A Predictive Time Series Analysis of Southern-Californian Road Traffic Injuries from 2002-2010

Harini Lakshmanan and Anusia Edward

Shiley-Marcos School of Engineering, University of San Diego

Abstract

Transportation related injuries, in terms of traffic collisions of motorcyclists, pedestrians, and bicyclists, are the second leading cause of death in California (Kenneth, 2022). Studying the time series trends of Southern California's traffic-related injuries is important in order to help mitigate and address this pressing issue. The dataset "Road-Traffic-Injuries-2002-2010" was sourced as a raw csv from Healthdata.gov. The purpose of this analysis is to aid the California government with forecasting the number of injuries for Southern Californians in order to determine the quantity and necessity of resource allocation for Southern California to reduce traffic related accidents. The following models were explored in order to carry out this study: Linear Regression Model, Holt-Exponential Smoothing AAA, Holt-Exponential Smoothing ANA, ARIMA, and Neural Network. It was found that the Neural Network model outperformed the other models based on the following success criteria: RMSE, MAE, MPE, MAPE, and MASE. The Neural Network forecasted an increasing trend in traffic-related injuries for Southern California. In the future, additional demographic factors such as high blood alcohol content, seatbelt use, phone use, gender of perpetrator, and time of day can be used to further analyze traffic-related injuries.

Keywords: Southern California, traffic-related injuries, RMSE, MAE, MPE, MAPE, MASE, Linear Regression, Holt Exponential Smoothing, ARIMA, and Neural Network.

Table of Contents

Introduction

Background and Practical Implications

Literature Review

Purpose, Objectives, and Hypothesis of Present Study

Methods

Data Collection, Wrangling, Preprocessing, and Splitting

Data Description and Sample Characteristics

Exploratory Data Analysis

Modeling

Model 1: Linear Regression

Model 2: Holt-Winter's Exponential Smoothing AAA

Model 3: Holt-Winter's Exponential Smoothing ANA

Model 4: Autoregressive Integrated Moving Average (ARIMA)

Model 5: Neural Network

Results

Model Performance and Final Model Selection

Discussion

Introduction

Background and Practical Implications

The ability to move individuals, products, or materials from one city to another, or even one state to another, was revolutionized through the use of motorized vehicles. Essentially, travel made possible by motor vehicles supports economic and social development in many countries. Although this development allowed for a lot of growth, it is necessary to note the downside of motorized vehicles. Each year 1.35 million people are killed on roadways globally (*Road Traffic Injuries and Deaths—A Global Problem*, 2020). Approximately 3,700 people are killed globally in crashes involving cars, buses, motorcycles, bicycles, trucks, or pedestrians each day, with more than half of those killed being pedestrians or motorcyclists (*Road Traffic Injuries and Deaths—A Global Problem*, 2020). On a more local level, on average, there are 4,018 deaths per year in California caused by transportation accidents (Kenneth, 2022). Transportation related injuries, in terms of traffic collisions of motorcyclists, pedestrians, and bicyclists, are the second leading cause of death in California (Kenneth, 2022). In 2010, the Centers for Disease Control and Prevention (CDC) estimated that fatal and nonfatal traffic-related injuries will cost the world economy approximately $1.8 trillion U.S. dollars from 2015 to 2030, which is essentially equivalent to a yearly tax of 0.12% on the global gross domestic product (*Road Traffic Injuries and Deaths—A Global Problem*, 2020). Based on this statistic, studying the time series trends of Southern California's traffic-related injuries is important in order to help mitigate and address this pressing issue on a local level.

Literature Review

Current studies, in relation to the analysis of traffic-related injuries within the United States, pertain to the exploration of traffic-related injuries in conjunction with weather

conditions, ARIMA modeling, and Neural Network modeling. More specifically, in the study *Characterizing the Role of Wind and Dust in Traffic Accidents in California*, the negative effects of wind on ground transportation is discussed. The study uses data from the California Highway Patrol reports in order to analyze how weather conditions potentially contribute to traffic accidents within the Mojave Desert and the Imperial Valley from 2006 to 2016. These Californian counties were chosen based on their high levels of wind, dust, and particulate matter. This study concludes that the percentage of people who died in wind-related accidents is approximately double the mortalities than that of the mortalities of non wind-related accidents. The probability of wind-related accidents was associated with low-visibility. This indicates that dust contributed to these incidents. Furthermore, ground visibility and dust optical depth were negatively correlated, which highlights the increased risk of accidents in the future (Bhattachan et al., 2019).

In contrast, the study *A Time Series Model for Assessing the Trend and Forecasting the Road Traffic Accident Mortality*, a decreased risk of accidents in the future was forecasted for Zanjan Province, Iran (Yousefzadeh-Chabok et al., 2016). In order to assess and forecast the road traffic accident mortality in Zanjan Province, Iran, the researcher built a number of models including Auto Regressive (AR), Moving Average (MA), Auto Regressive Moving Average (ARMA), and Autoregressive Moving Integrated Moving Average (ARIMA). Through the use of autocorrelation function (ACF) plots, partial autocorrelation function (PACF) plots, least mean square error (MSE), and residuals, the ARIMA model was determined as the most proficient model for forecasting. The study, using the ARIMA model, was able to conclude and forecast that there was a decreasing trend for traffic accident-related mortality. The study proposed that the implementation of some interventions in the recent decade may have had a positive effect on

the decline in road traffic accident fatalities.

Furthermore, in the study, *Smoothing Strategies Combined with ARIMA and Neural Networks to Improve the Forecasting of Traffic Accidents*, the researchers delve closely into a time series analysis of traffic accidents within Valparaiso, Chile, using Autoregressive Neural Network (ANN) models and ARIMA models (Barba et al., 2014). The study works on improving these two models through the addition of either a 3-point MA smoothing or a singular value decomposition of the Hankel matrix. Additionally, two ANNs for one-step-ahead time series forecasting are used, where the coefficients of the first ANN are estimated through the particle swarm optimization learning algorithm. The coefficients of the second ANN are estimated with the resilient backpropagation learning algorithm. Within the study, the researchers concluded that the model that resulted in the best forecasts were given by the combination of a single value decomposition of the Hankel matrix with an ARIMA model.

Purpose, Objective, and Hypothesis of Present Study

The purpose and business objective of this analysis is to aid the California government with forecasting the number of injuries for Southern Californians in order to determine the quantity and necessity of resource allocation for Southern California to reduce traffic related accidents. In order to forecast the overall trend, based on the number of injuries, the machine learning objective of this predictive time series analysis is to create and tune five time series models that can accurately predict the prospective number of injuries for Southern California. The following five time series models were employed to carry out the objective of this study: Linear Regression, Holt-Winter's Exponential Smoothing AAA, Holt-Winter's Exponential Smoothing ANA, ARIMA, and Neural Network. The success criteria of this analysis is to create a model with the lowest RMSE, MAE, MPE, MAPE, and MASE scores. The general hypothesis

of this study is that at least one of the five time series modeling techniques utilized will result in a model that forecasts the overall trend of injuries within Southern California. More specifically, a secondary hypothesis is that the best model will be the ARIMA model with the success criteria being low scores for RMSE, MAE, MPE, MAPE, and MASE.

Methods

Data Collection, Wrangling, Preprocessing, and Splitting

The dataset "Road-Traffic-Injuries-2002-2010" was sourced as a raw csv from Healthdata.gov, which is an official website of the United States government dedicated to making health-related data more accessible. The data contains information from 2002 to 2010 about traffic-related injuries within different California based counties. The first step taken, in terms of preprocessing, was to perform dimensionality reduction. This was carried out because there were a total of thirty-one variables with 448,950 records. The columns were removed based on redundancy. For example, columns containing the variables *region_name* and *region_code* both included information regarding which region of California was being specified. Based on this information, the variable *region_code* was removed and the variable *region_name* was kept. This was also the case for the following variables: *reotype, geotypevalue, geoname, country_name, county_fips, LL95CI_poprate, UL95CL_poprate, poprate_se, decile_pop, LL95CI_avmtrate, avmtrate_se, avmtrate_rse,* and *CA_decile_avmtrate.* Additionally, columns were removed based on irrelevancy. For example, the column *version* contained the same value of the version of the data. Since it didn't contain any relevant information, it was removed. After this step the following columns remained: *reportyear, county_name, region_name, mode, severity, injuries, totalpop,* and *poprate.* In order to further focus the study, the data was subsetted into Southern Californian data using the *region_name* variable. Following this step, the

following variables remained: *reportyear, county_name, mode, severity, injuries, totalpop,* and *poprate.*

Next, the data's format was corrected in terms of the removal of random spacings and the removal of backslashes within the data. This step was carried out to ensure that the quality of the data analysis would not be compromised. The remaining data had a total of 267 missing values. The missing values for the *totalpop* and *poprate* columns were filled by using the U.S. government census information. Additional missing values were handled using the VIM package's *kNN* function to impute missing values. Outliers were evaluated using boxplots and then addressed using IQR (Appendix A). Additionally, skewness was evaluated using histograms and adjusted by normalizing the numeric data by applying a transformation (Appendix A). This transformation was carried out using the *BoxCoxTrans* function. Then, the categorical variables were dummy coded using the *dummy_cols* function. Subsequently, time plots were also created to analyze the data. A correlation plot of all of the variables was also produced to ensure that there wouldn't be any concern of multicollinearity (Appendix A). Finally, the data was split from 2002 to 2009 for the training set and 2010 for the validation set.

Data Description and Sample Characteristics

The dataset contains information on the annual number of fatal and severe road traffic injuries for California. More specifically, the dataset is composed of California's regions, counties, county divisions, cities, towns, and census tracts. The injury data is from the Statewide Integrated Traffic Records System (SWITRS), California Highway Patrol (CHP), and Transportation Injury Mapping System (TIMS). The data contains information from 2002 to 2010 along with the race/ethnicity of the person injured in a traffic-related injury. Additionally,

the data contains the number of injuries in the geographic area by severity (killed, severe injury) and mode of transportation of the victim (bicyclist, bus, car, motorcycle, pedestrian, truck, vehicles). The total population, rate of injuries over total population, and rate of injuries over annual miles traveled were also included within the dataset. The total number of variables for this dataset is thirty-one, which include the following variables: *ind_id, ind_definition, reportyear, race_eth_code, race_eth_name, genotype, geotypevalue, geoname, country_name, country_fips, region_name, region_code, mode, severity, injuries, totoalpop, poprate, LL95cl_poprate, UL95Cl_proprate, prprate_se, poprate_rse, CA_decile_pop, CA_RR_poprate, avmttotal, avmtrate, LL95Cl_avmtrate, UL95Cl_avmtrate, avmtrate_se, avmtrate_rse, CA_decile_avmt, and CA_RR_avmtrate*.

The variables employed in the analysis after dimensionality reduction and further preprocessing, as discussed in the previous section, include the following: *reportyear, county_name, mode, severity, injuries, totalpop,* and *poprate.* The variable *reportyear* is a discrete variable that indicates the years that the accident was reported, while the variable *county_name* is a categorical variable that indicates the name of the county where the accident took place. The variable *mode* is a categorical variable indicating the mode of transportation of the victim (bicyclist, bus, car, motorcycle, pedestrian, truck, vehicles), while the variable *severity* is a binary variable which indicates the severity of the victim's injuries (killed or severely injured). The variable *injuries* is a numeric variable that indicates the number of injuries reported, while the variable *totalpop* is a numeric variable that indicates the total population reported in the geographic area. The variable *poprate* is a numeric variable indicating the rate of injuries over the total population in the geographic area.

Exploratory Data Analysis

In addition to preprocessing and wrangling the data, exploratory data analysis was also performed on the dataset. Exploratory data analysis was conducted in order to gain insight of patterns and possible relationships within the dataset. From the seven independent variables, three were identified as quantitative variables. The three variables are as follows: *injuries*, *totalpop*, and *poprate*. The descriptive statistics for each of the quantitative variables can be seen in Table 1. Within Table 1, the mean, standard deviation, minimum, first quartile value, median, third quartile value, and maximum were included. From the table, it can be seen that the average number of injuries for Southern California were approximately seven per report. The maximum number of injuries for Southern California per report was seventy-five, while the minimum number of injuries per report was one. Additionally, on average, the total population for Southern California was approximately 255,247 people, while the average rate of injury per population was 27%.

Table 1.

Descriptive Statistics of the Quantitative Variables

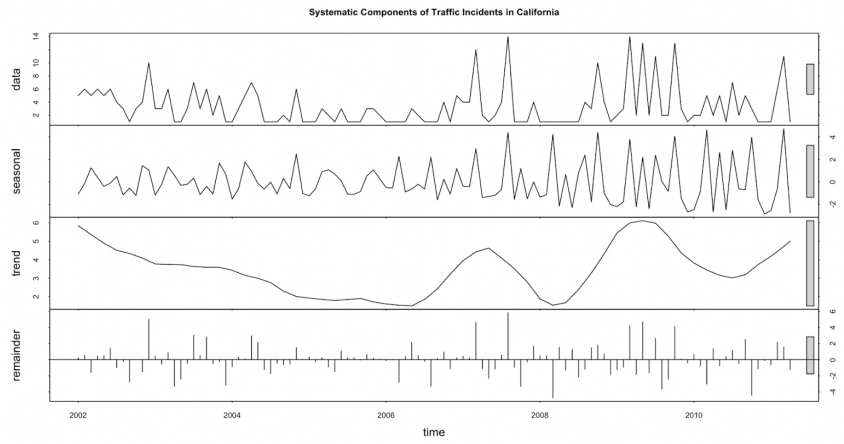
	Mean	Standard Deviation	Minimum	Q1 (25%)	Median	Q3 (75%)	Maximum
injuries	6.79E+00	1.22E+02	1.00E+00	6.70E-01	1.00E+00	2.00E+00	7.50E+01
totalpop	2.55E+05	1.86E+06	9.21E+02	4.42E+03	9.17E+03	6.25E+04	3.57E+05
poprate	2.77E+01	2.44E+02	0.00E+00	3.79E+00	7.93E+00	1.91E+01	3.80E+01

Additionally, a general time series analysis was performed on the data as seen in Figure 1. From this figure, it is evident that this data consists of a non-stationary series. Furthermore, it should be noted that the time plot produced is a decomposition of the time series data. This decomposition plot essentially works by breaking down the data into its systematic components of the data itself, the seasonal trends, overall trend, and irregular components using Loess (Bonaros, 2022). Based on this breakdown, the components present in this series are the four components of level, seasonality, trend, and noise. The level of the series is the average value for

the specific time period, which in this case is from 2002 to 2010. It can be determined by finding the average value of each of the four quarters for the time series. Next, within Figure 1, the seasonality of the series can be seen in the second row of the decomposition plot. It is evident, based on the plot, that the data spikes in such a way that indicates an upward linear trend with additive seasonality. Following the seasonality section in Figure 1 is the trend section. In looking at the trend section for the data, it is evident that the amount of traffic-related injuries steadily decreased from 2002 to 2006 and then fluctuated in an upward trend from 2006 to 2010. There is also evidence of noise present as seen in the last section of the decomposition plot.

Figure 1.

Time Series Analysis Decomposition Plot



Modeling

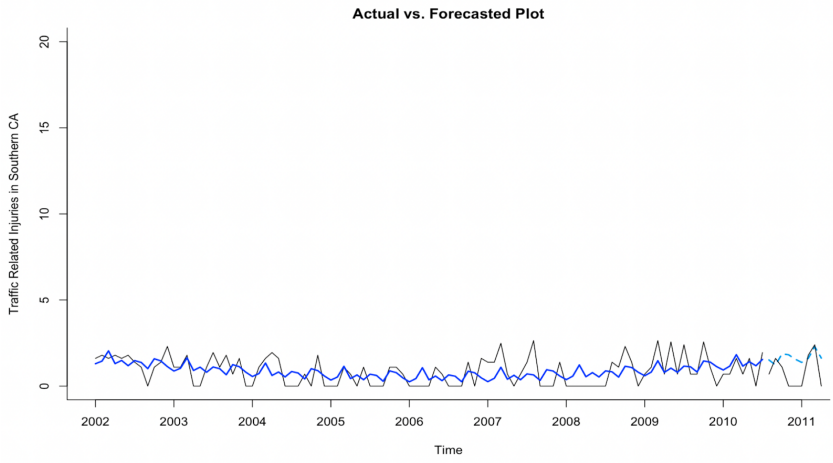
Model 1: Linear Regression

The first model explored within this time series analysis of traffic-related injuries within Southern California was a linear regression model. One key aspect of linear regression models is that they are a popular forecasting technique that capitalizes on suitable predictors to capture trends as well as other patterns (Shmueli & Lichtendah Jr., 2018). This is important because one of the key findings determined from the initial exploratory analysis of the data was that it

contained a downward trend from 2002 to 2006 and then had an upward trend. These trends would be better captured by a linear regression forecasting model, which is why the linear regression model was employed as the first model. The way in which this model was set up was by employing the time series linear model (*tslm)* function, with the formula *train_f.ts ~ trend + I(trend^2) + season,* to produce a quadratic trend model with seasonality. The *y* variable was specified as the number of injuries, while the two *x* variables were the trend and square of trend along with seasonality. The *forecast* function was then employed to make predictions on the *train_f.lm.trend.season* model where *h = nValid_f* steps ahead in the validation period. The forecasting results were then plotted as seen in Figure 2.

Figure 2.

Linear Regression Model Forecasting Plot



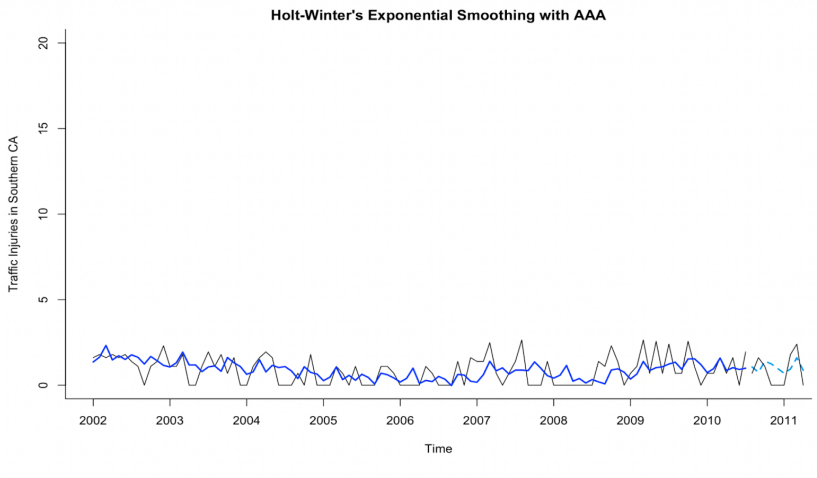
Model 2: Holt-Winter's Exponential Smoothing AAA

The second model explored within this time series analysis of traffic-related injuries within Southern California was a Holt-Winter's Exponential Smoothing AAA model. In general, for a series that contains a trend, a double exponential smoothing model is recommended (Shmueli & Lichtendah Jr., 2018). A double exponential smoothing model is commonly referred to as Holt's linear trend model. An adaptive method of Holt's linear trend model is Holt-Winter's

exponential smoothing model (Shmueli & Lichtendah Jr., 2018). This model is sufficient for capturing level, trend, and seasonality patterns that change over time. As previously discussed in the *Exploratory Data Analysis* section, the data requires the capturing of level, trend, and seasonality (Figure 1); therefore, this model is appropriate to be explored for this data. The way in which this model was set up was by using the *ets* function with the option model set to "AAA" to fit Holt-Winter's exponential smoothing model with additive error, additive trend, and additive seasonality. The *forecast* function was then used to make predictions on the *hwin_AAAf* model where $h = nValid\_f$ steps ahead in the validation period. The forecasting results were then plotted as seen in Figure 3.

Figure 3.

Holt-Winter's Exponential Smoothing AAA Model Forecasting Plot



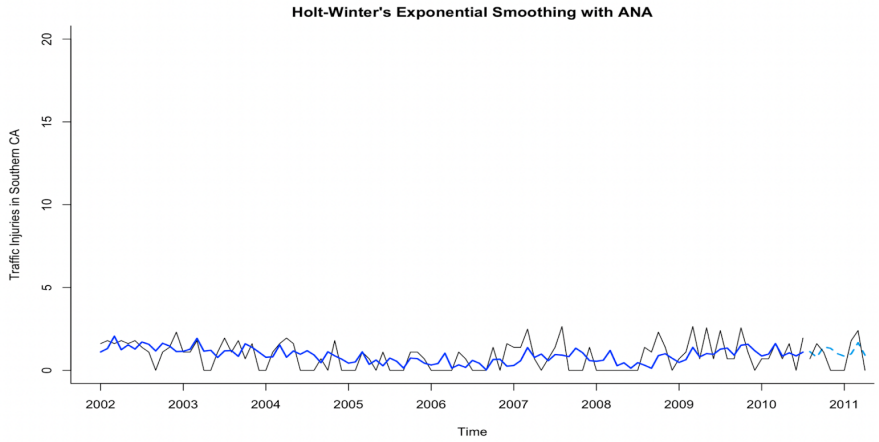
Model 3: Holt-Winter's Exponential Smoothing ANA

The third model explored within this time series analysis of traffic-related injuries within Southern California was a Holt-Winter's Exponential Smoothing ANA model. As previously discussed in the *Model 2: Holt-Winter's Exponential Smoothing AAA* section, Holt-Winter's Exponential Smoothing model was employed as it readily captures level, trend, and seasonality

patterns. The difference in this model, as opposed to the previous model, was adjusting the trend from additive to absent in order to explore and examine how this would affect the forecasted results. The way in which this model was set up was by using the *ets* function with the option model set to "ANA" to fit Holt-Winter's exponential smoothing model with additive error, no trend, and additive seasonality. The *forecast* function was then employed to make predictions on the *hwin_ANAf* model where $h = nValid\_f$ steps ahead in the validation period. The forecasting results were then plotted as seen in Figure 4.

Figure 4.

Holt-Exponential Smoothing ANA Model Forecasting Plot



Model 4: Autoregressive Integrated Moving Average (ARIMA) Model

The fourth model explored was the ARIMA Model, which is generally used for forecasting time series data. The ARIMA model typically accounts for three important characteristics: p, d, and q. These refer to the number of autoregressive terms, the number of MA terms, and the number of integrative terms respectively (Shmueli & Lichtendah Jr., 2018). The ARIMA model predicts values by using past data. In order to construct the ARIMA model, the training data was first used and fitted to a second-order linear model using the *tslm* function in R. This accounted for trend and seasonality. This was then implemented onto the ARIMA model for

forecasting. ACF and PACF residual plots were utilized to identify the order of the ARIMA model as seen in Figure 5. The order was determined to be as follows using a 95% confidence interval: p=3, d=2 and q=3. The *forecast* function was then used to predict the future values from 2010 to 2011, and the trend can be seen in Figure 6 where the actual values and the forecasted values are plotted.

Figure 5

ACF and PACF plots

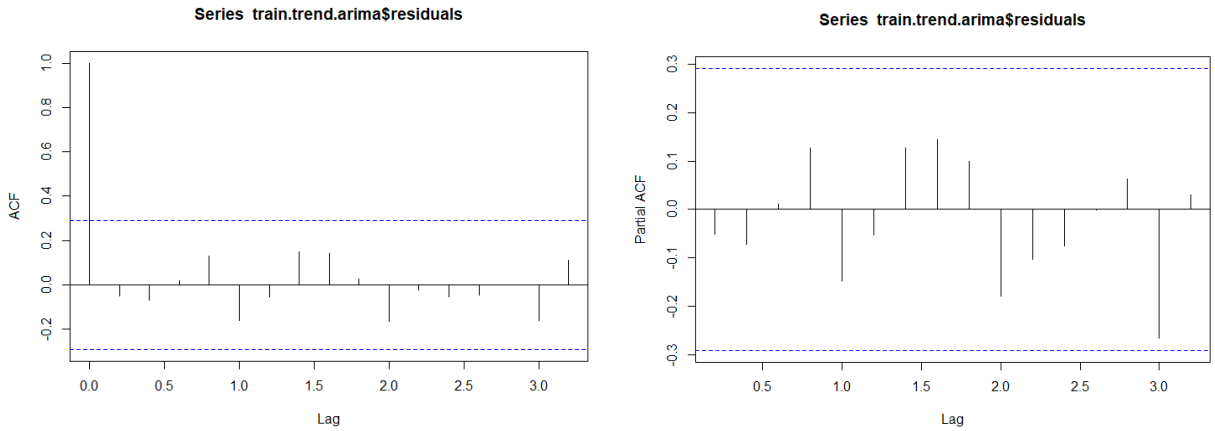
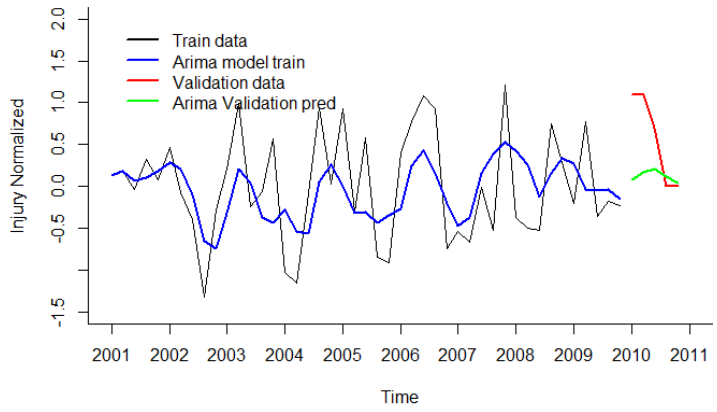


Figure 6.

ARIMA Model Forecasting Plot

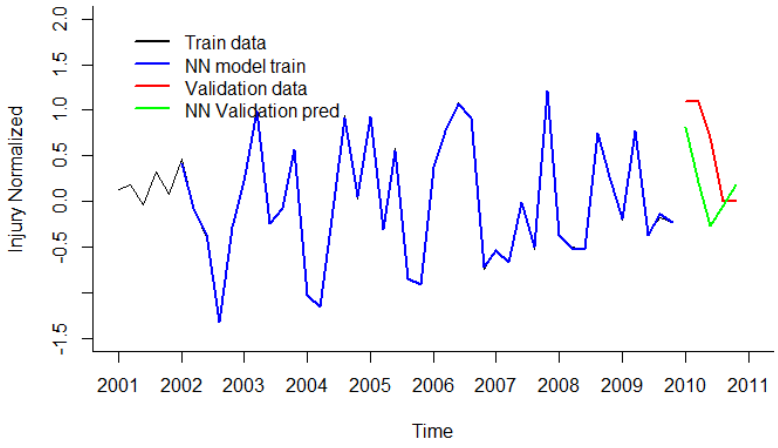


Model 5: Neural Network

The fifth model explored within this time series analysis of traffic-related injuries within Southern California was a Neural Network model. In order to construct the Neural Network model, the training data was first used and fitted to a second-order linear model using the *tslm* function. This accounted for trend, trend squared, and seasonality. The residuals of the second-order linear model were then implemented onto the Neural Network model for forecasting. The Neural Network modeling itself was carried out using the *nnetar* function with the following specifications: the number of seasonal lags to be used as inputs (P) set to 1, number of nodes in the hidden layer (size) set to 7, and the number of networks to fit with different random starting weights (repeats) set to 20. The Neural Network forecasting results can be visualized in Figure 7.

Figure 7.

Neural Network Model Forecasting Plot



Results

The models employed for this study were the Linear Regression Model, Holt-Winter's Exponential Smoothing AAA Model, Holt-Winter's Exponential Smoothing ANA Model,

ARIMA Model, and Neural Network Model. In order to evaluate the performance of the models the following evaluation metrics were observed: RMSE score, MAE score, MPE score, MAPE score, and MASE score. The Linear Regression Model had an RMSE score of 1.14, an MAE score of 0.96, an MPE score of -0.81, an MAPE score of 1.14, and an MASE score of 0.96. The Holt-Winter's Exponential Smoothing AAA Model obtained an RMSE score of 0.82, an MAE score of 0.77, an MPE score of -0.21, an MAPE score of 0.82, and an MASE score of 0.77. The Holt-Winter's Exponential Smoothing ANA Model obtained an RMSE score of 0.89, an MAE score of 0.80, an MPE score of -0.18, an MAPE score of 0.89, and an MASE score of 0.80. The ARIMA Model obtained an RMSE score of 0.54, an MAE score of 0.46, an MPE score of 94.35, an MAPE score of 187.45, and an MASE score of 0.53. The Neural Network Model obtained an RMSE score of 0.24, an MAE score of 0.16, an MPE score of -0.81, an MAPE score of 4.08, and an MASE score of 0.18. It should be noted from the results that the Holt-Winters Exponential Smoothing models of AAA and ANA performed of similar proficiency. Additionally, the model that performed the best of all the models, based on the aforementioned evaluation metrics, was the Neural Network Model, while the model that performed inferiorly was the Linear Regression Model. All of the aforementioned results are summarized in Table 2.

Table 2.

Evaluation metric results for each of the models.

	RMSE	MAE	MPE	MAPE	MASE
Linear Regression	1.14	0.96	-0.81	1.14	0.96
AAA	0.82	0.77	-0.21	0.82	0.77
ANA	0.89	0.8	-0.18	0.89	0.8
ARIMA	0.54	0.46	94.35	187.45	0.53
Neural Network	0.24	0.16	-0.81	4.08	0.18

Discussion

The general hypothesis of this study is that at least one of the five time series modeling techniques utilized will result in a model that forecasts the overall trend of injuries within Southern California. More specifically, a secondary hypothesis is that the best model will be the ARIMA model with the success criteria being low scores for RMSE, MAE, MPE, MAPE, and MASE. The general hypothesis was achieved through the Neural Network model, which predicted an upward trend of traffic-related injuries as seen in Figure 7. The secondary hypothesis did not have sufficient evidence for ARIMA to be the best model, as the Neural Network model outperformed the ARIMA model constructed based on the following evaluation metrics RMSE, MAE, MPE, MAPE, and MASE. The proficient performance of the Neural Network model seen in this study may be a result of this model's ability to identify intricate patterns despite the complexity of the time series data it is working with. Furthermore, Neural Network models are essentially the updated versions of linear regression and autoregressive models. This is the case as the incorporation of the hidden layers is one of key characteristics that allow Neural Networks to be more advanced than the linear regression and autoregressive models (Shmueli & Lichtendah Jr., 2018). In the future, researchers can look at the incorporation of variables such as perpetrators of high blood alcohol content, seatbelt use, phone use, gender of perpetrator, and time of day in order to improve the study. Additionally, the incorporation of more high frequency data may help improve the forecasted results. Furthermore, the exploration of combining a 3-point MA smoothing or a singular value decomposition of the Hankel matrix with the Neural Network model may produce interesting results to be explored (Barba et. al., 2014).

References

Barba, L., Rodríguez, N., & Montt, C. (2014). *Smoothing strategies combined with ARIMA and neural networks to improve the forecasting of traffic accidents.* The Scientific World Journal, 2014. https://www.hindawi.com/journals/tswj/2014/152375/

Bhattachan, A., Okin, G. S., Zhang, J., Vimal, S., & Lettenmaier, D. P. (2019). *Characterizing the role of wind and dust in traffic accidents in California*. GeoHealth, 3(10), 328-336. https://agupubs.onlinelibrary.wiley.com/doi/full/10.1029/2019GH000212

Bonaros, B. (2022, July 4). *Time Series Decomposition In Python - Towards Data Science*. Medium.https://towardsdatascience.com/time-series-decomposition-in-python-8acac385a

Kenneth M. Sigelman & Associates. (2022, October 3). 2022 *California Car Accident Statistics - Traffic Fatalities CA*. https://sigelmanassociates.com/california-car-accident-statistics/

Road Traffic Injuries and Deaths—A Global Problem. (2020, December 14). Centers for Disease Control and Prevention. https://www.cdc.gov/injury/features/global-road-safety/index.html

Shmueli, G. & Lichtendahl Jr., K.C. (2018). *Practical time series forecasting with R: A hands-on guide (2nd ed.)*. Axelrod Schnall Publishers.

Yousefzadeh-Chabok, S., Ranjbar-Taklimie, F., Malekpouri, R., & Razzaghi, A. (2016). *A time series model for assessing the trend and forecasting the road traffic accident mortality.* https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5079210/