όσο και για μηνιαία χρονική ανάλυση.

Το δεύτερο κρίσιμο στάδιο αφορούσε τη διαμόρφωση των dataset για την εκπαίδευση ενός LSTM δικτύου. Εδώ δόθηκε ιδιαίτερη έμφαση στην κανονικοποίηση των μεταβλητών μέσω StandardScaler, με παράλληλο σεβασμό στη φύση των binary μεταβλητών που διατηρήθηκαν ανεπηρέαστες από τη διαδικασία. Ο διαχωρισμός των δεδομένων έγινε χρονικά, με το διάστημα έως το 2017 να χρησιμοποιείται για εκπαίδευση του NN και το διάστημα από το 2018 και μετά για αξιολόγηση, εξασφαλίζοντας μια ρεαλιστική προσέγγιση πρόβλεψης στο χρόνο (out-of-sample evaluation). Η δημιουργία των ακολουθιών (sequences) για το LSTM έγινε είτε με `time_steps=1` είτε με μεγαλύτερα βάθη (π.χ. 192), ανάλογα με την εκάστοτε εφαρμογή, ώστε να διερευνηθεί η επίδραση της χρονικής μνήμης του μοντέλου στην απόδοσή του.

Στο τρίτο στάδιο, σχεδιάστηκε και υλοποιήθηκε ένα σύστημα εύρεσης βέλτιστων υπερπαραμέτρων (hyperparameter tuning) μέσω grid search, το οποίο αξιολόγησε χιλιάδες συνδυασμούς διαστάσεων μοντέλου (πλήθος LSTM και dense units), dropout, τύπου loss function (MSE/Huber), ρυθμού εκμάθησης, μεγέθους παρτίδας (batch size) και πλήθους εποχών εκπαίδευσης. Για κάθε συνδυασμό καταγράφονταν οι βασικοί δείκτες απόδοσης, MAE, RMSE, R^2 , MAPE, SMAPE, MASE, WAPE. Η δυνατότητα εκτέλεσης των πειραμάτων σε GPU Environment με χρήση του εργαλείου CUDA της Nvidia αξιοποιήθηκε πλήρως, μειώνοντας σημαντικά τον χρόνο υπολογισμού και επιτρέποντας την εκτέλεση ακόμη και των πιο υπολογιστικά σύνθετων παραμετρικών συνδυασμών.

Μετά την επιλογή των βέλτιστων υπερπαραμέτρων, η μεθοδολογία προχώρησε στην πλήρη εκπαίδευση του LSTM μοντέλου στο σύνολο των δεδομένων εκπαίδευσης.

Ακολουθώντας την αρχιτεκτονική που είχε αποδώσει καλύτερα στο grid search, το τελικό μοντέλο αξιολογήθηκε σε unseen δεδομένα (εβδομάδες 2018–2022), με στόχο να μετρηθεί η γενικευσιμότητά του. Παράλληλα, παραχθήκαν διαγράμματα Actual vs Predicted, scatter plots και ιστογράμματα σφάλματος, για την ποιοτική αξιολόγηση του μοντέλου.

Τέλος, με σκοπό την πρακτική αξιοποίηση του μοντέλου, υλοποιήθηκε μια αναδρομική διαδικασία πρόβλεψης (recursive forecasting) για την περίοδο 2023–2026. Η διαδικασία αυτή στηρίζεται στο γεγονός ότι, σε προβλεπτικό σενάριο, τα μελλοντικά χαρακτηριστικά είναι άγνωστα και πρέπει να ανακατασκευαστούν δυναμικά με βάση τις ίδιες τις προβλέψεις του μοντέλου. Έτσι, κάθε προβλεπόμενη τιμή χρησιμοποιείται ως είσοδος για τον υπολογισμό των εξαρτημένων μεταβλητών, των rolling/lagged χαρακτηριστικών και όλων των σχετικών δεικτών για την επόμενη χρονική στιγμή.

7.1.1 DATASET PREPARATION KAI HYPERPARAMETER GRID SEARCH

Η επιλογή και ρύθμιση των υπερπαραμέτρων (hyperparameters) σε ένα νευρωνικό δίκτυο τύπου LSTM αποτελεί έναν από τους πλέον κρίσιμους παράγοντες για την επιτυχία του τελικού μοντέλου πρόβλεψης. Στην παρούσα φάση της ανάλυσης, η προσοχή στράφηκε στην ανάπτυξη και αξιολόγηση ενός μεγάλου συνόλου υποψήφιων LSTM μοντέλων μέσω grid search. Αντικείμενο της διαδικασίας ήταν η βελτιστοποίηση της αρχιτεκτονικής και των παραμέτρων εκπαίδευσης του μοντέλου στο περιβάλλον της Python με χρήση της βιβλιοθήκης PyTorch.

Η αρχιτεκτονική που επιλέχθηκε για τα μοντέλα αποτελείται από δύο διαδοχικά LSTM layers με dropout, ακολουθούμενα από ένα τριπλό fully connected μπλοκ (με δύο κρυφά επίπεδα και ένα τελικό output layer), με συνάρτηση ενεργοποίησης την ReLU. Η χρήση δύο επιπέδων LSTM επιτρέπει τη μοντελοποίηση πιο σύνθετων χρονικών σχέσεων και δυναμικών στη σειρά των παρατηρήσεων, ενώ τα ενδιάμεσα dense layers προσφέρουν την απαραίτητη μη γραμμικότητα για να αποτυπωθούν πολυπαραγοντικές επιδράσεις. Ο αριθμός των μονάδων στα LSTM και dense επίπεδα, όπως και το dropout ποσοστό, δεν επιλέγονται αυθαίρετα, αλλά αποτελούν αντικείμενο της grid search διαδικασίας.

Ο συνδυαστικός έλεγχος των υπερπαραμέτρων υλοποιήθηκε με χρήση της `itertools.product()`, επιτρέποντας τον εξαντλητικό σχηματισμό όλων των πιθανών συνδυασμών των εξής παραμέτρων: αριθμός βημάτων χρόνου (time steps), LSTM και Dense units, dropout, loss function (MSE ή Huber), ρυθμός εκμάθησης (learning rate), μέγεθος batch και αριθμός

εποχών (epochs). Για κάθε συνδυασμό παραμέτρων, πραγματοποιείται ξεχωριστή εκπαίδευση του μοντέλου, πρόβλεψη στις τιμές test set και αξιολόγηση με επτά συνολικά μετρικές: MAE, RMSE, R^2 , MAPE, SMAPE, MASE και WAPE.

Η υλοποίηση των μοντέλων έγινε εξολοκλήρου στο Pycharm IDE με την βιβλιοθήκη PyTorch με αξιοποίηση της υποστήριξης CUDA για χρήση GPU. Πιο συγκεκριμένα, η διαδικασία εκτελέστηκε με χρήση κάρτας γραφικών Nvidia GTX 1660 Super, γεγονός που επιτάχυνε δραματικά τη διαδικασία εκπαίδευσης και δοκιμής των εκατοντάδων συνδυασμών παραμέτρων. Το κάθε μοντέλο εκπαιδεύτηκε με τυπικό optimization loop σε batch mode με χρήση του Adam optimizer.

Ιδιαίτερη σημασία δόθηκε και στην ανακατασκευή των πραγματικών τιμών για σκοπούς ερμηνείας. Τα αποτελέσματα των προβλέψεων και των πραγματικών τιμών μετατράπηκαν πίσω στην αρχική τους κλίμακα μέσω αντίστροφης κανονικοποίησης (inverse transform), ώστε οι μετρικές σφάλματος να είναι άμεσα συγκρίσιμες και ερμηνεύσιμες. Κάθε εκπαιδευμένο μοντέλο αξιολογήθηκε και καταγράφηκε, μαζί με τις αντίστοιχες παραμέτρους του, σε έναν συγκεντρωτικό πίνακα αποτελεσμάτων, ο οποίος επέτρεψε την εύκολη αναγνώριση του καλύτερου μοντέλου με βάση πολλαπλά κριτήρια.

Αξίζει να σημειωθεί ότι η διαδικασία αυτή υλοποιήθηκε σε τέσσερα διαφορετικά σενάρια πρόβλεψης: για εβδομαδιαία και μηνιαία δεδομένα, τόσο για το πλήθος των ατυχημάτων όσο και για τον αριθμό των θανάτων από αυτά. Σε κάθε περίπτωση, τα μοντέλα που εκπαιδεύτηκαν και αξιολογήθηκαν ήταν απολύτως ανεξάρτητα, ενώ η ίδια δομή και μεθοδολογική συνέπεια εφαρμόστηκε ομοιόμορφα.

```
281 #Hyperparameter Grid
282
283 time_steps_list = [1]
284 lstm_units_list = [64, 128, 256]
285 dense_units_list = [64, 128, 256]
286 dropout_list = [0.2, 0.3]
287 loss_list = ['mse', 'huber']
288 learning_rates = [0.001, 0.005]
289 batch_size_list = [16, 32, 64]
290 epoch_list = [50, 100, 200]
291 results = []
292
```

Σχήμα 7.1.1-1: Hyperparameter Grid για το Grid Search Βέλτιστου Μοντέλου

7.1.2 TRAINING KAI PREDICTIONS TOY LSTM ΒΑΣΗ ΤΩΝ ΒΕΛΤΙΣΤΩΝ HYPERPARAMETERS

Η διαδικασία εκπαίδευσης και αρχικής πρόβλεψης του τελικού LSTM μοντέλου πραγματοποιήθηκε σε ξεχωριστό pipeline, μετά την επιλογή του βέλτιστου συνδυασμού υπερπαραμέτρων μέσω grid search. Σε αυτό το στάδιο, το εκάστοτε μοντέλο επανεκπαιδεύεται με τις κατάλληλες παραμέτρους στο πλήρες εκπαιδευτικό σύνολο (1996 - 2017) και στη συνέχεια αξιολογείται πάνω στο test set (2018 - 2022), ενώ παράλληλα πραγματοποιείται και η πρώτη πραγματική πρόβλεψη για την αμέσως επόμενη χρονική μονάδα (εβδομάδα ή μήνας αντίστοιχα).

Το εκπαιδευτικό pipeline αξιοποιεί εκ νέου το προκαθαρισμένο dataset και επανακανονικοποιεί τα δεδομένα βάσει των επιλεγμένων χαρακτηριστικών. Διαχωρίζονται οι αριθμητικές μεταβλητές από τις δυαδικές, όπως και πριν, ώστε να μην αλλοιωθεί η φύση τους κατά την κανονικοποίηση. Η διαδικασία δημιουργίας ακολουθιών (create_sequences) προσαρμόζεται στο επιλεγμένο πλήθος χρονικών βημάτων, ενώ γίνεται χρήση τυποποιημένων TensorDataset και DataLoader για τη διαχείριση των batches εισόδου. Ολόκληρη η εκπαίδευση

υλοποιείται όπως και προαναφέρθηκε με GPU Nvidia GTX 1660 Super με χρήση του CUDA Toolkit για μέγιστη αποδοτικότητα.

Το ίδιο το μοντέλο LSTM βασίζεται στην επιλεγμένη αρχιτεκτονική δύο διαδοχικών επιπέδων LSTM, ενδιάμεσων dropout layers και ενός τριπλού πυκνού μπλοκ (dense layers) στο τελικό μέρος. Η εκπαίδευση γίνεται με χρήση του Adam optimizer και μιας συνάρτησης κόστους (είτε MSE είτε Huber), ανάλογα με τις παραμέτρους που έχουν προκύψει από το grid search. Το μοντέλο εκπαιδεύεται για προκαθορισμένο αριθμό εποχών και καταγράφεται η πορεία του loss ανά εποχή.

Μετά την εκπαίδευση, το μοντέλο αξιολογείται στο test set με το πλήρες σύνολο μετρικών απόδοσης που αναφέρθηκε προηγουμένως. Οι προβλέψεις και τα πραγματικά δεδομένα επανέρχονται στην αρχική τους κλίμακα με χρήση της inverse_transform, ενώ αποτυπώνονται και σε γραφήματα που επιτρέπουν την απεικόνιση της ακρίβειας του μοντέλου, την κατανομή των σφαλμάτων, τη διασπορά τους και τη χρονική εξέλιξη του μέσου σφάλματος με rolling μέσο όρο.

Τέλος, η εκπαιδευμένη εκδοχή του μοντέλου αξιοποιείται για να πραγματοποιήσει μία άμεση πρόβλεψη για την επόμενη χρονική περίοδο, την πρώτη εβδομάδα ή τον πρώτο μήνα του 2023, αναλόγως της χρονικής ανάλυσης. Η πρόβλεψη αυτή γίνεται χρησιμοποιώντας την τελευταία διαθέσιμη παρατήρηση από τα δεδομένα του training set.

```

=====
Layer (type:depth-idx)      Output Shape      Param #
=====
LSTMModel                  [64, 1]           --
├─LSTM: 1-1                 [64, 1, 256]      349,184
├─Dropout: 1-2              [64, 1, 256]      --
├─LSTM: 1-3                 [64, 1, 256]      526,336
├─Dropout: 1-4              [64, 1, 256]      --
├─Linear: 1-5               [64, 256]         65,792
├─ReLU: 1-6                 [64, 256]         --
├─Linear: 1-7               [64, 128]         32,896
├─ReLU: 1-8                 [64, 128]         --
├─Linear: 1-9               [64, 1]           129
=====
Total params: 974,337
Trainable params: 974,337
Non-trainable params: 0
Total mult-adds (Units.MEGABYTES): 62.36
=====
Input size (MB): 0.02
Forward/backward pass size (MB): 0.46
Params size (MB): 3.90
Estimated Total Size (MB): 4.38
=====

```

Σχήμα 7.1.2-1: Αρχιτεκτονική NN LSTM

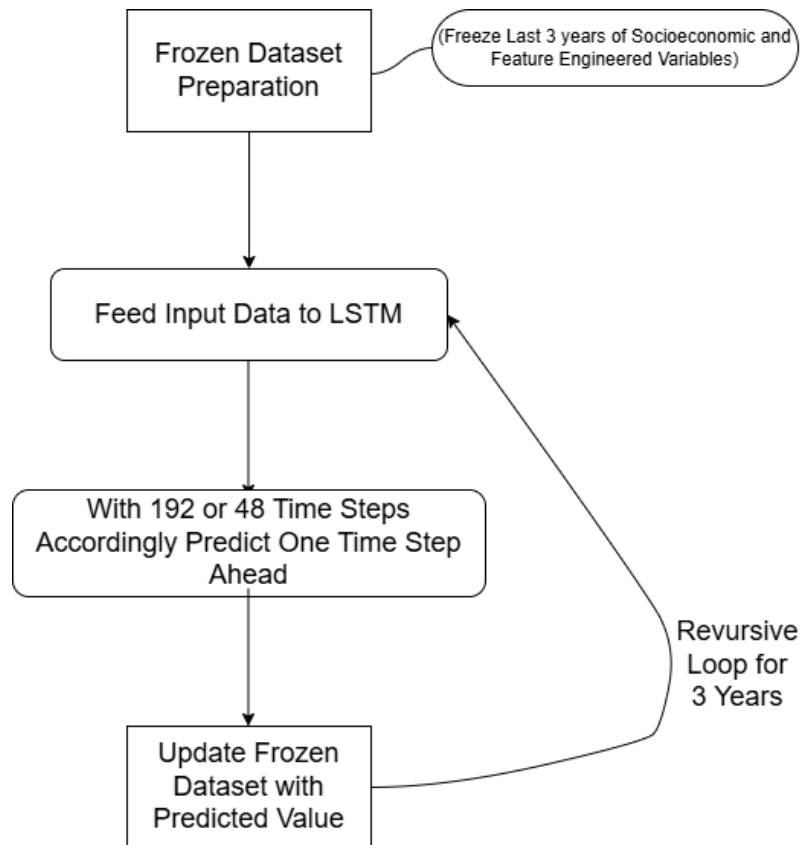
7.1.3 ΑΝΑΔΡΟΜΙΚΗ ΠΡΟΒΛΕΨΗ ΚΑΙ ΣΤΑΣΙΜΑ ΣΥΝΟΛΑ ΔΕΔΟΜΕΝΩΝ

Η μέθοδος του recursive forecasting εφαρμόστηκε στην συνέχεια για την πρόβλεψη μακροχρόνιων χρονοσειρών, με στόχο να επεκταθούν οι προβλέψεις έως το τέλος του 2026. Η διαδικασία ακολουθήθηκε συστηματικά σε τέσσερις ανεξάρτητες περιπτώσεις: εβδομαδιαία και μηνιαία χρονοσειρά για τα τροχαία ατυχήματα και αντίστοιχα εβδομαδιαία και μηνιαία χρονοσειρά για τους θανάτους από αυτά. Η επιλογή αυτής της μεθοδολογίας οφείλεται στην ανάγκη για δημιουργία ρεαλιστικών σεναρίων πρόβλεψης, όπου κάθε προβλεπόμενη τιμή χρησιμοποιείται ως είσοδος για την πρόβλεψη της επόμενης χρονικής στιγμής, επιτρέποντας έτσι την αναπαραγωγή της δυναμικής του φαινομένου με αυξανόμενη χρονική εμβέλεια.

Βασικό χαρακτηριστικό της μεθόδου είναι η χρήση ενός frozen dataset, το οποίο αποτελεί μια προκατασκευασμένη χρονική δομή δεδομένων που περιλαμβάνει όχι μόνο το παρελθόν (π.χ. έτη 1996–2022) αλλά και "παγωμένες" εγγραφές για τις χρονικές στιγμές του μέλλοντος, δηλαδή από το 2023 έως το 2026. Ο όρος "frozen" αναφέρεται στο γεγονός ότι όλες οι κοινωνικοοικονομικές μεταβλητές και αυτές που προέκυψαν από το feature engineering, όπως lags, rolling averages, rolling sums, σωρευτικά μεγέθη cumulative features, έχουν ήδη υπολογιστεί ή αντιγραφεί για το μέλλον, με σκοπό να παρέχουν σταθερή βάση εισόδου στις επερχόμενες προβλέψεις. Οι μεταβλητές αυτές είτε δεν αλλάζουν με την πάροδο του χρόνου (όπως πληθυσμός ή θεσμικές μεταρρυθμίσεις), είτε καθορίζονται σε προηγούμενα χρονικά βήματα και μπορούν να μεταφέρονται ή να υπολογίζονται εκ νέου σε κάθε βήμα πρόβλεψης.

Η διαδικασία ξεκινά με την εκπαίδευση του LSTM μοντέλου στα ιστορικά δεδομένα, χρησιμοποιώντας παράθυρο χρονικής οπισθοδρόμησης τριών χρόνων, ώστε να αξιοποιούνται οι μακροχρόνιες εξαρτήσεις. Κατά τη διάρκεια της recursive πρόβλεψης, για κάθε χρονικό βήμα από το 2023 και μετά, το μοντέλο λαμβάνει ως είσοδο το σύνολο των μεταβλητών των προηγούμενων χρονικών βημάτων από το frozen dataset, το οποίο ενημερώνεται σταδιακά. Η πρόβλεψη που παράγεται για την τρέχουσα εβδομάδα ή μήνα χρησιμοποιείται άμεσα ως είσοδος για το επόμενο χρονικό βήμα, ενώ παράλληλα ενημερώνονται όλες οι εξαρτώμενες μεταβλητές (όπως τα lags, οι rolling μέσοι όροι, τα cumulatives, κ.λπ.) βάσει της νέας τιμής.

Η διαδικασία συνεχίζεται διαδοχικά για όλα τα μελλοντικά χρονικά βήματα έως το τέλος του 2026, καλύπτοντας 192 εβδομάδες ή 48 μήνες, ανάλογα με το επίπεδο της χρονικής ανάλυσης. Το τελικό αποτέλεσμα αποθηκεύεται σε ξεχωριστά αρχεία πρόβλεψης και μπορεί να συνδυαστεί με τα ιστορικά δεδομένα για τη δημιουργία συνεχών χρονοσειρών.



Σχήμα 7.1.3-1: Recursive Predictions Flowchart

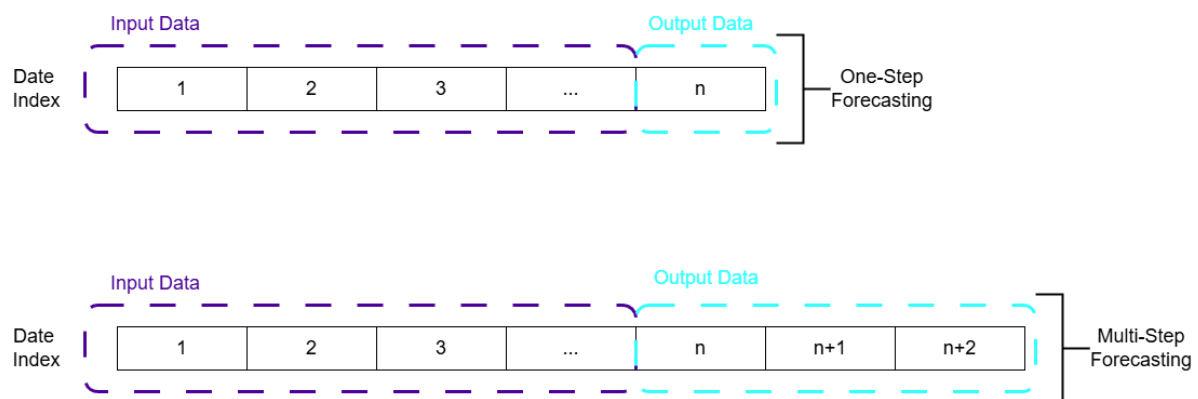
7.1.4 ONE-STEP VS MULTI-STEP FORECASTING

Εκτός των παραπάνω, δοκιμάστηκε επιπλέον μία διαφορετική προσέγγιση πρόβλεψης, η multi-step forecasting (πολλαπλών βημάτων). Στο multi-step forecasting το μοντέλο καλείται να προβλέψει μια ολόκληρη ακολουθία μελλοντικών τιμών (π.χ. τις επόμενες 4 ή 12 περιόδους) σε μία μόνο πρόβλεψη και όχι μόνο μία μελλοντική τιμή όπως γίνεται στο one-step forecasting.

Η βασική πρόκληση που προκύπτει με την μέθοδο του multi-step forecasting είναι η συσσώρευση σφάλματος. Καθώς αυξάνεται ο χρονικός ορίζοντας των προβλέψεων, το μοντέλο δυσκολεύεται να διατηρήσει ακρίβεια στις εκτιμήσεις του, ιδιαίτερα όταν οι προβλεπόμενες τιμές εξαρτώνται από μεταβλητές που επίσης προβλέπονται και όχι από πραγματικά παρατηρημένα δεδομένα. Αυτό δημιουργεί ένα είδος σύνθετης εξάρτησης, όπου το μοντέλο στηρίζεται σε δικές του εκτιμήσεις που μπορεί να είναι ανακριβείς, και αυτές με τη σειρά τους επηρεάζουν τις επόμενες. Αυτό το φαινόμενο μεταφράστηκε σε αρκετά υποβαθμισμένες

επιδόσεις των μοντέλων. Δηλαδή, οι τιμές των δεικτών R^2 ήταν χαμηλές, ενώ τα σφάλματα MAPE και SMAPE ήταν σημαντικά υψηλότερα συγκριτικά με τις one-step προβλέψεις.

Η διαπίστωση αυτών των περιορισμών οδήγησε στην απόφαση να ακολουθηθεί μία πιο αξιόπιστη προσέγγιση: η εφαρμογή one-step forecasting για μία μόνο μελλοντική πρόβλεψη και το recursive forecasting για προβλέψεις έως το 2026. Η επιλογή αυτής της προσέγγισης one-step και recursive forecasting συνέβαλε στο να παραμείνει η μεθοδολογία συνεπής και επεκτάσιμη σε μεγάλο βάθος χρόνου.



Σχήμα 7.1.4-1: Διαφορά Μεταξύ One-Step vs Multi-Step Forecasting

7.2 Αποτελέσματα Εφαρμογής NN LSTM

Σε αυτό το στάδιο παρουσιάζονται αναλυτικά τα αποτελέσματα των LSTM μοντέλων που εφαρμόστηκαν σε όλα τα επεξεργασμένα σύνολα δεδομένων, τόσο για τα τροχαία ατυχήματα όσο και για τους θανάτους από αυτά, σε εβδομαδιαία και μηνιαία βάση. Για κάθε ένα από αυτά τα datasets, καταγράφονται με λεπτομέρεια οι βέλτιστες υπερπαραμέτροι που προέκυψαν μέσα από το grid search, με στόχο τη βελτιστοποίηση της απόδοσης των μοντέλων. Τα hyperparameters που αναφέρονται περιλαμβάνουν, μεταξύ άλλων, τον αριθμό των χρονικών βημάτων (time steps), τις μονάδες του LSTM και των πλήρως συνδεδεμένων στρωμάτων, τα ποσοστά dropout, τη συνάρτηση κόστους, το learning rate, καθώς και τις παραμέτρους εκπαίδευσης όπως το batch size και τα epochs.

7.2.1 ONE-STEP ΠΡΟΒΛΕΨΗ ΕΒΔΟΜΑΔΙΑΙΩΝ ΤΡΟΧΑΙΩΝ ΑΤΥΧΗΜΑΤΩΝ

Πίνακας 7.2.1-1: Καλύτερες Hyperparameters από το Grid Search για το LSTM με το Εβδομαδιαίο Dataset Τροχαίων Ατυχημάτων

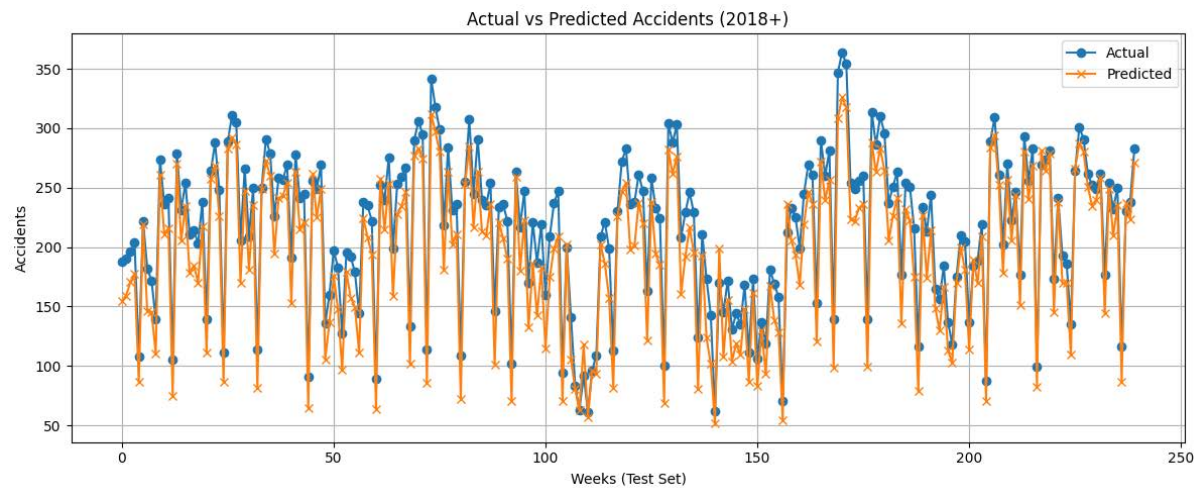
<u>Hyperparameter</u>	<u>Value</u>
Time Steps	1
LSTM Units	256
Dense Units	256
Dropout	0.2
Loss Function	MSE
Learning Rate	0.001
Batch Size	64
Epochs	200

Πίνακας 7.2.1-2: Μετρικές LSTM NN για το Εβδομαδιαίο Dataset Τροχαίων Ατυχημάτων

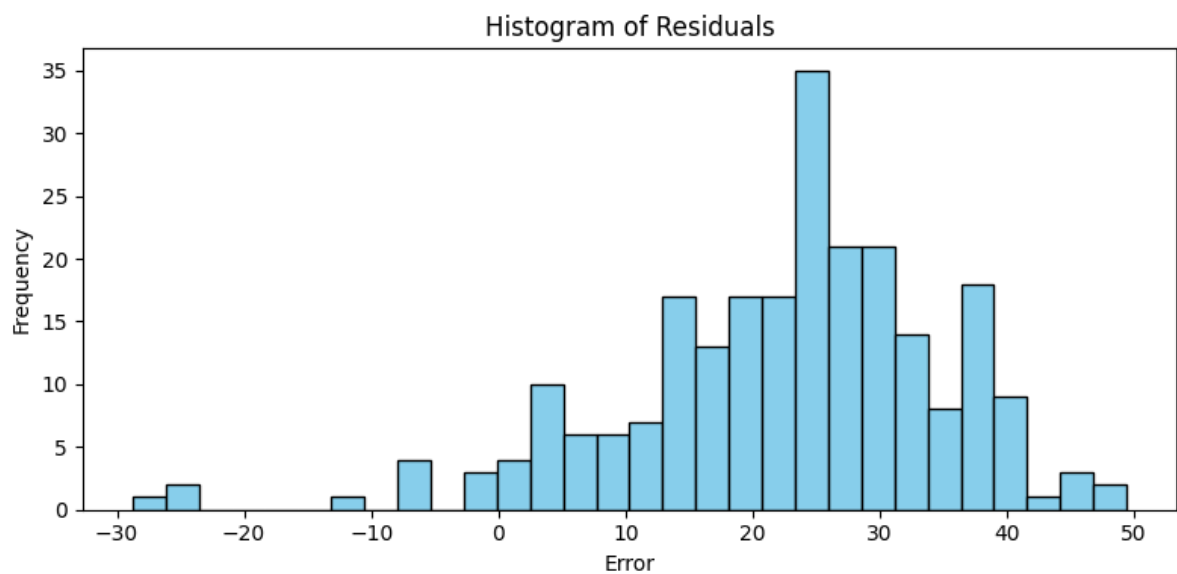
<u>Metric</u>	<u>Value</u>
MAE	15.54
RMSE	20.10
R^2	0.9006
MAPE	10.44%
SMAPE	9.17%
MASE	0.1839
WAPE	7.25

Πίνακας 7.2.1-3: Πρόβλεψη Τροχαίων Ατυχημάτων με το LSTM NN για την 1^η εβδομάδα του 2023

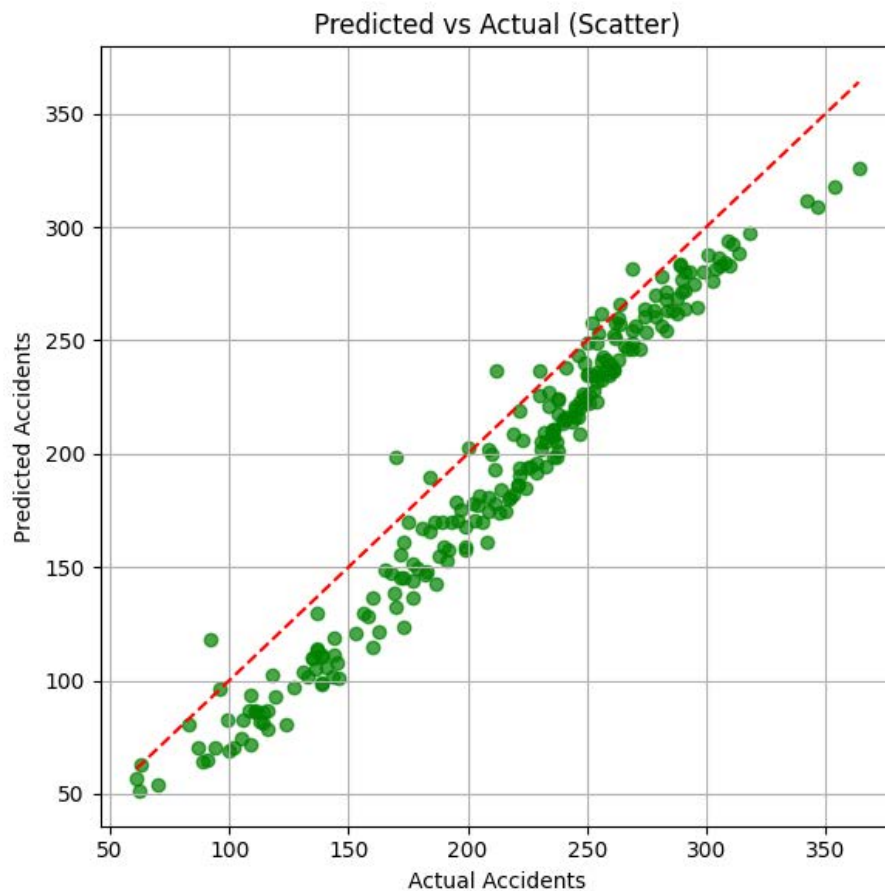
<u>Prediction</u>	<u>Value</u>
Week 1 of 2023	272



Σχήμα 7.2.1-1: Διάγραμμα Actual vs Predicted του LSTM NN για το Εβδομαδιαίο Dataset Τροχαίων Ατυχημάτων



Σχήμα 7.2.1-2: Ιστόγραμμα των Residuals του LSTM NN για το Εβδομαδιαίο Dataset Τροχαίων Ατυχημάτων



Σχήμα 7.2.1-3: Scatter Plot των Predicted vs Actual του LSTM NN για το Εβδομαδιαίο Dataset Τροχαίων Ατυχημάτων

Οι μετρικές αξιολόγησης του LSTM μοντέλου, όπως παρουσιάζονται στον Πίνακα 7.2.1-2, καταδεικνύουν πολύ ικανοποιητική απόδοση για το εβδομαδιαίο σύνολο δεδομένων τροχαίων ατυχημάτων. Συγκεκριμένα, ο δείκτης μέσου απόλυτου σφάλματος (MAE) ανέρχεται στις 15.54 μονάδες, γεγονός που σημαίνει ότι κατά μέσο όρο οι προβλέψεις αποκλίνουν κατά περίπου 16 ατυχήματα από τις πραγματικές τιμές. Το RMSE, το οποίο δίνει μεγαλύτερη έμφαση σε ακραίες αποκλίσεις, διαμορφώνεται στις 20.10 μονάδες. Η τιμή του συντελεστή προσδιορισμού R^2 φτάνει το 0.9006, αποδεικνύοντας ότι το μοντέλο εξηγεί το 90% της διακύμανσης των εβδομαδιαίων ατυχημάτων στην περίοδο ελέγχου. Επιπλέον, οι ποσοστιαίες μετρικές MAPE και SMAPE παραμένουν σε χαμηλά επίπεδα, 10.44% και 9.17% αντίστοιχα, ενισχύοντας την αξιοπιστία του μοντέλου στη σχετική απόδοση. Τέλος, η εξομάλυνση ως προς την παύση πρόβλεψη ($MASE = 0.1839$) και το σταθμισμένο απόλυτο σφάλμα ($WAPE = 7.25\%$) δείχνουν ότι το LSTM υπερέχει σημαντικά.

Το συγκεκριμένο αποτέλεσμα επιτεύχθηκε με ένα σύνολο προσεκτικά επιλεγμένων υπερπαραμέτρων που εντοπίστηκαν μέσω εκτενούς grid search στο προηγούμενο στάδιο της ανάλυσης. Συγκεκριμένα, η αρχιτεκτονική του μοντέλου βασίστηκε σε ένα one-step forecasting πλαίσιο με χρονικό παράθυρο ενός βήματος (Time Steps = 1), γεγονός που διευκόλυνε τη σταθερότητα των προβλέψεων. Ο αριθμός μονάδων LSTM καθώς και ο αριθμός νευρώνων στο πλήρως συνδεδεμένο στρώμα (Dense Units) ορίστηκαν και στις δύο περιπτώσεις στις 256, υποστηρίζοντας την επαρκή εκπροσώπηση της πολύπλοκης σχέσης μεταξύ κοινωνικών, οικονομικών και εποχικών μεταβλητών και του αριθμού ατυχημάτων. Το dropout ποσοστό 0.2 συνέβαλε στον περιορισμό της υπερπροσαρμογής, ενώ η λειτουργία απώλειας επιλέχθηκε ως Mean Squared Error (MSE), η οποία είναι ιδιαίτερα ευαίσθητη σε μεγάλες αποκλίσεις και ευνοεί την αποδοτικότητα στις ακραίες περιπτώσεις. Ο αλγόριθμος βελτιστοποίησης ήταν ο Adam με learning rate 0.001, ενώ το μοντέλο εκπαιδεύτηκε για 200 εποχές με batch size 64, ώστε να διασφαλιστεί επαρκής σύγκλιση χωρίς υπερβολική επιβάρυνση μνήμης ή χρόνου υπολογισμού.

Όσον αφορά τα διαγράμματα, στο Actual vs Predicted, παρατηρούμε την εξέλιξη των προβλέψεων ανά εβδομάδα. Η καμπύλη των προβλέψεων σε πολλές περιπτώσεις σχεδόν εφάπτεται με αυτή των πραγματικών τιμών, κάτι που δείχνει ότι το μοντέλο κατάφερε να αποτυπώσει και να μαθει την δομή των δεδομένων και έτσι να προβλέψει με ακρίβεια τις unseen τιμές του test set.

Στο Ιστόγραμμα των Residuals, φαίνεται η κατανομή των σφαλμάτων του συγκεκριμένου LSTM NN. Η συγκέντρωση των residuals γύρω από το εύρος των 15-30 ατυχημάτων δείχνει ότι οι περισσότερες προβλέψεις κυμαίνονται κοντά στις πραγματικές τιμές, αλλά με μια σταθερή θετική απόκλιση, υποδηλώνοντας ελαφρά υποεκτίμηση από την πλευρά του μοντέλου. Η ουρά προς τα δεξιά υποδεικνύει ότι σε ορισμένες περιπτώσεις το μοντέλο απέτυχε να αποτυπώσει αιχμές στον αριθμό των ατυχημάτων, ιδίως σε περιόδους με απότομες αυξήσεις.

Στο Scatter Plot των Predicted vs Actual, η γραμμική συσχέτιση μεταξύ των προβλεπόμενων και πραγματικών τιμών επιβεβαιώνει τη συνολική καλή προσαρμογή του μοντέλου. Η πλειοψηφία των σημείων βρίσκεται κοντά στην διαγώνιο, αποδεικνύοντας ότι το LSTM μοντέλο κατάφερε να μάθει τη σχέση μεταξύ εισόδων και στόχου με ικανοποιητική ακρίβεια.

7.2.2 ONE-STEP ΠΡΟΒΛΕΨΗ ΕΒΔΟΜΑΔΙΑΙΩΝ ΘΑΝΑΤΩΝ ΑΠΟ ΤΡΟΧΑΙΑ ΑΤΥΧΗΜΑΤΑ

Πίνακας 7.2.2-1: Καλύτερες Hyperparameters από το Grid Search για το LSTM με το Εβδομαδιαίο Dataset Θανάτων από Τροχαία Ατυχήματα

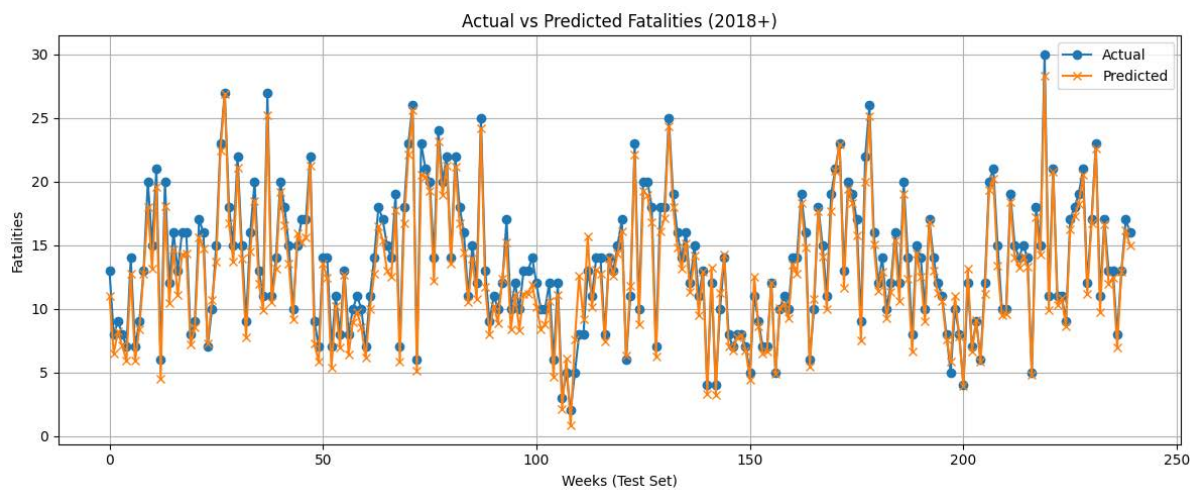
<u>Hyperparameter</u>	<u>Value</u>
Time Steps	1
LSTM Units	256
Dense Units	64
Dropout	0.2
Loss Function	MSE
Learning Rate	0.001
Batch Size	16
Epochs	50

Πίνακας 7.2.2-2: Μετρικές LSTM NN για το Εβδομαδιαίο Dataset Θανάτων από Τροχαία Ατυχήματα

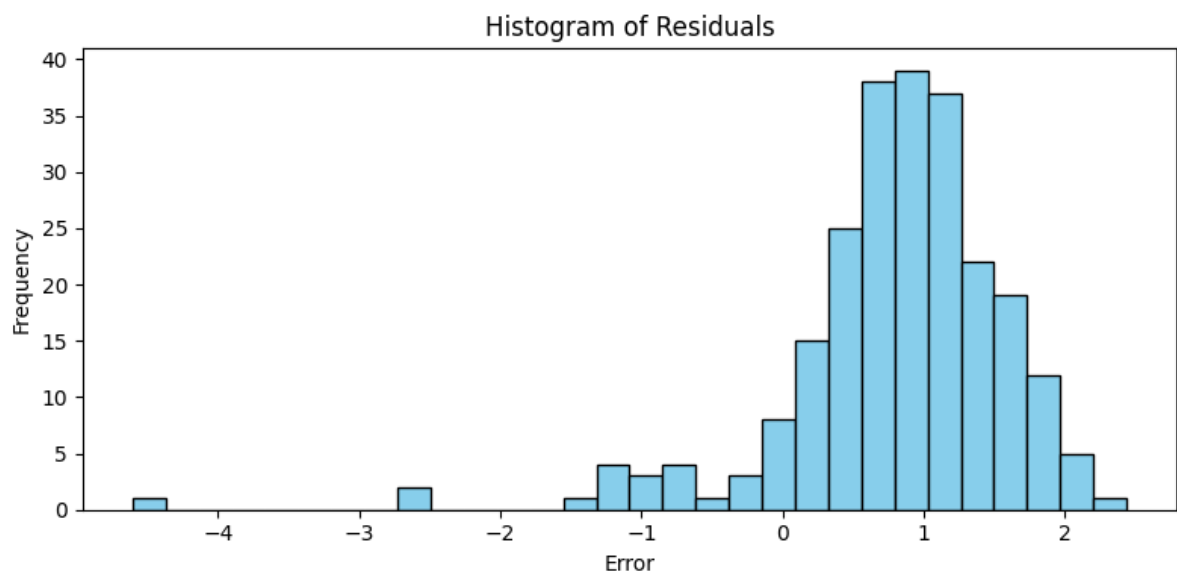
<u>Metric</u>	<u>Value</u>
MAE	0.9880
RMSE	1.1393
R ²	0.9528
MAPE	8.56%
SMAPE	8.94%
MASE	0.0994
WAPE	7.30%

Πίνακας 7.2.2-3: Πρόβλεψη Θανάτων από Τροχαία Ατυχήματα με το LSTM NN για την 1^η εβδομάδα του 2023

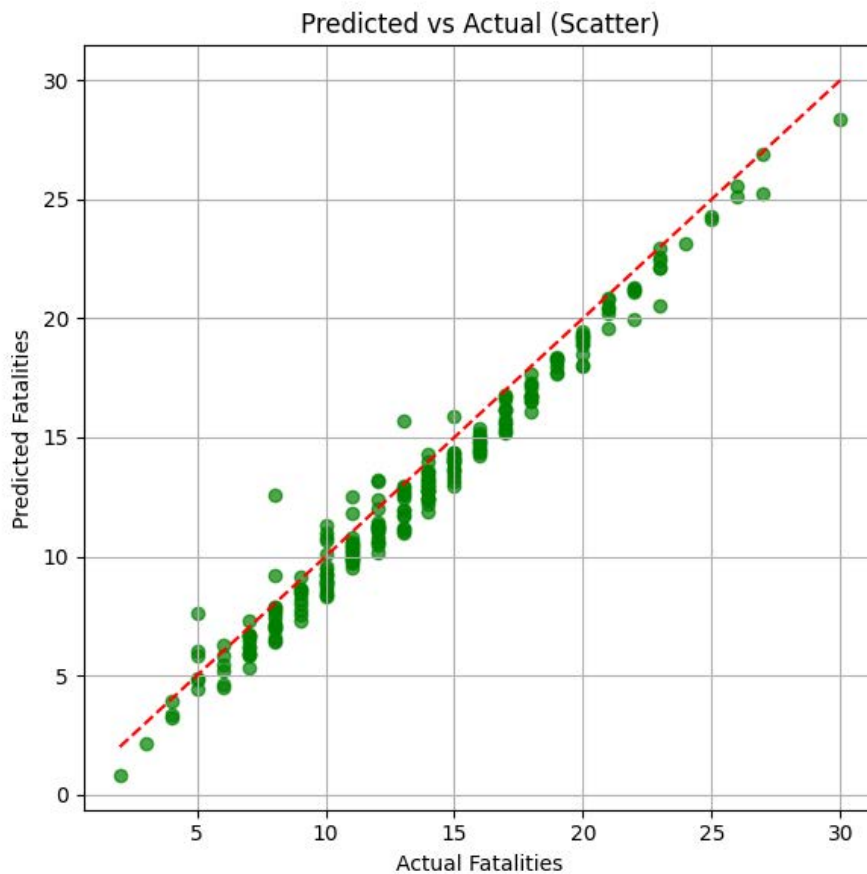
<u>Prediction</u>	<u>Value</u>
Week 1 of 2023	14



Σχήμα 7.2.2-1: Διάγραμμα Actual vs Predicted του LSTM NN για το Εβδομαδιαίο Dataset Θανάτων από Τροχαία Ατυχήματα



Σχήμα 7.2.2-2: Ιστόγραμμα των Residuals του LSTM NN για το Εβδομαδιαίο Dataset Θανάτων από Τροχαία Ατυχήματα



Σχήμα 7.2.2-3: Scatter Plot των Predicted vs Actual του LSTM NN για το Εβδομαδιαίο Dataset Θανάτων από Τροχαία Ατυχήματα

Οι ποσοτικοί δείκτες αξιολόγησης για το LSTM NN του εβδομαδιαίου dataset με τους θανάτους από τα τροχαία ατυχήματα, καταδεικνύουν εξαιρετικά αποτελέσματα. Πιο συγκεκριμένα, η τιμή του μέσου απόλυτου σφάλματος (MAE) διαμορφώνεται στις 0.988 μονάδες. Το RMSE (1.1393) παραμένει χαμηλό, επιβεβαιώνοντας ότι δεν υπάρχουν συστηματικές μεγάλες αποκλίσεις. Ο συντελεστής προσδιορισμού (R^2) φτάνει το 0.9528, γεγονός που σημαίνει ότι σχεδόν το 95% της διακύμανσης στους εβδομαδιαίους θανάτους εξηγείται από το μοντέλο. Παράλληλα, οι μετρικές MAPE και SMAPE καταγράφονται σε πολύ ικανοποιητικά επίπεδα, 8.56% και 8.94% αντίστοιχα, ενισχύοντας την εμπιστοσύνη στη σχετική ακρίβεια των προβλέψεων.

Οι υπερπαραμέτροι του μοντέλου καθορίστηκαν, όπως και προηγουμένως, έπειτα από στοχευμένη αναζήτηση στο hyperparameter grid με την διαδικασία grid search. Το LSTM εκπαιδεύτηκε για one-step forecasting, με χρονικό παράθυρο ενός βήματος (Time Steps = 1),

επιτρέποντας στο δίκτυο να επικεντρωθεί στη βέλτιστη πρόβλεψη της αμέσως επόμενης εβδομάδας. Ο αριθμός των LSTM units καθορίστηκε σε 256, ενώ το πλήρως συνδεδεμένο στρώμα περιορίστηκε σε 64 units για την αποφυγή υπερπροσαρμογής. Η λειτουργία dropout στα 0.2 προστάτευσε τη γενίκευση του μοντέλου, και η επιλογή του MSE ως loss function διασφάλισε ευαισθησία σε ακραίες αποκλίσεις. Ο αλγόριθμος Adam με learning rate 0.001 και training για 50 εποχές με batch size 16 οδήγησε σε επαρκή εκπαίδευση με χαμηλή υπολογιστική επιβάρυνση.

Η γραφική απεικόνιση της πραγματικής έναντι της προβλεπόμενης τιμής (Actual vs Predicted Fatalities) αναδεικνύει την εξαιρετική προσαρμογή του μοντέλου στα ιστορικά δεδομένα. Οι δύο καμπύλες, αυτή των πραγματικών και αυτή των προβλεπόμενων θανάτων, κινούνται σχεδόν παράλληλα για το σύνολο της περιόδου ελέγχου. Το μοντέλο καταφέρνει να παρακολουθήσει με υψηλή αξιοπιστία τις αυξομειώσεις που σχετίζονται με εποχικές μεταβολές, κοινωνικοοικονομικά γεγονότα ή θεσμικές παρεμβάσεις.

Το διάγραμμα διασποράς μεταξύ των πραγματικών και των προβλεπόμενων θανάτων ενισχύει τη διαγνωστική ισχύ της αξιολόγησης. Η πλειονότητα των σημείων συγκεντρώνεται γύρω από τη διαγώνιο της τέλει πρόβλεψης, υποδηλώνοντας ότι το μοντέλο αναπαράγει με συνέπεια τις target values χωρίς ιδιαίτερες αποκλίσεις. Η μικρή απόσταση από την ιδανική γραμμή σε ορισμένα σημεία είναι φυσιολογική, καθώς αναμένονται διακυμάνσεις σε σπάνια ή εξαιρετικά γεγονότα.

Τέλος, το ιστόγραμμα των residuals προσφέρει χρήσιμες πληροφορίες για την κατανομή των σφαλμάτων. Η έντονη συγκέντρωση γύρω από την τιμή 1 και η έλλειψη σημαντικής ουράς προς τα αρνητικά δείχνουν ελαφρά υποεκτίμηση, αλλά χωρίς σημαντικά σφάλματα ή outliers. Το μοντέλο παρουσιάζει προβλεψιμότητα με σχετικά ομοιόμορφο τρόπο, γεγονός που ενισχύει την αξιοπιστία του σε πραγματικές εφαρμογές. Συνολικά, η απόδοση του LSTM για τα εβδομαδιαία δεδομένα θανάτων χαρακτηρίζεται ως ιδιαίτερα ικανοποιητική τόσο από άποψη μετρικών όσο και από ποιοτική σύγκριση προβλεπόμενων και πραγματικών τιμών.

7.2.3 ONE-STEP ΠΡΟΒΛΕΨΗ ΜΗΝΙΑΙΩΝ ΤΡΟΧΑΙΩΝ ΑΤΥΧΗΜΑΤΩΝ

Πίνακας 7.2.3-1: Καλύτερες Hyperparameters από το Grid Search για το LSTM με το Μηνιαίο Dataset Τροχαίων Ατυχημάτων

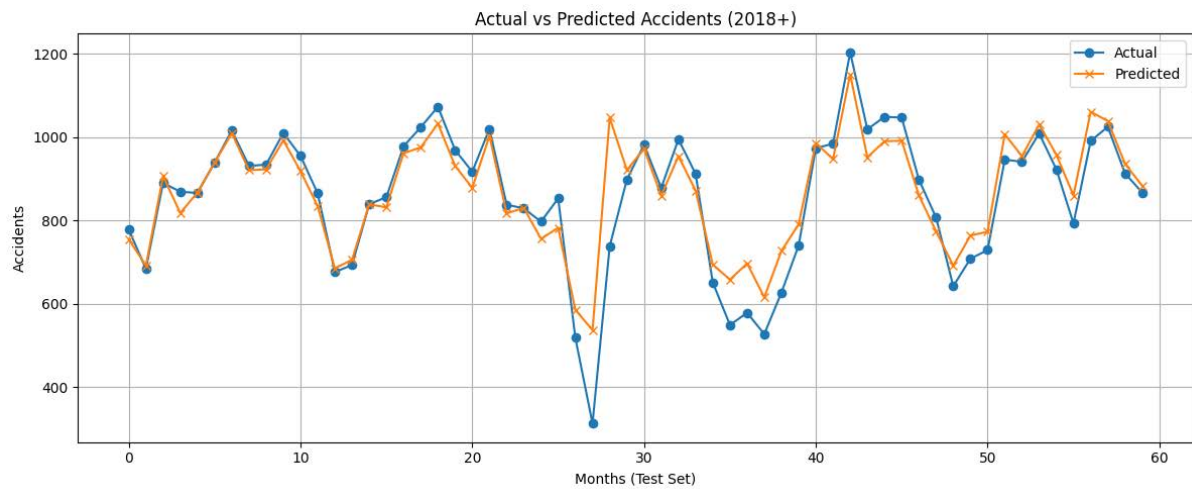
<u>Hyperparameter</u>	<u>Value</u>
Time Steps	1
LSTM Units	256
Dense Units	64
Dropout	0.2
Loss Function	MSE
Learning Rate	0.001
Batch Size	64
Epochs	200

Πίνακας 7.2.3-2: Μετρικές LSTM NN για το Μηνιαίο Dataset Τροχαίων Ατυχημάτων

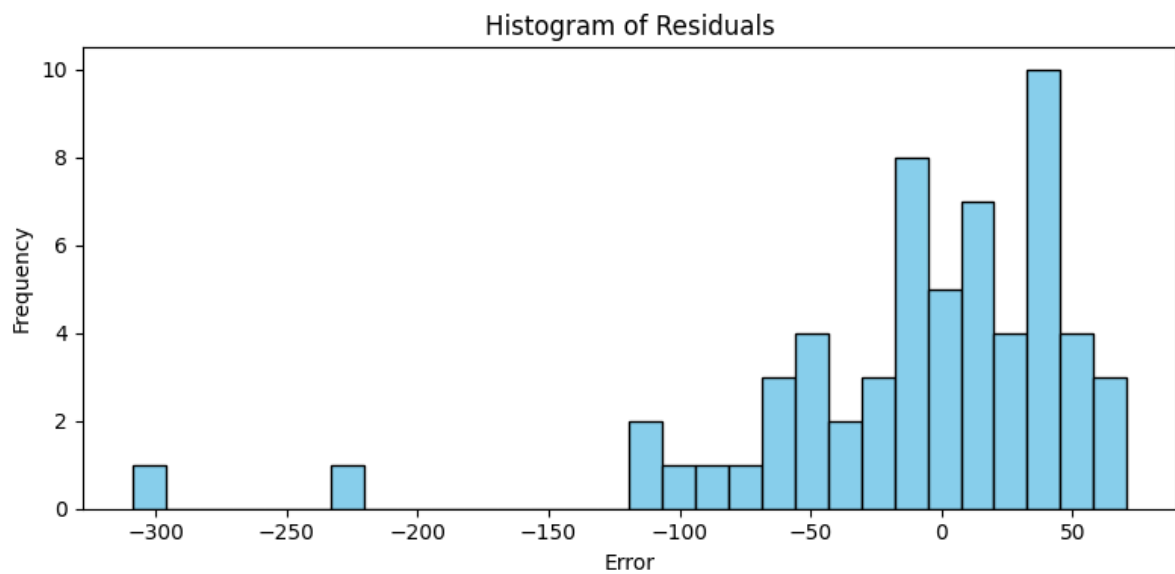
<u>Metric</u>	<u>Value</u>
MAE	44.3160
RMSE	66.5681
R ²	0.8340
MAPE	6.43%
SMAPE	5.87%
MASE	0.3685
WAPE	5.17%

Πίνακας 7.2.3-3: Πρόβλεψη Τροχαίων Ατυχημάτων με το LSTM NN για τον 1^ο μήνα του 2023

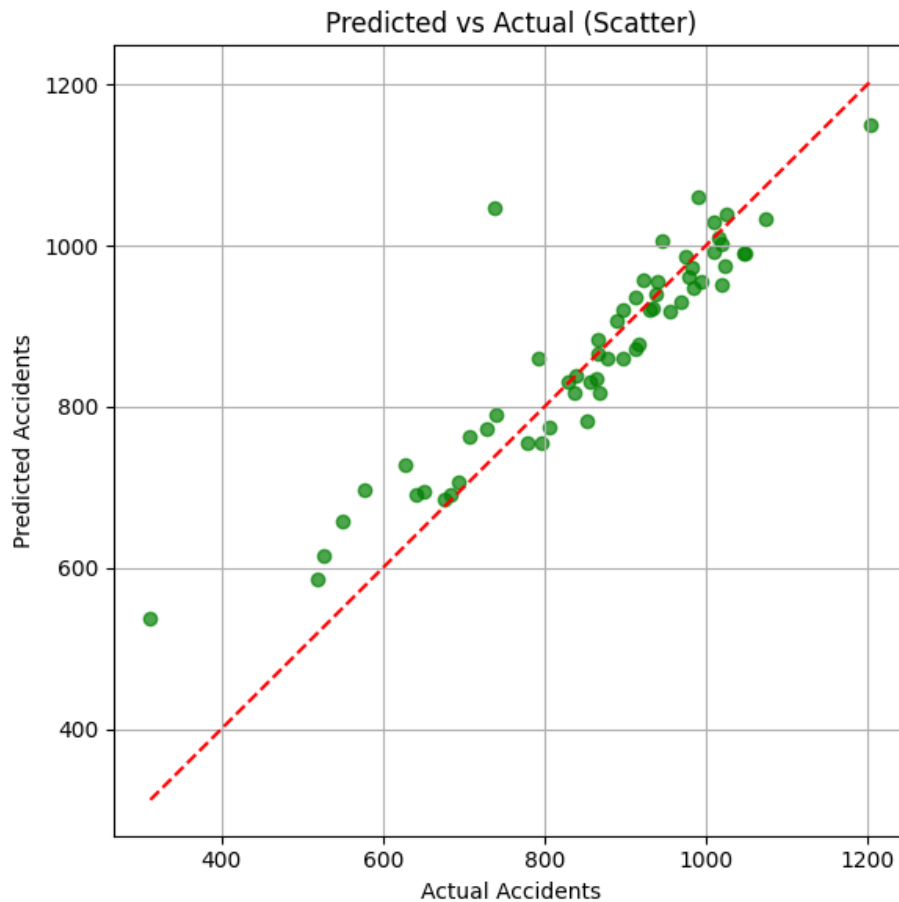
<u>Prediction</u>	<u>Value</u>
Month 1 of 2023	876



Σχήμα 7.2.3-1: Διάγραμμα Actual vs Predicted του LSTM NN για το Μηνιαίο Dataset Τροχαίων Ατυχημάτων



Σχήμα 7.2.3-2: Ιστόγραμμα των Residuals του LSTM NN για το Μηνιαίο Dataset Τροχαίων Ατυχημάτων



Σχήμα 7.2.3-3: Scatter Plot των Predicted vs Actual του LSTM NN για το Μηνιαίο Dataset Τροχαίων Ατυχημάτων

Η απόδοση του LSTM μοντέλου για το μηνιαίο σύνολο δεδομένων τροχαίων ατυχημάτων εκτιμάται ως ιδιαίτερα ικανοποιητική με βάση το σύνολο των προβλεπτικών μετρικών και την οπτική αξιολόγηση των αποτελεσμάτων. Η διαδικασία εκπαίδευσης στηρίχθηκε και πάλι, όπως και πριν, σε μια one-step προσέγγιση. Παρότι οι τιμές είναι σαφώς υψηλότερες λόγω της μηνιαίας κλίμακας (σε απόλυτα μεγέθη), το μοντέλο έδειξε σταθερότητα και καλή ευθυγράμμιση με τις πραγματικές τιμές, όπως αποτυπώνεται στα διαγράμματα και στις μετρικές.

Πιο συγκεκριμένα, το MAE των 44.32 και το RMSE των 66.57 αποτυπώνουν την απόσταση των προβλέψεων από τις πραγματικές τιμές, η οποία, δεδομένου του μεγέθους των μηνιαίων τιμών, κρίνεται ικανοποιητική. Η τιμή R^2 στο 0.8340 επιβεβαιώνει ότι το 83.4% της διακύμανσης εξηγείται από το μοντέλο. Το MAPE (6.43%) και SMAPE (5.87%) φανερώνουν μικρό σχετικό σφάλμα, ενώ η τιμή του MASE (0.3685) καταδεικνύει σημαντική υπεροχή του

μοντέλου σε σχέση με την παύση πρόβλεψη. Το WAPE 5.17% επιβεβαιώνει τη συνέπεια των προβλέψεων.

Η επίδοση αυτή επιτεύχθηκε με ένα σετ υπερπαραμέτρων που εντοπίστηκαν μέσω του grid search. Χρησιμοποιήθηκε όπως και προηγουμένως time steps = 1. Το LSTM στρώμα περιείχε 256 μονάδες, ακολουθούμενο από ένα πυκνό στρώμα 64 νευρώνων. Το dropout 0.2 προστάτευσε από overfitting, ενώ η απώλεια υπολογίστηκε με MSE για ακρίβεια σε μεγάλα λάθη. Το μοντέλο εκπαιδεύτηκε για 200 epochs με learning rate 0.001 και batch size 64.

Εκτός αυτού, η σύγκριση των πραγματικών με τις προβλεπόμενες τιμές, όπως αποτυπώνεται στο scatter plot, αποδεικνύει ισχυρή συσχέτιση ανάμεσα στις μεταβλητές-στόχους. Η πλειοψηφία των σημείων τοποθετείται κοντά στη διαγώνιο, υποδηλώνοντας ότι το μοντέλο μπορεί να γενικεύσει τις δυναμικές σχέσεις στο σύνολο δεδομένων, ακόμη και για εξάρσεις ή πτώσεις στη συχνότητα ατυχημάτων. Παρότι παρατηρούνται μερικές αποκλίσεις για ακραίες τιμές, η γενική κατανομή δεν παρουσιάζει υπερεκτίμηση ή υποεκτίμηση.

Επίσης, η χρονική κατανομή των πραγματικών και προβλεπόμενων τιμών επιβεβαιώνει την ευαισθησία του LSTM στη σύλληψη των κυκλικών και εποχικών τάσεων των μηνιαίων ατυχημάτων. Το μοντέλο καταφέρνει να εντοπίζει τις περιόδους υψηλής και χαμηλής κινητικότητας. Οι αποκλίσεις σε περιόδους έντονης διακύμανσης (όπως απότομες αυξήσεις ή μειώσεις) παραμένουν περιορισμένες και δεν επηρεάζουν σημαντικά την αξιοπιστία των συνολικών προβλέψεων.

Τέλος, η κατανομή των σφαλμάτων μέσω του ιστογράμματος των residuals αποκαλύπτει μια ελαφρώς δεξιόστροφη ασυμμετρία, αλλά με συγκέντρωση των σφαλμάτων κοντά στο μηδέν και το εύρος μεταξύ ± 50 . Αν και υπάρχουν ορισμένες μεμονωμένες ακραίες αποκλίσεις (outliers), η συνολική εικόνα υποδηλώνει σταθερή απόδοση με περιορισμένη συστηματική απόκλιση. Αυτό ενισχύει την εμπιστοσύνη ότι το μοντέλο δεν είναι προκατειλημμένο προς συγκεκριμένες τιμές και παραμένει ευέλικτο σε διαφορετικές χρονικές περιόδους.

7.2.4 ONE-STEP ΠΡΟΒΛΕΨΗ ΜΗΝΙΑΙΩΝ ΘΑΝΑΤΩΝ ΑΠΟ ΤΡΟΧΑΙΑ ΑΤΥΧΗΜΑΤΑ

Πίνακας 7.2.4-1: Καλύτερες Hyperparameters από το Grid Search για το LSTM με το Μηνιαίο Dataset Θανάτων από Τροχαία Ατυχήματα

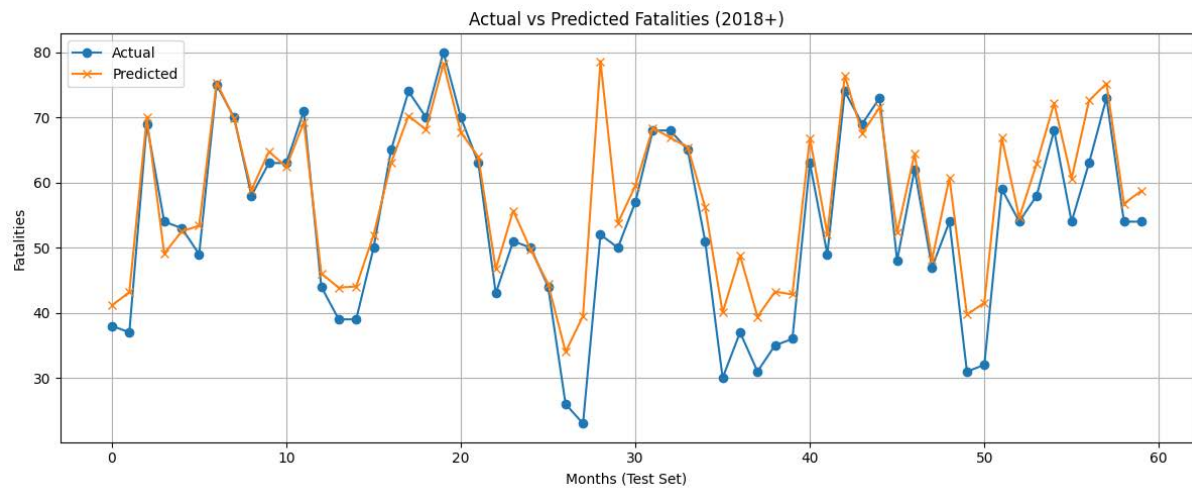
<u>Hyperparameter</u>	<u>Value</u>
Time Steps	1
LSTM Units	256
Dense Units	256
Dropout	0.2
Loss Function	huber
Learning Rate	0.001
Batch Size	64
Epochs	200

Πίνακας 7.2.4-2: Μετρικές LSTM NN για το Μηνιαίο Dataset Θανάτων από Τροχαία Ατυχήματα

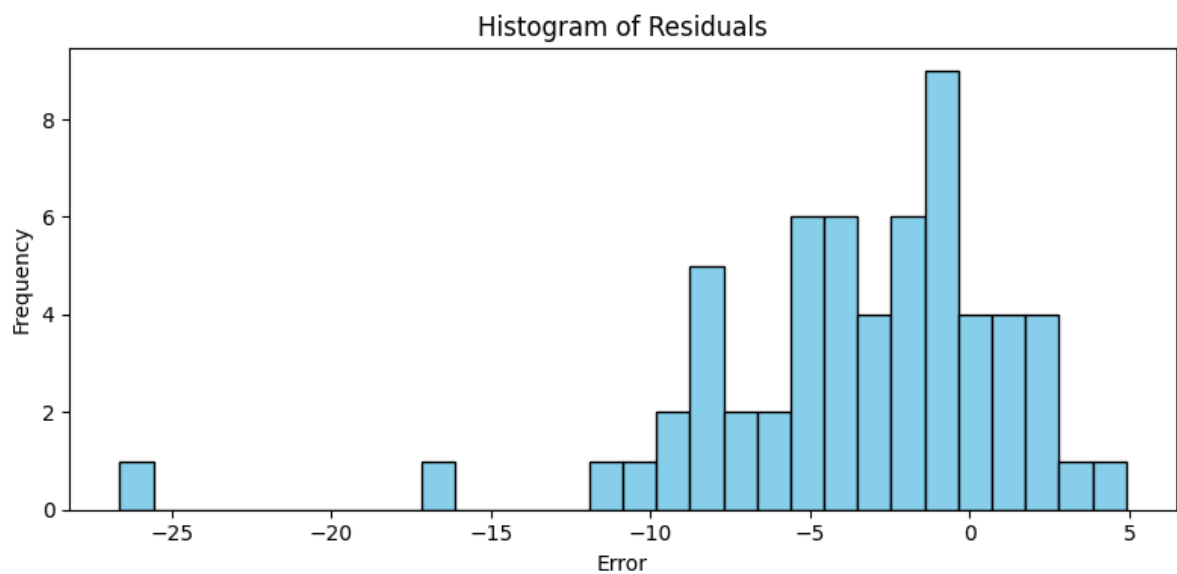
<u>Metric</u>	<u>Value</u>
MAE	4.3304
RMSE	6.1811
R ²	0.8069
MAPE	10.30%
SMAPE	9.18%
MASE	0.2352
WAPE	7.99%

Πίνακας 7.2.4-3: Πρόβλεψη Θανάτων από Τροχαία Ατυχήματα με το LSTM NN για τον 1^ο μήνα του 2023

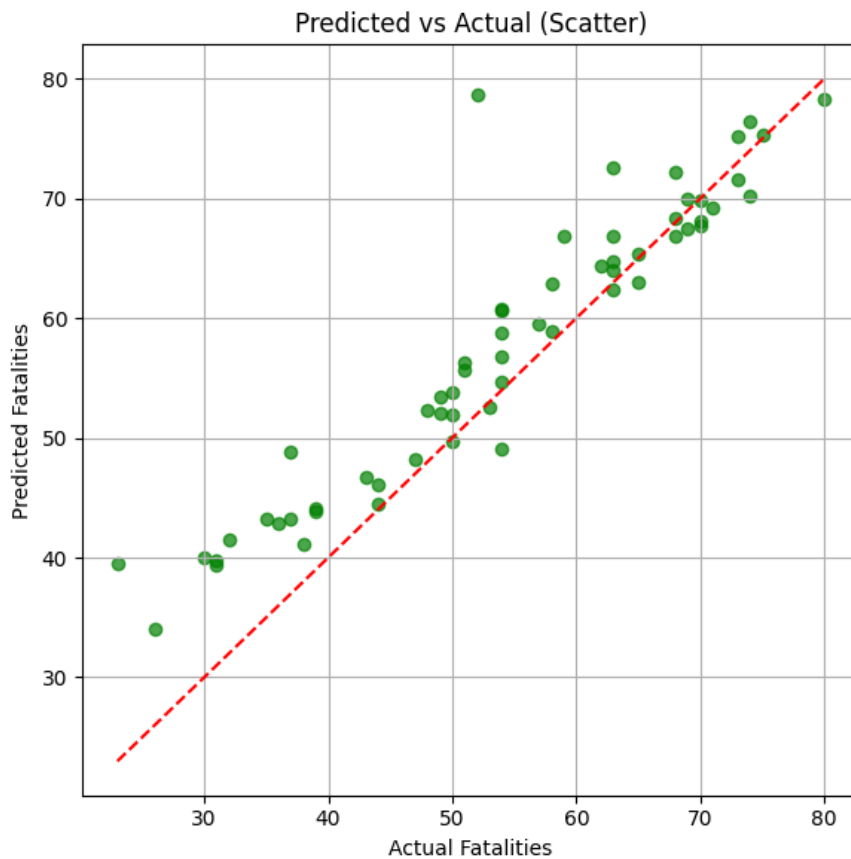
<u>Prediction</u>	<u>Value</u>
Month 1 of 2023	55



Σχήμα 7.2.4-1: Διάγραμμα Actual vs Predicted του LSTM NN για το Μηνιαίο Dataset Θανάτων από Τροχαία Ατυχήματα



Σχήμα 7.2.4-2: Ιστόγραμμα των Residuals του LSTM NN για το Μηνιαίο Dataset Θανάτων από Τροχαία Ατυχήματα



Σχήμα 7.2.4-3: Scatter Plot των Predicted vs Actual του LSTM NN για το Μηνιαίο Dataset Θανάτων από Τροχαία Ατυχήματα

Οι προβλέψεις του LSTM για το μηνιαίο σύνολο δεδομένων θανάτων από τροχαία ατυχήματα παρουσιάζουν εξίσου ικανοποιητική απόδοση, όπως τεκμηριώνεται τόσο από τις ποσοτικές μετρικές αξιολόγησης όσο και από τα διαγράμματα αποτελεσμάτων. Το μοντέλο πέτυχε MAE ίσο με 4.33 και ο δείκτης RMSE διαμορφώθηκε στις 6.18 μονάδες. Το R^2 φτάνει το 0.8069, επιβεβαιώνοντας ότι το 80.7% της συνολικής διακύμανσης των θανάτων αποδίδεται από το μοντέλο.

Οι ποσοστιαίες μετρικές MAPE (10.30%) και SMAPE (9.18%) ενισχύουν την αξιοπιστία της πρόβλεψης σε σχετική κλίμακα, αποκαλύπτοντας πως το μοντέλο έχει γενικά καλή προσαρμοστικότητα σε διαφορετικά επίπεδα τιμών. Το MASE (0.2352) υποδεικνύει ότι η απόδοση είναι σχετικά καλή, ενώ η χαμηλή τιμή WAPE (7.99%) υποδεικνύει σταθερή και αποτελεσματική πρόβλεψη σε σύγκριση με το συνολικό μέγεθος των παρατηρούμενων θανάτων.

Όσον αφορά τα διαγράμματα, η κατανομή των σφαλμάτων όπως παρουσιάζεται στο ιστόγραμμα των residuals υποδεικνύει ότι η πλειοψηφία των προβλέψεων απέχει από τις πραγματικές τιμές κατά λιγότερο από 10 θανάτους, με μικρή ασυμμετρία προς την αρνητική πλευρά. Αυτό υποδηλώνει τάση μικρής υπερεκτίμησης από το μοντέλο σε ορισμένες περιόδους. Τα λίγα ακραία αρνητικά residuals (π.χ. -25) πιθανόν αντανakλούν περιπτώσεις ξαφνικής μείωσης των θανάτων που το μοντέλο δεν κατάφερε να προβλέψει.

Η συσχέτιση μεταξύ των πραγματικών και προβλεπόμενων τιμών στο διάγραμμα διασποράς είναι εξαιρετικά έντονη. Η πλειονότητα των σημείων ευθυγραμμίζεται με τη διαγώνιο γραμμή, ενώ οι αποκλίσεις είναι σχετικά μικρές και δεν εμφανίζουν συστηματική τάση. Το μοντέλο αποδίδει σταθερά σε όλο το εύρος τιμών, τόσο σε χαμηλές όσο και σε υψηλές περιόδους θνησιμότητας. Η συγκέντρωση των σημείων σε ένα γραμμικό μοτίβο καταδεικνύει ότι το LSTM έχει μάθει τη δυναμική της χρονοσειράς.

Η χρονική εξέλιξη των προβλέψεων στο διάγραμμα Actual vs Predicted αποτυπώνει τη συνεπή συμπεριφορά του μοντέλου με τις πραγματικές τιμές. Οι κορυφώσεις της σειράς διατηρούνται με ικανοποιητική πιστότητα, χωρίς σημαντική υστέρηση ή πρόωρη αντίδραση. Οι αποκλίσεις που παρατηρούνται οφείλονται κατά κύριο λόγο σε περιόδους με έντονη μεταβλητότητα, όπου κοινωνικοοικονομικοί και πολιτικοί παράγοντες, όπως νέα μέτρα οδικής ασφάλειας ή αυξομειώσεις στην κυκλοφορία, επηρεάζουν απότομα τους θανάτους. Παρά τις προκλήσεις, το μοντέλο καταφέρνει να διατηρήσει υψηλό βαθμό συνοχής και να προσφέρει αξιόπιστες μηνιαίες προβλέψεις.

Η παραπάνω επίδοση, επιτεύχθηκε με χρήση μιας αποδοτικής αρχιτεκτονικής LSTM. Το μοντέλο εκπαιδεύτηκε με χρονικό βήμα ενός μήνα (Time Steps = 1), ενώ τόσο τα LSTM όσο και τα dense units χρησιμοποιούν 256 νευρώνες, προσφέροντας τη δυνατότητα επεξεργασίας πολύπλοκων συσχετίσεων μεταξύ εισόδων και στόχου. Η χρήση Huber loss αντί της MSE χρησιμοποιήθηκε για να περιοριστεί η ευαισθησία σε ακραίες τιμές, βελτιώνοντας τη γενίκευση. Η εκπαίδευση σε 200 epochs με batch size 64 εξασφάλισε επαρκή αριθμό επαναλήψεων για σταθερή σύγκλιση, ενώ το σταθερό learning rate 0.001 και η χρήση του αλγόριθμου Adam συνέβαλαν σε ομαλή και αποτελεσματική ενημέρωση των βαρών.

7.3 Υλοποίηση του Recursive Forecasting

Για την τελική φάση της ανάλυσης και πρόβλεψης, υλοποιήθηκε η μέθοδος recursive forecasting για όλα τα μοντέλα που είχαν προηγουμένως εκπαιδευτεί με one-step προσέγγιση. Σε αυτό το στάδιο, ο στόχος δεν ήταν πλέον μόνο η πρόβλεψη της επόμενης χρονικής στιγμής, αλλά η επέκταση της πρόβλεψης στο μέλλον για μια εκτεταμένη χρονική περίοδο που καλύπτει τα έτη 2023 έως και 2026. Συνεπώς, το κάθε μοντέλο χρησιμοποιήθηκε για να παράγει συνεχόμενες προβλέψεις, όπου η πρόβλεψη της κάθε εβδομάδας ή μήνα τροφοδοτούσε ως είσοδο τα χαρακτηριστικά της επόμενης, μαζί με τις υπόλοιπες σταθερές (frozen) τιμές των υπόλοιπων μεταβλητών.

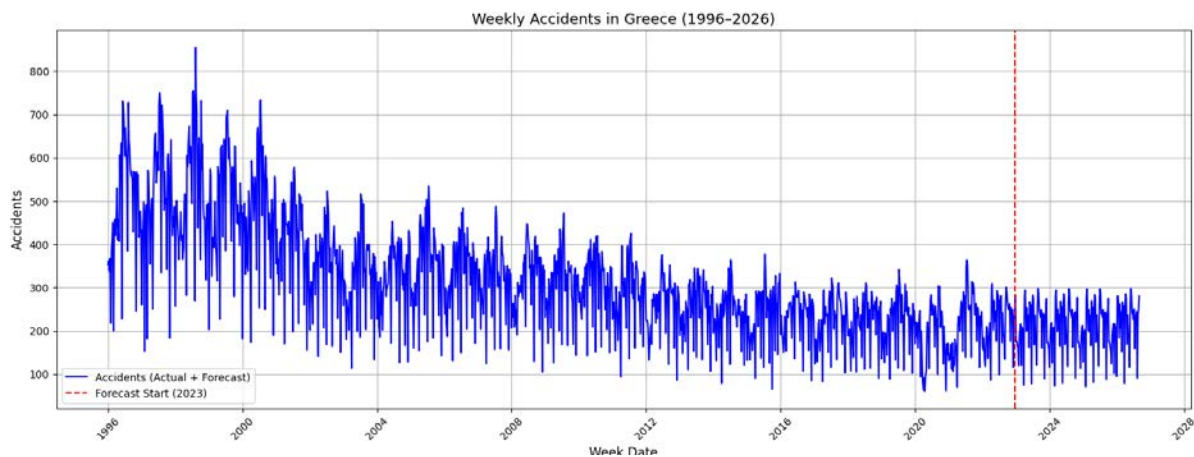
Σε γενικά πλαίσια, αξιοποιήθηκαν οι βέλτιστες υπερπαράμετροι που είχαν προκύψει μέσα από τη διαδικασία του grid search για κάθε επιμέρους μοντέλο. Η μοναδική διαφορά υπάρχει στη ρύθμιση των time steps, η οποία προσαρμόστηκε ώστε να καλύπτει ένα παράθυρο τριών ετών δεδομένων, δηλαδή 192 εβδομάδες για τα εβδομαδιαία datasets και 48 μήνες για τα μηνιαία. Αυτή η επιλογή έγινε για να διασφαλιστεί ότι κάθε πρόβλεψη βασίζεται σε ένα επαρκές ιστορικό παρατηρήσεων, το οποίο είναι κρίσιμο για τη διατήρηση της σταθερότητας και αξιοπιστίας των προβλέψεων σε βάθος χρόνου.

Συνολικά, η μεθοδολογία του recursive forecasting που εφαρμόστηκε, προσφέρει μια πιο ολοκληρωμένη και ρεαλιστική προσέγγιση για μελλοντική πρόβλεψη, ιδιαίτερα σε περιβάλλοντα όπου η διαθεσιμότητα πραγματικών δεδομένων για το μέλλον είναι περιορισμένη.

7.3.1 ΠΡΟΒΛΕΨΕΙΣ ΕΒΔΟΜΑΔΙΑΙΩΝ ΤΡΟΧΑΙΩΝ ΑΤΥΧΗΜΑΤΩΝ

Πίνακας 7.3.1-1: Ενδεικτικές Recursive Predictions (2023+) για το Dataset των Εβδομαδιαίων Τροχαίων Ατυχημάτων

<u>Date</u>	<u>Prediction Value</u>
1 st week of 2023	187
2 nd week of 2023	176
3 rd week of 2023	191
4 th week of 2023	177
5 th week of 2023	129
6 th week of 2023	215
7 th week of 2023	223
8 th week of 2023	240
9 th week of 2023	179
10 th week of 2023	215
11 th week of 2023	207
12 th week of 2023	221

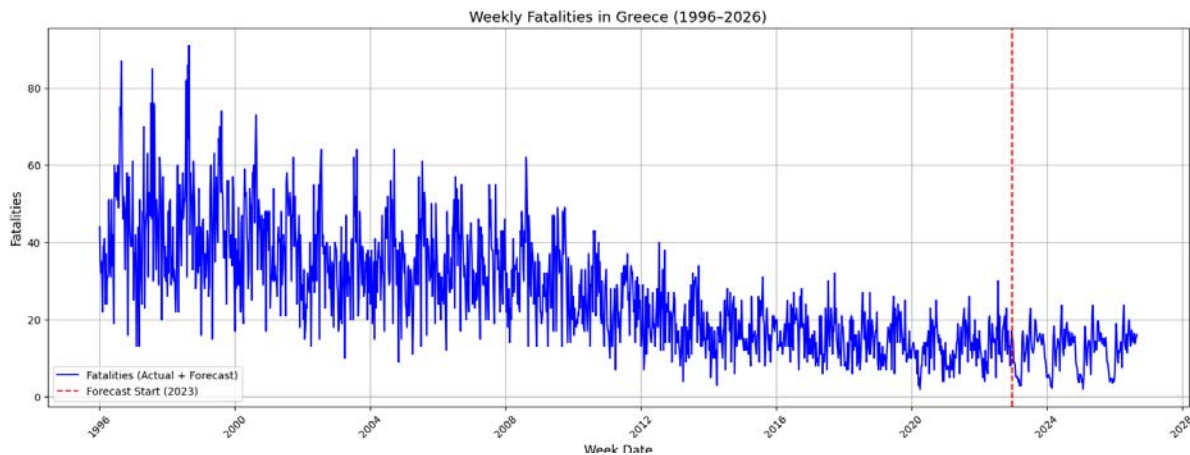


Σχήμα 7.3.1-1: Recursive Predictions Εβδομαδιαίων Τροχαίων Ατυχημάτων έως το 2026

7.3.2 ΠΡΟΒΛΕΨΕΙΣ ΕΒΔΟΜΑΔΙΑΙΩΝ ΘΑΝΑΤΩΝ ΑΠΟ ΤΡΟΧΑΙΑ ΑΤΥΧΗΜΑΤΑ

Πίνακας 7.3.2-1: Ενδεικτικές Recursive Predictions (2023+) για το Dataset των Εβδομαδιαίων Θανάτων από Τροχαία Ατυχήματα

<u>Date</u>	<u>Prediction Value</u>
1 st week of 2023	13
2 nd week of 2023	10
3 rd week of 2023	10
4 th week of 2023	10
5 th week of 2023	8
6 th week of 2023	9
7 th week of 2023	9
8 th week of 2023	9
9 th week of 2023	8
10 th week of 2023	11
11 th week of 2023	8
12 th week of 2023	8

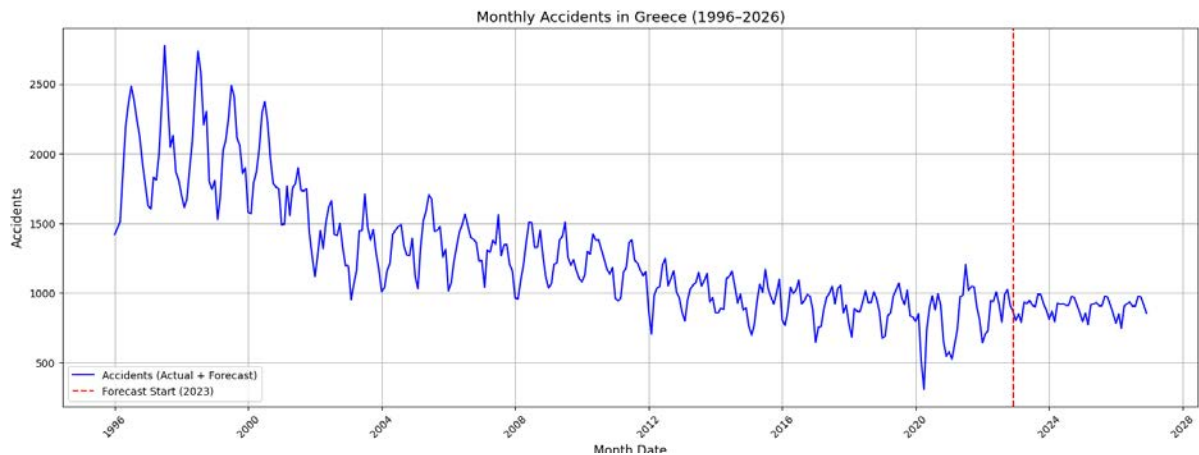


Σχήμα 7.3.2-1: Recursive Predictions Εβδομαδιαίων Θανάτων από Τροχαία Ατυχήματα έως το 2026

7.3.3 ΠΡΟΒΛΕΨΕΙΣ ΜΗΝΙΑΙΩΝ ΤΡΟΧΑΙΩΝ ΑΤΥΧΗΜΑΤΩΝ

Πίνακας 7.3.3-1: Ενδεικτικές Recursive Predictions (2023+) για το Dataset των Μηνιαίων Τροχαίων Ατυχημάτων

<u>Date</u>	<u>Prediction Value</u>
1 st month of 2023	802
2 nd month of 2023	871
3 rd month of 2023	777
4 th month of 2023	881
5 th month of 2023	907
6 th month of 2023	943
7 th month of 2023	904
8 th month of 2023	844
9 th month of 2023	951
10 th month of 2023	987
11 th month of 2023	907
12 th month of 2023	868

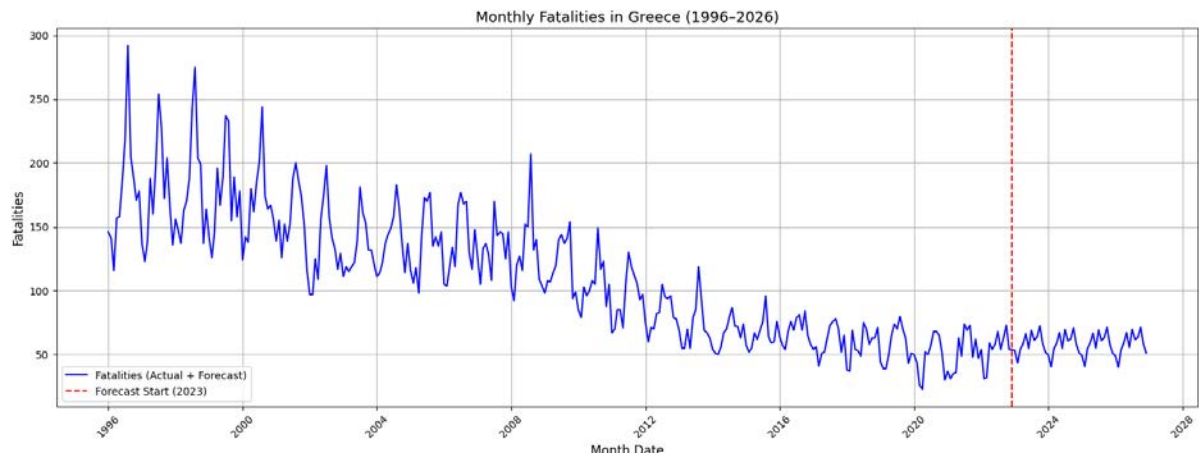


Σχήμα 7.3.3-1: Recursive Predictions Μηνιαίων Τροχαίων Ατυχημάτων έως το 2026

7.3.4 ΠΡΟΒΛΕΨΕΙΣ ΜΗΝΙΑΙΩΝ ΘΑΝΑΤΩΝ ΑΠΟ ΤΡΟΧΑΙΑ ΑΤΥΧΗΜΑΤΑ

Πίνακας 7.3.4-1: Ενδεικτικές Recursive Predictions (2023+) για το Dataset των Μηνιαίων Θανάτων από Τροχαία Ατυχήματα

<u>Date</u>	<u>Prediction Value</u>
1 st month of 2023	52
2 nd month of 2023	41
3 rd month of 2023	45
4 th month of 2023	58
5 th month of 2023	65
6 th month of 2023	55
7 th month of 2023	69
8 th month of 2023	60
9 th month of 2023	62
10 th month of 2023	71
11 th month of 2023	56
12 th month of 2023	52



Σχήμα 7.3.4-1: Recursive Predictions Μηνιαίων Θανάτων από Τροχαία Ατυχήματα έως το 2026

Κεφάλαιο 8 Σύγκριση Αποτελεσμάτων

Τα μοντέλα ARIMA και SARIMA, αν και θεωρούνται θεμέλια της στατιστικής μοντελοποίησης χρονοσειρών, παρουσίασαν σαφείς περιορισμούς στην ανάλυση της συγκεκριμένης χρονοσειράς τροχαίων ατυχημάτων και θανάτων. Παρά τις ενδείξεις ικανοποιητικής προσαρμογής ορισμένων μοντέλων στις log-μετασχηματισμένες εκδοχές, η πλειονότητα των υποδειγμάτων απέτυχε να ικανοποιήσει τα στατιστικά κριτήρια επάρκειας των residuals.

Ενδεικτικά, σε αρκετές περιπτώσεις το MAPE κυμάνθηκε σε πολύ υψηλά επίπεδα ενώ ακόμα και τα πιο υποσχόμενα μοντέλα, όπως το Log SARIMA(2,1,1)(0,1,1)[48], δεν κρίθηκαν πλήρως αποδεκτά λόγω κακής συμπεριφοράς των residuals. Η αδυναμία αυτή αναδεικνύει τα όρια των γραμμικών και εποχικών μοντέλων σε δεδομένα με μη γραμμικότητα, θόρυβο και σύνθετες εξωτερικές επιρροές, όπως αυτά των τροχαίων ατυχημάτων και θανάτων.

Αντιθέτως, το μοντέλο LSTM έδειξε σημαντική υπεροχή στις προβλέψεις εβδομαδιαίων και μηνιαίων θανάτων, κατορθώνοντας να ενσωματώσει την πληροφορία από το πλήρες σύνολο των διαθέσιμων χαρακτηριστικών. Με τη χρήση ενός απλού χρονικού παραθύρου (time steps = 1), 256 LSTM μονάδων και πυκνού στρώματος 64 νευρώνων, επιτεύχθηκαν εξαιρετικές μετρικές ακρίβειας, όπως MAPE = 6.43%, SMAPE = 5.87% και MASE = 0.3685, τιμές που υποδεικνύουν εντυπωσιακή απόδοση σε σχέση με τα ARIMA / SARIMA μοντέλα.

Η υπεροχή του LSTM μοντέλου πηγάζει από δύο θεμελιώδη πλεονεκτήματα. Την ικανότητά του να συλλαμβάνει μη γραμμικές αλληλεπιδράσεις και την ευελιξία του να ενσωματώνει πολυδιάστατη πληροφορία χωρίς την ανάγκη αυστηρής υπόθεσης στασιμότητας ή εποχικότητας. Ενώ τα ARIMA / SARIMA προϋποθέτουν έλεγχο στασιμότητας και αυστηρή επιλογή παραμέτρων p, d, q και P, D, Q, s , το LSTM εκπαιδεύεται απευθείας πάνω στο δομημένο dataset που προκύπτει από τη διαδικασία feature engineering, αξιοποιώντας μεταβλητές όπως χρονικά lags, κυλιόμενους μέσους όρους, εποχικά χαρακτηριστικά και dummy variables για πολιτικές παρεμβάσεις.

8.1 ARIMA vs SARIMA Μοντέλα

Η εφαρμογή των ARIMA και SARIMA μοντέλων στα εβδομαδιαία και μηνιαία datasets ανέδειξε ενδιαφέρουσες διαφορές στις προβλεπτικές τους επιδόσεις. Τα βασικά ARIMA μοντέλα, ακόμα και με χρήση log-transform, αδυνατούσαν να ενσωματώσουν πλήρως τις περιοδικές αυξομειώσεις των δεδομένων, με αποτέλεσμα να εμφανίζουν στατιστικά σημαντικά residuals σε αρκετά χρονικά σημεία. Συγκεκριμένα, στο μοντέλο ARIMA(4,1,1) για εβδομαδιαία ατυχήματα, παρατηρήθηκε διακύμανση στο διάστημα πρόβλεψης, με υποεκτιμήσεις σε περιόδους αυξημένης κυκλοφορίας, όπως οι καλοκαιρινοί μήνες και οι περίοδοι εορτών. Η αδυναμία αυτή δείχνει το ότι οι μη εποχικοί όροι του ARIMA δεν επαρκούν για να συλλάβουν τις εναλλαγές της συχνότητας των ατυχημάτων που εμφανίζουν περιοδικότητα σε ετήσια, μηνιαία ή και εβδομαδιαία βάση.

Αντιθέτως, τα SARIMA μοντέλα, μέσω της ενσωμάτωσης εποχικών παραμέτρων, προσέφεραν πιο αποδοτική προσαρμογή στη δομή των συγκεκριμένων δεδομένων. Το μοντέλο SARIMA(2,1,1)(0,1,1)[48], το οποίο εφαρμόστηκε στην εβδομαδιαία χρονοσειρά, κατάφερε να συλλάβει τη περιοδικότητα, εξομαλύνοντας τις προβλέψεις σε περιόδους με έντονη διακύμανση. Ωστόσο, ακόμη και με αυτή τη βελτίωση, τα διαγνωστικά των residuals δεν πληρούσαν σε όλες τις περιπτώσεις τα απαραίτητα κριτήρια λευκού θορύβου, γεγονός που υποδηλώνει ότι το SARIMA, αν και ανώτερο σε σχέση με το ARIMA, δεν επαρκεί από μόνο του για να αντιμετωπίσει την πολυπλοκότητα της συγκεκριμένης χρονοσειράς.

Η σύγκριση σε επίπεδο μετρικών ακρίβειας επιβεβαίωσε τις παραπάνω παρατηρήσεις. Σε όλες τις περιπτώσεις, οι τιμές των δεικτών RMSE, MAPE και MAE ήταν ελαφρώς βελτιωμένες στα SARIMA μοντέλα σε σχέση με τα αντίστοιχα ARIMA. Για παράδειγμα, στη μηνιαία πρόβλεψη τροχαίων ατυχημάτων, το μοντέλο SARIMA(2,1,1)(0,1,1)[12] υπερείχε του ARIMA(3,1,2) τόσο σε προβλεπτική σταθερότητα όσο και σε κατανομή των residuals.

Αν και τα SARIMA μοντέλα εμφανίζονται καταλληλότερα σε περιβάλλοντα με καθορισμένη εποχικότητα, όπως είναι οι χρονοσειρές των τροχαίων ατυχημάτων, η εφαρμογή τους πρέπει να γίνεται με προσοχή και έλεγχο των παραδοχών. Τα ARIMA μοντέλα, παρότι πιο απλά και πιο ευέλικτα σε μη εποχικές περιπτώσεις, υστερούν σημαντικά όταν το πρόβλημα εμφανίζει καθαρές κυκλικές τάσεις. Στη συγκεκριμένη περίπτωση, η παρουσία ισχυρής εποχικότητας στην εβδομαδιαία και μηνιαία συχνότητα των τροχαίων ατυχημάτων και θανάτων, καθιστά τα SARIMA μοντέλα ανώτερα σε θεωρητικό επίπεδο.

8.2 ARIMA / SARIMA Μοντέλα vs LSTM NN

Η απόδοση των LSTM στην συγκεκριμένη περίπτωση, υπερείχε ξεκάθαρα έναντι των ARIMA/SARIMA σε όλα σχεδόν τα σενάρια πρόβλεψης (εβδομαδιαία και μηνιαία δεδομένα, ατυχήματα και θάνατοι). Ενδεικτικά, στο εβδομαδιαίο dataset των θανάτων από τροχαία ατυχήματα, το LSTM πέτυχε $MAPE = 6.43\%$, $MAE = 1.34$ και $WAPE = 5.17\%$, σε αντίθεση με τα αντίστοιχα SARIMA μοντέλα, των οποίων οι τιμές MAPE ξεπερνούσαν συχνά το 20%, ενώ το MASE πλησίαζε ή υπερέβαινε το 1.0. Οι διαφορές αυτές αντανακλούν την ουσιαστική ικανότητα του LSTM να μάθει την αλληλεπίδραση μεταξύ πολλών input και target values, χωρίς αυστηρές προϋποθέσεις στασιμότητας ή εποχικότητας.

Επίσης, σε αντίθεση με τα ARIMA/SARIMA, τα LSTM μοντέλα επωφελούνται από το δυναμικό πολυμεταβλητής επεξεργασίας και από την ενσωμάτωση των feature engineered μεταβλητών. Τα LSTM μοντέλα αξιοποίησαν όλο το σύνολο των διαθέσιμων inputs, τα οποία ενισχύουν την πληροφορία της χρονοσειράς και επιτρέπουν στο LSTM να κατασκευάσει μια πιο σύνθετη πρόβλεψη. Έτσι, το LSTM κρίνεται περισσότερο κατάλληλο για προβλήματα όπου η εξεταζόμενη περίπτωση αλληλεπιδρά με εξωγενείς μεταβλητές και παρουσιάζει μη γραμμικά μοτίβα, όπως συμβαίνει στις πραγματικές καταστάσεις πρόβλεψης τροχαίων ατυχημάτων και θανάτων.

Εκτός αυτού, το LSTM δείχνει ευελιξία τόσο σε επίπεδο αρχιτεκτονικής όσο και σε επίπεδο προβλεπτικής μεθοδολογίας. Με τη δυνατότητα εκπαίδευσης σε one-step ή multi-step πλαίσια, το μοντέλο μπορεί να προσαρμοστεί σε διαφορετικούς χρονικούς ορίζοντες πρόβλεψης. Στην συγκεκριμένη περίπτωση, αν και εφαρμόστηκε κυρίως σε one-step forecasts, η μέθοδος υποστηρίζει και recursive ή direct multi-output προσεγγίσεις, προσφέροντας προοπτικές μελλοντικής επεκτασιμότητας. Αντίθετα, τα ARIMA/SARIMA περιορίζονται σε σταθερού μήκους βήματα και απαιτούν ξεχωριστή προσαρμογή για κάθε επίπεδο πρόβλεψης.

Σε γενικά πλαίσια, η ανάλυση δείχνει ότι τα LSTM υπερέχουν σε ακρίβεια, ευελιξία και ικανότητα ενσωμάτωσης πληροφορίας, καθιστώντας τα ιδιαίτερα κατάλληλα για προβλήματα όπως η πρόβλεψη τροχαίων ατυχημάτων, όπου επικρατεί έντονη μη γραμμικότητα. Παρ' όλα αυτά, τα ARIMA και SARIMA εξακολουθούν να έχουν αξία σε περιβάλλοντα όπου η διαφάνεια και η στατιστική ορθότητα αποτελούν προτεραιότητα.

8.3 Ακρίβεια Προβλέψεων στα Εβδομαδιαία vs Μηνιαία Dataset

Εκτός των παραπάνω, η ανάλυση των αποτελεσμάτων έδειξε επίσης, ότι τα εβδομαδιαία datasets προσφέρουν ανώτερη ακρίβεια πρόβλεψης σε περιβάλλοντα όπου εφαρμόζονται ευέλικτα μοντέλα μηχανικής μάθησης, όπως τα LSTM. Πιο συγκεκριμένα, οι προβλέψεις εβδομαδιαίων θανάτων με χρήση LSTM απέδωσαν χαμηλότερες τιμές σφαλμάτων σε όλους τους βασικούς δείκτες. Αντίθετα, η μικρότερη πυκνότητα του μηνιαίου dataset περιορίσε τη δυνατότητα του LSTM να μάθει τις σύνθετες εξαρτήσεις μεταξύ εισόδων και στόχου, οδηγώντας σε πιο γενικευμένες και λιγότερο ακριβείς προβλέψεις.

Από την άλλη πλευρά, στα παραδοσιακά στατιστικά μοντέλα ARIMA και SARIMA, η μηνιαία χρονοσειρά φαίνεται να λειτουργεί ευνοϊκά, καθώς περιορίζει την επίδραση του θορύβου και βελτιώνει τα residuals, ιδιαίτερα σε μοντέλα όπου η εποχικότητα είναι εμφανής (π.χ. SARIMA(2,1,1)(0,1,1)[12]). Παρά τις χαμηλότερες επιδόσεις τους έναντι των LSTM, τα στατιστικά μοντέλα εμφάνισαν πιο σταθερή συμπεριφορά με το μηνιαίο dataset, με μικρότερη ευαισθησία στις ακραίες τιμές και στα βραχυχρόνια συμβάντα.

Αυτό σημαίνει ότι η επιλογή μεταξύ εβδομαδιαίου και μηνιαίου βάθους πρόβλεψης δεν είναι απόλυτη, αλλά εξαρτάται από το μοντέλο και τον στόχο της ανάλυσης. Εάν ο στόχος είναι η αξιόπιστη πρόβλεψη και η άμεση υποστήριξη λήψης αποφάσεων τα εβδομαδιαία δεδομένα με χρήση LSTM προσφέρουν ανώτερη πληροφόρηση και ακρίβεια. Αντίθετα, για στρατηγικές μακροπρόθεσμου σχεδιασμού, οι μηνιαίες χρονοσειρές σε συνδυασμό με πιο σταθερά στατιστικά μοντέλα αποτελούν μια καλύτερη επιλογή.

Κεφάλαιο 9 Συμπεράσματα και Προτάσεις για Μελλοντική Έρευνα

Η παρούσα εργασία επιχείρησε μια πολυδιάστατη ανάλυση της χρονοσειράς τροχαίων ατυχημάτων και θανάτων στην Ελλάδα, αξιοποιώντας ένα ευρύ φάσμα μεθόδων πρόβλεψης, που εκτείνονται από τα κλασικά στατιστικά υποδείγματα ARIMA και SARIMA έως σύγχρονες τεχνικές βαθιάς μάθησης όπως τα νευρωνικά δίκτυα τύπου LSTM. Μέσα από συστηματική ανάλυση δεδομένων, εκτεταμένη διαδικασία feature engineering, καθώς και συνεχή αξιολόγηση των μοντέλων με βάση πλήθος μετρικών (MAPE, RMSE, MAE, SMAPE, MASE, WAPE), επιχειρήθηκε να εντοπιστεί η πιο κατάλληλη μεθοδολογική προσέγγιση για την αξιόπιστη πρόβλεψη της εξέλιξης των τροχαίων περιστατικών.

Τα συμπεράσματα της μελέτης συγκλίνουν στην ανάδειξη της υπεροχής των LSTM μοντέλων έναντι των παραδοσιακών στατιστικών εναλλακτικών. Πιο συγκεκριμένα, σε προβλέψεις εβδομαδιαίων θανάτων, το LSTM πέτυχε τιμές σφαλμάτων ιδιαίτερα χαμηλές επιβεβαιώνοντας την ικανότητά του να συλλαμβάνει σύνθετες και μη γραμμικές εξαρτήσεις που δεν είναι δυνατόν να μοντελοποιηθούν αποτελεσματικά από γραμμικά ή εποχιακά στατιστικά μοντέλα. Αντιθέτως, τα μοντέλα ARIMA και SARIMA εμφάνισαν περιορισμούς τόσο στη δομική τους ευκαμψία όσο και στην ικανότητα πρόβλεψης σε δεδομένα με έντονη μεταβλητότητα και πολυπαραγοντική δυναμική. Η στατιστική διάγνωση των residuals αποκάλυψε παραβιάσεις βασικών παραδοχών, γεγονός που αμφισβήτησε την εγκυρότητα των στατιστικών προβλέψεων σε πολλά σενάρια.

Παράλληλα, η ανάλυση ανέδειξε τη σημασία της χρονικής κλίμακας στην ποιότητα των προβλέψεων. Η χρήση εβδομαδιαίων δεδομένων επέτρεψε μεγαλύτερη ευκρίνεια στις βραχυχρόνιες τάσεις και στις επιδράσεις εποχικών ή πολιτικών μεταβλητών, προσφέροντας καλύτερες επιδόσεις κυρίως στα LSTM μοντέλα. Αντιθέτως, τα μηνιαία δεδομένα προσέφεραν μεγαλύτερη σταθερότητα, διευκολύνοντας την εφαρμογή απλούστερων υποδειγμάτων αλλά σε βάρος της ακρίβειας σε βραχυχρόνιο επίπεδο. Το εύρημα αυτό καθιστά σαφές ότι η επιλογή χρονικής κλίμακας δεν είναι ουδέτερη, αλλά αλληλεπιδρά άμεσα με τη φύση του μοντέλου και τον σκοπό της πρόβλεψης.

Σε πρακτικό επίπεδο, τα ευρήματα της παρούσας μελέτης μπορούν να αξιοποιηθούν ως εργαλείο υποστήριξης λήψης αποφάσεων από κρατικούς και τοπικούς φορείς οδικής ασφάλειας. Η δυνατότητα του LSTM να ενσωματώνει πληροφορία από κοινωνικοοικονομικούς δείκτες, πολιτικές παρεμβάσεις και κυκλικά μοτίβα εποχικότητας, προσφέρει σημαντικά πλεονεκτήματα στην πρόβλεψη περιόδων αυξημένου κινδύνου. Αυτό μπορεί να επιτρέψει στοχευμένες παρεμβάσεις, όπως ενίσχυση της αστυνόμευσης ή εφαρμογή έκτακτων μέτρων πρόληψης, ιδίως σε χρονικά διαστήματα με σταθερά υψηλή προβλεπόμενη επικινδυνότητα.

9.1 Περιορισμοί της Μελέτης

Παρά τη συνολική επιτυχία της μελέτης στην ανάπτυξη ενός ισχυρού πλαισίου πρόβλεψης για τροχαία ατυχήματα και θανάτους, οφείλουν να αναγνωριστούν ορισμένοι περιορισμοί. Ένας από τους σημαντικότερους, αφορά την αδυναμία πραγματοποίησης έγκυρου multi-step ή recursive forecasting με το μοντέλο LSTM πέραν του τελευταίου χρονικού σημείου του dataset (2022), εξαιτίας της έλλειψης διαθέσιμων ή προβλεπόμενων τιμών για τις κοινωνικοοικονομικές και πολιτικές μεταβλητές που χρησιμοποιούνται ως είσοδοι. Δεδομένου ότι το μοντέλο εξαρτάται από πολυδιάστατη πληροφορία, η μη διαθεσιμότητα μελλοντικών τιμών των μεταβλητών αυτών περιορίζει τη δυνατότητα του μοντέλου να επεκταθεί σε έγκυρες προβλέψεις για τα έτη 2023 και μετά, χωρίς την ενσωμάτωση σεναρίων ή άλλων τεχνικών πρόβλεψης.

Ένας επιπλέον περιορισμός αφορά την ευαισθησία των LSTM μοντέλων στην ποιότητα και στατιστική σταθερότητα των δεδομένων εισόδου. Παρά τις προηγμένες τεχνικές κανονικοποίησης, ορισμένες μεταβλητές παρουσίαζαν σημαντική ετεροσκεδαστικότητα ή ασυμμετρία, που ενδέχεται να επηρέασαν την προσαρμογή και τη γενίκευση του μοντέλου. Επιπροσθέτως, το μικρό πλήθος παρατηρήσεων στο μηνιαίο dataset (μόλις 324 μήνες) περιορίσε την αποτελεσματικότητα της εκπαίδευσης των LSTM υποδειγμάτων, τα οποία, ως μοντέλα βαθιάς μάθησης, απαιτούν σημαντικό όγκο δεδομένων για να εκπαιδευτούν επαρκώς.

Εκτός των παραπάνω, ένας ουσιαστικός περιορισμός της μελέτης σχετίζεται με τις αδυναμίες των μοντέλων ARIMA και SARIMA στο πλαίσιο της πρόβλεψης τροχαίων ατυχημάτων και θανάτων. Τα συγκεκριμένα μοντέλα, παρότι ισχυρά σε χρονοσειρές με σταθερή στατιστική συμπεριφορά, εμφανίζουν περιορισμένη προσαρμοστικότητα σε δεδομένα

που χαρακτηρίζονται από έντονη μη γραμμικότητα, εξωτερικές παρεμβάσεις και εποχιακές ασυμμετρίες. Επιπλέον, το γεγονός ότι τα ARIMA/SARIMA λειτουργούν αποκλειστικά ως univariate μοντέλα συνεπάγεται πως αδυνατούν να ενσωματώσουν επιπλέον πηγές πληροφορίας, που ενδέχεται να εξηγούν σημαντικό μέρος της διακύμανσης της χρονοσειράς. Ως αποτέλεσμα, τα μοντέλα αυτά βασίζονται εξ ολοκλήρου στην εσωτερική δομή του ιστορικού σήματος και παραμένουν ευάλωτα σε εξωγενείς μεταβολές που δεν είχαν εκδηλωθεί κατά την περίοδο εκπαίδευσης.

Επίσης, αξίζει να σημειωθεί ότι εφόσον υπήρχαν διαθέσιμες ή αξιόπιστα προβλεπόμενες μελλοντικές τιμές των κοινωνικοοικονομικών και πολιτικών μεταβλητών, θα υπήρχε η δυνατότητα εφαρμογής πιο εξελιγμένων στατιστικών υποδειγμάτων, όπως το SARIMAX (Seasonal ARIMA with exogenous variables). Το υπόδειγμα αυτό επεκτείνει το κλασικό SARIMA, επιτρέποντας την ενσωμάτωση εξωγενών μεταβλητών ως inputs στη διαδικασία πρόβλεψης. Μέσω αυτής της προσέγγισης, σημαντικοί κοινωνικοοικονομικοί δείκτες θα μπορούσαν να λειτουργήσουν ως προβλεπτικοί παράγοντες και να ενισχύσουν την ακρίβεια των στατιστικών μοντέλων σε περιόδους με έντονες εξωτερικές επιδράσεις.

9.2 Προτάσεις για Μελλοντική Έρευνα

Μία από τις βασικότερες προτάσεις για μελλοντική έρευνα αφορά την επέκταση του συνόλου των ανεξάρτητων μεταβλητών που χρησιμοποιούνται για την πρόβλεψη των τροχαίων ατυχημάτων και θανάτων. Ειδικότερα, η ενσωμάτωση μετεωρολογικών δεδομένων (όπως βροχόπτωση, θερμοκρασία, παγετός, ομίχλη, ηλιοφάνεια) θα μπορούσε να ενισχύσει σημαντικά την ακρίβεια των μοντέλων, καθώς έχει αποδειχθεί ότι οι καιρικές συνθήκες επηρεάζουν άμεσα τη συμπεριφορά των οδηγών και τη συχνότητα των ατυχημάτων. Παράλληλα, μεταβλητές όπως η ώρα του περιστατικού, η κατηγορία της ημέρας (εργάσιμη, αργία, Σαββατοκύριακο), αλλά και η ημερήσια χρονική ανάλυση των δεδομένων θα επέτρεπαν μια πολύ πιο λεπτομερή προσέγγιση, οδηγώντας σε μοντέλα πρόβλεψης με μεγαλύτερη ευαισθησία στις χρονικές διακυμάνσεις του κινδύνου. Η μετάβαση από εβδομαδιαία ή μηνιαία χρονοσειρά σε ημερήσια ανάλυση θα αύξανε τον όγκο των δεδομένων και θα επέτρεπε τη χρήση τεχνικών εντοπισμού ανωμαλιών (anomaly detection) και real-time prediction.

Επιπλέον, ένας κρίσιμος άξονας για την επέκταση της έρευνας είναι η ανάπτυξη μοντέλων πρόβλεψης και για τις κοινωνικοοικονομικές και πολιτικές μεταβλητές, που σήμερα

χρησιμοποιούνται μόνο ως ιστορικά inputs. Εφόσον καταστεί δυνατή η αξιόπιστη πρόβλεψη αυτών των μεταβλητών, τότε θα μπορέσουν να υλοποιηθούν πλήρως multi-step και recursive προβλέψεις με χρήση LSTM ή SARIMAX. Η ενσωμάτωση τέτοιων μοντέλων (π.χ. VAR, Prophet, ακόμα και LSTM για κάθε μεμονωμένο input) θα επιτρέψει τον υπολογισμό μελλοντικών inputs, ενισχύοντας την ικανότητα πρόβλεψης σε ορίζοντες μεγαλύτερους των 1-2 εβδομάδων ή μηνών.

Επιπλέον, μελλοντική έρευνα μπορεί να στραφεί στη διερεύνηση εναλλακτικών μοντέλων μηχανικής μάθησης, όπως τα Support Vector Regression (SVR), Random Forest Regressors, Gradient Boosting Machines (GBM) ή XGBoost, που έχουν αποδείξει εξαιρετική απόδοση σε προβλήματα παρόμοιας φύσης. Τα μοντέλα αυτά προσφέρουν μεγαλύτερη ερμηνευσιμότητα σε σύγκριση με τα LSTM και είναι λιγότερο ευαίσθητα σε ζητήματα μικρών dataset ή overfitting. Η συγκριτική αξιολόγηση των παραπάνω αλγορίθμων με το υπάρχον LSTM υπόδειγμα μπορεί να αποκαλύψει σενάρια στα οποία μοντέλα με μικρότερη πολυπλοκότητα πετυχαίνουν ισοδύναμα ή και καλύτερα αποτελέσματα, ιδίως σε συνθήκες περιορισμένων δεδομένων ή υψηλού θορύβου.

Ιδιαίτερο ενδιαφέρον παρουσιάζει και η προοπτική υλοποίησης υβριδικών μοντέλων, τα οποία συνδυάζουν τη στατιστική ακρίβεια των ARIMA/SARIMA με τη μη γραμμική δύναμη των μοντέλων ML. Για παράδειγμα, μπορεί να υλοποιηθεί ένας υβριδικός αλγόριθμος όπου το στατιστικό υπόδειγμα προβλέπει τη γραμμική και εποχική συνιστώσα της χρονοσειράς, ενώ το LSTM ή ένα άλλο ML μοντέλο εκπαιδεύεται για να αποδώσει τα μη γραμμικά στοιχεία.

Τέλος, στο πλαίσιο της πρακτικής αξιοποίησης των προβλέψεων, προτείνεται η ανάπτυξη διαδραστικών εργαλείων οπτικοποίησης (dashboard-based interfaces) που θα επιτρέπουν στους φορείς χάραξης πολιτικής να παρακολουθούν σε πραγματικό χρόνο τις τάσεις, να αξιολογούν την πρόβλεψη επικινδυνότητας ανά εβδομάδα ή μήνα και να λαμβάνουν αποφάσεις βασισμένες σε δεδομένα. Η ενσωμάτωση ενός συστήματος προειδοποίησης (early warning) βασισμένο στις έγκυρες προβλέψεις των LSTM ή SARIMAX μπορεί να αποτελέσει πολύτιμο εργαλείο για την πρόληψη ατυχημάτων και την ορθολογική κατανομή πόρων.

Βιβλιογραφία

- Angadi, V. S., & Halyal, S. (2024). Forecasting Road Accidents Using Deep Learning Approach: Policies to Improve Road Safety. *Journal of Soft Computing in Civil Engineering*, 8(4), 27 - 53.
- Azhari, A. H., Zaidi, F. A., Anuar, M. H., & Othman, J. (2022). A Statistical Analysis of Road Accident Fatalities in Malaysia. *International Journal of Academic Research in Business and Social Sciences*, 12(5), 1224 - 1230.
- Balawi, M., & Tenekeci, G. (2024). Time series traffic collision analysis of London hotspots: Patterns, predictions and prevention strategies. *Heliyon*, 10.
- Dalal, K., Lin, Z., Gifford, M., & Svanström, L. (2013). Economics of Global Burden of Road Traffic Injuries and Their Relationship with. *International Journal of Preventive Medicine*, 1442-1450.
- Deretic, N., Stanimirovic, D., Awadh, M. A., Vujanovic, N., & Djukic, A. (2022). SARIMA Modelling Approach for Forecasting of Traffic Accidents. *Sustainability*, 14(4403).
- European Commission. (2023). *Country Profile Greece*. Road Safety Observatory. Brussels: European Commission, Directorate General for Transport.
- Folla, K. (2023, December 14). The Greek National Road Safety Plan 2021-2030. *Workshop: RSPR MD policy dialogue and how to move forward road safety system*. Chisinau: National Technical University of Athens, Hellenic Ministry of Infrastructure and Transport.
- Getahun, K. A. (2021). Time series modeling of road traffic accidents in Amhara Region. *Journal of Big Data*, 8(102).
- Mpogas, K., Kopelias, P., Mitropoulos, L., & Kepaptsoglou, K. (2016). Road Safety in urban areas in Greece during economy downturn. A before – after comparison. *3rd Conference on Sustainable Urban Mobility* (pp. 228 – 234). Volos: Elsevier.

- Parvareh, M., Karimi, A., Rezaei, S., Woldemichael, A., Nili, S., Nouri, B., & Nasab, N. E. (2018). Assessment and prediction of road accident injuries trend using time-series models in Kurdistan. *Burns & Trauma*, 6(9). doi:<https://doi.org/10.1186/s41038-018-0111-6>
- Peden, M., Scurfield, R., Sleet, D., Mohan, D., Hyder, A. A., Jarawan, E., & Mathers, C. (2004). *World report on road traffic injury prevention*. Geneva: World Health Organization.
- Qian, Y., Zhang, X., Fei, G., Sun, Q., Li, X., Stallones, L., & Xiang, H. (2020). Forecasting deaths of road traffic injuries in China using an artificial neural network. *Traffic Injury Prevention*, 21(6), 407 - 412. doi:10.1080/15389588.2020.1770238
- Road Safety Observatory. (2025, March 18). Basic road safety figures, Greece 2014 - 2024. Αθήνα, Αττική, Ελλάδα: National Technical University of Athens.
- Slight increase in road fatalities in 2024, Greece*. (2025, March 19). Retrieved from nrso.ntua: <https://www.nrso.ntua.gr/slight-increase-in-road-fatalities-in-2024-greece/>
- World Health Organization. (2023). *Global status report on road safety 2023*. Geneva: World Health Organization.
- Yannis, G., Papadimitriou, E., & Folla, K. (2014). Effect of GDP changes on road traffic fatalities. *Safety Science*, 42 - 49.
- Zhang, Z., Yang, W., & Wushour, S. (2020). Traffic Accident Prediction Based on LSTM-GBRT Model. *Journal of Control Science and Engineering*, 2020.
- ΣΕΣ Επιτροπή Οδικών Μεταφορών. (2025, January 27). Οι συγκοινωνιολόγοι συζητούν και προτείνουν: Υφιστάμενη κατάσταση και προοπτικές. Αθήνα, Αττική, Ελλάδα: Συγκοινωνιολόγοι Ελλάδος ΣΕΣ.

ΠΑΡΑΡΤΗΜΑ: Ενδεικτικά Σημεία Κώδικα

9.3 SQL Queries Εισαγωγής Δεδομένων στην ΒΔ

9.3.1 ΕΙΣΑΓΩΓΗ ΕΒΔΟΜΑΔΙΑΙΩΝ ΔΕΔΟΜΕΝΩΝ ΣΤΗΝ ΒΔ

```
UPDATE weekly_accidents
SET fatalities =
CASE
    WHEN week = 1 THEN ...
    WHEN week = 2 THEN ...
    WHEN week = 3 THEN ...
    WHEN week = 4 THEN ...
    WHEN week = 5 THEN ...
    WHEN week = 6 THEN ...
    ELSE fatalities
END
WHERE year = ...
AND month = ...
AND week BETWEEN 1 AND 6;
```

9.3.2 AGGREGATE ΜΗΝΙΑΙΩΝ ΔΕΔΟΜΕΝΩΝ ΣΤΗΝ ΒΔ

```
UPDATE monthly_accidents
SET fatalities = (
    SELECT SUM(COALESCE(wd.fatalities, 0))
    FROM weekly_accidents wd
    WHERE wd.year = monthly_accidents.year
```

```
AND wd.month = monthly_accidents.month
)
WHERE year = ... ;
```

9.3.3 AGGREGATE ΕΤΗΣΙΩΝ ΔΕΔΟΜΕΝΩΝ ΣΤΗΝ ΒΔ

```
UPDATE yearly_accidents
SET fatalities = (
    SELECT SUM(COALESCE(md.fatalities, 0))
    FROM monthly_accidents md
    WHERE md.year = yearly_accidents.year
)
WHERE year = ... ;
```

9.4 SQL Queries Εξαγωγής Δεδομένων στην ΒΔ

9.4.1 ΕΞΑΓΩΓΗ ΕΒΔΟΜΑΔΙΑΙΟΥ DATASET ΤΡΟΧΑΙΩΝ ΑΤΥΧΗΜΑΤΩΝ

```
SELECT *
FROM weekly_accidents
ORDER BY year, month, week;
```

9.4.2 ΕΞΑΓΩΓΗ ΕΒΔΟΜΑΔΙΑΙΟΥ DATASET ΤΡΟΧΑΙΩΝ ΑΤΥΧΗΜΑΤΩΝ ΜΕ ΚΟΙΝΩΝΙΚΟΟΙΚΟΝΟΜΙΚΟΥΣ ΠΑΡΑΓΟΝΤΕΣ

```
SELECT wi.*,
    yi.total_gdp, yi.gdp_per_capita, yi.unemployment, yi.population,
    yi.environmental_taxes, yi.vehicles, yi.petrol_cons, yi.ng_cons,
    yi.el_cons, yi.fatal_per_million_veh, yi.speed_infr, yi.drink_infr,
    yi.belt_infr, yi.helmet_infr, yi.inflation_rate, yi.gini_index
FROM weekly_accidents wi
JOIN yearly_indicators yi ON wi.year = yi.year
```

ORDER BY wi.year, wi.month, wi.week;

9.5 Ενδεικτικά Σημεία Κώδικα στην R

9.5.1 ΕΝΔΕΙΚΤΙΚΟΣ ΚΩΔΙΚΑΣ ΕΙΣΑΓΩΓΗΣ DUMMY VARIABLES ΓΙΑ ΝΟΜΟΘΕΣΙΕΣ

```
acc.w.ind.y_mod <- acc.w.ind.y_mod %>%
  mutate(
    speed_limits = ifelse(year >= 1999, 1, 0),
    seat_belt = ifelse(year >= 1999, 1, 0),
    helmet = ifelse(year >= 1999, 1, 0),
    alcohol = ifelse(year >= 1999, 1, 0),
    phone = ifelse(year >= 1999, 1, 0),
    licensing = ifelse(year >= 1999, 1, 0),
    airbags = ifelse(year >= 2009, 1, 0),
    abs = ifelse(year >= 2009, 1, 0),
    kid_seat = ifelse(year >= 1999, 1, 0),
    fines = ifelse(year >= 2007, 1, 0),
    pedestrians = ifelse(year >= 2011, 1, 0),
    kteo = ifelse(year >= 1999, 1, 0),
    point_system = ifelse(year >= 2007, 1, 0)
  )
```

9.5.2 ΣΥΝΑΡΤΗΣΗ ΥΠΟΛΟΓΙΣΜΟΥ DESCRIPTIVE STATISTICS

```
compute_statistics <- function(df) {
  numeric_df <- df %>% select(where(is.numeric))
  stats <- data.frame(
    Variable = colnames(numeric_df),
    Std_Dev = apply(numeric_df, 2, sd, na.rm = TRUE),
    Count = colSums(!is.na(numeric_df)),
    Length = nrow(numeric_df),
```

```

Max = apply(numeric_df, 2, max, na.rm = TRUE),
Min = apply(numeric_df, 2, min, na.rm = TRUE),
Mode = apply(numeric_df, 2, function(x) {
  ux <- unique(na.omit(x))
  ux[which.max(tabulate(match(x, ux)))]
}),
Median = apply(numeric_df, 2, median, na.rm = TRUE),
Range = apply(numeric_df, 2, function(x) max(x, na.rm = TRUE) - min(x, na.rm = TRUE)),
Variance = apply(numeric_df, 2, var, na.rm = TRUE),
IQR = apply(numeric_df, 2, IQR, na.rm = TRUE),
CV = apply(numeric_df, 2, function(x) sd(x, na.rm = TRUE) / mean(x, na.rm = TRUE)),
Q1 = apply(numeric_df, 2, function(x) quantile(x, 0.25, na.rm = TRUE)),
Q2 = apply(numeric_df, 2, function(x) quantile(x, 0.50, na.rm = TRUE)),
Q3 = apply(numeric_df, 2, function(x) quantile(x, 0.75, na.rm = TRUE)),
Skewness = apply(numeric_df, 2, function(x) skewness(x, na.rm = TRUE)),
Kurtosis = apply(numeric_df, 2, function(x) kurtosis(x, na.rm = TRUE)),
p10 = apply(numeric_df, 2, function(x) quantile(x, 0.10, na.rm = TRUE)),
p25 = apply(numeric_df, 2, function(x) quantile(x, 0.25, na.rm = TRUE)),
p50 = apply(numeric_df, 2, function(x) quantile(x, 0.50, na.rm = TRUE)),
p75 = apply(numeric_df, 2, function(x) quantile(x, 0.75, na.rm = TRUE)),
p90 = apply(numeric_df, 2, function(x) quantile(x, 0.90, na.rm = TRUE))
)
return(stats)
}

```

9.5.3 ΕΝΔΕΙΚΤΙΚΟ BLOCK ΚΩΔΙΚΑ ΓΙΑ MANUAL ARIMA / SARIMA MODEL SELECTION

```

train_ts_arima_acc_w <- window(acc_w_ts, end = c(2017, 29))
test_ts_arima_acc_w <- window(acc_w_ts, start = c(2017, 30))
model_arima_acc_w <- Arima(train_ts_arima_acc_w, order = c(3,1,2))
fc_arima_acc_w <- forecast(model_arima_acc_w, h = length(test_ts_arima_acc_w))
rmse_val_arima_acc_w <- rmse(test_ts_arima_acc_w, fc_arima_acc_w$mean)

```

```

mae_val_arma_acc_w <- mae(test_ts_arma_acc_w, fc_arma_acc_w$mean)
mape_val_arma_acc_w <- mape(test_ts_arma_acc_w, fc_arma_acc_w$mean) * 100
summary(model_arma_acc_w)
cat("AIC:", AIC(model_arma_acc_w), "\n")
cat("BIC:", BIC(model_arma_acc_w), "\n")
cat("RMSE:", rmse_val_arma_acc_w, "\n")
cat("MAE:", mae_val_arma_acc_w, "\n")
cat("MAPE:", mape_val_arma_acc_w, "%\n")
AIC(model_arma_acc_w)
BIC(model_arma_acc_w)

```

9.5.4 ΕΝΔΕΙΚΤΙΚΟ BLOCK ΚΩΔΙΚΑ ΓΙΑ ΕΦΑΡΜΟΓΗ ΤΗΣ AUTO.ARIMA() ΣΥΝΑΡΤΗΣΗΣ

```

model_auto_arma_acc_w <- auto.arima(
  train_ts_arma_acc_w,
  seasonal = FALSE,
  stepwise = FALSE,
  approximation = FALSE,
  trace = TRUE
)
fc_auto_arma_acc_w <- forecast(model_auto_arma_acc_w, h = length(test_ts_arma_acc_w))
summary(fc_auto_arma_acc_w)
summary(model_auto_arma_acc_w)
rmse_val_auto_arma_acc_w <- rmse(test_ts_arma_acc_w, fc_auto_arma_acc_w$mean)
mae_val_auto_arma_acc_w <- mae(test_ts_arma_acc_w, fc_auto_arma_acc_w$mean)
mape_val_auto_arma_acc_w <- mape(test_ts_arma_acc_w, fc_auto_arma_acc_w$mean) *
100
cat("AIC:", AIC(model_auto_arma_acc_w), "\n")
cat("BIC:", BIC(model_auto_arma_acc_w), "\n")
cat("RMSE:", rmse_val_auto_arma_acc_w, "\n")
cat("MAE:", mae_val_auto_arma_acc_w, "\n")
cat("MAPE:", mape_val_auto_arma_acc_w, "%\n")

```

9.5.5 ΕΝΔΕΙΚΤΙΚΟ BLOCK ΚΩΔΙΚΑ ΓΙΑ ΤΑ RESIDUAL CHECKS

```
Box.test(residuals(model_final_arma_manual_acc_w), lag = 10, type = 'Ljung-Box',fitdf =
length(model_final_arma_manual_acc_w$coef))
Box.test(residuals(model_final_arma_manual_acc_w), lag=12)
checkresiduals(model_final_arma_manual_acc_w)
Box.test(residuals(model_final_arma_auto_acc_w), lag = 10, type = 'Ljung-Box',fitdf =
length(model_final_arma_auto_acc_w$coef))
checkresiduals(model_final_arma_auto_acc_w)
Box.test(residuals(model_final_arma_auto_acc_w), lag=12)
```

9.5.6 ΕΝΔΕΙΚΤΙΚΟ BLOCK ΚΩΔΙΚΑ ΓΙΑ GRID SEARCH

```
acc_w_ts_gs <- ts(acc.w.ind.y_4wmod$accidents, start = c(1996, 1), frequency = 48)
train_ts_arma_acc_w_gs <- window(acc_w_ts_gs, end = c(2017, 29))
test_ts_arma_acc_w_gs <- window(acc_w_ts_gs, start = c(2017, 30))
p_gs_acc <- 0:3; d_gs_acc <- 0:3; q_gs_acc <- 0:3
results_gs_arma_acc <- list()
tried_models_gs_arma_acc <- 0
total_combinations_gs_arma_acc <- length(p_gs_acc) * length(d_gs_acc) * length(q_gs_acc)
cat(" Total combinations to try:", total_combinations_gs_arma_acc, "\n")
for (i in p_gs_acc) {
  for (j in d_gs_acc) {
    for (k in q_gs_acc) {
      tried_models_gs_arma_acc <- tried_models_gs_arma_acc + 1
      order <- c(i, j, k)
      cat("\nTrying          model",          tried_models_gs_arma_acc,          "of",
total_combinations_gs_arma_acc,
      "-> ARIMA(", paste(order, collapse = ","), ") \n")
      tryCatch({
        fit <- Arima(train_ts_arma_acc_w_gs,
```

```

        order = order,
        method = "ML")
h <- length(test_ts_arma_acc_w_gs)
fc <- forecast(fit, h = h)
mape <- accuracy(fc, test_ts_arma_acc_w_gs)["Test set", "MAPE"]
lb_test <- Box.test(residuals(fit), lag = 10, type = "Ljung-Box", fitdf = length(fit$coef))
bp_test <- Box.test(residuals(fit), lag = 12, type = "Box-Pierce", fitdf = length(fit$coef))
if (lb_test$p.value > 0.05 && bp_test$p.value > 0.05) {
  cat(" Model passed residual tests. Saving...\n")
  results_gs_arma_acc <- append(results_gs_arma_acc, list(list(
    order = paste0("(", paste(order, collapse = ","), ")"),
    AIC = AIC(fit),
    BIC = BIC(fit),
    MAPE = mape,
    LB_p = lb_test$p.value,
    BP_p = bp_test$p.value
  )))
} else {
  cat(" Residuals failed white noise tests.\n")
}
}, error = function(e) {
  message(" Error for ARIMA(", paste(order, collapse = ","), "): ", e$message)
})
}
}
}

```

9.6 Ενδεικτικά Σημεία Κώδικα στην Python

9.6.1 ΕΝΔΕΙΚΤΙΚΟ HYPERPARAMETER GRID ΚΑΙ ΑΡΧΙΤΕΚΤΟΝΙΚΗ LSTM NN

```
time_steps_list = [1]
```

```
lstm_units_list = [64, 128, 256]
dense_units_list = [64, 128, 256]
dropout_list = [0.2, 0.3]
loss_list = ['mse', 'huber']
learning_rates = [0.001, 0.005]
batch_size_list = [16, 32, 64]
epoch_list = [50, 100, 200]
results = []
```

```
class LSTMMModel(nn.Module):
    def __init__(self, input_dim, lstm_units, dense_units, dropout):
        super(LSTMMModel, self).__init__()
        self.lstm1 = nn.LSTM(input_dim, lstm_units, batch_first=True)
        self.dropout1 = nn.Dropout(dropout)
        self.lstm2 = nn.LSTM(lstm_units, lstm_units, batch_first=True)
        self.dropout2 = nn.Dropout(dropout)
        self.fc1 = nn.Linear(lstm_units, dense_units)
        self.relu1 = nn.ReLU()
        self.fc2 = nn.Linear(dense_units, dense_units // 2)
        self.relu2 = nn.ReLU()
        self.fc3 = nn.Linear(dense_units // 2, 1)
    def forward(self, x):
        x, _ = self.lstm1(x)
        x = self.dropout1(x)
        x, _ = self.lstm2(x)
        x = self.dropout2(x)
        x = x[:, -1, :] # Get last time step
        x = self.relu1(self.fc1(x))
        x = self.relu2(self.fc2(x))
        x = self.fc3(x)
        return x
```

9.6.2 GITHUB REPOSITORY

Όλο το διαθέσιμο υλικό που αναπτύχθηκε και χρησιμοποιήθηκε στο πλαίσιο της παρούσας διπλωματικής εργασίας είναι διαθέσιμο δημόσια μέσω του Repository στο GitHub: https://github.com/PavlosGkougkoulis/Road_Accident_Analysis_and_Prediction.