Time series analysis of motor vehicle collisions in the United States

Joanne Micaela R. Dizo, Joan Isabel V. Yu, Dianne C. Yumol

Abstract

The United States faces a significant challenge with motor vehicle crashes due to its car-centric culture and high ownership rates. Over 6 million collisions annually result in thousands of fatalities, with drunk driving and speeding as leading causes. With strained emergency resources, addressing this issue through estimating anticipated accidents can enable better resource allocation and proactive interventions to ensure safer roads for all. The 'US Accidents' dataset offers comprehensive traffic incident records from 48 states, spanning January 2016 to March 2023. Exploratory tests indicated autocorrelation and non-stationarity, necessitating further preprocessing. After data transformation, researchers employed a range of time series models to analyze and forecast traffic accident occurrences. Identifying significant lags led to the ARMA(4, 2) model being deemed most suitable. The model's effectiveness suggests that accident counts are influenced by past values and errors, capturing systematic trends and abrupt disturbances. The decrease in accidents during August and September 2020 aligns with the COVID-19 pandemic's impact, highlighting the model's ability to reflect changes in traffic patterns due to lockdowns. Furthermore, the increased variance post-pandemic underscores the need for adaptive modeling techniques to address evolving traffic conditions and ensure road safety.

1 Introduction

1.1 Background

In 2022, there were approximately 283 million private, public, and commercial cars, trucks, buses, and motorcycles registered in the United States, a figure that continues to grow annually (Rogers, 2023). Remarkably, 91.7% of households owned at least one vehicle, leaving only about 10.9 million out of the total 131.2 million American households without a single vehicle (Valentine, 2024). This underscores how car ownership has become as a symbol of the American dream. It is deeply ingrained in their culture and history that driving a private vehicle is seen as a display of status and economic capability. Consequently, the government prioritizes investment in road infrastructure over the state of public transport, perpetuating the car-centric culture.

This level of car traffic significantly contributes to the number of road accidents each year. In fact, there are estimated to be over 6 million collisions annually, with drunk driving and speeding remaining the leading causes (Bieber, 2023). This phenomenon results in thousands of deaths in the American population, totaling 42,795 in 2022 alone. A staggering 32% of these fatalities involve an intoxicated driver, with blood alcohol levels (BAC) of .08 g/dL or higher. Additionally, statistics on teenagers driving under the influence are concerning, with over 10% of all American teenagers admitting to having experienced drinking and driving, while 17% have stated they have been driven by an individual who had consumed alcohol beforehand. This has led to approximately 29% of teen drivers involved in fatal motor accidents testing positive for being over the legal alcohol limit. On the other hand, 29% of deadly car collisions occur due to speeding drivers, as speeding not only increases the risk of collisions but also significantly raises the chances of severe or fatal injuries.

Submission for MATH 62.2 Time Series and Forecasting, AY 2022-2023 Second Semester

All these numbers add up to a fatal car accident occurring every 15 minutes in the United States. Notably, these figures do not solely involve participants in the country's car culture, as pedestrians account for about 17.7% of all car accident deaths. In 2022, 437,677 minors sustained injuries in motor collisions, with 4,414 succumbing to their wounds. This makes motor vehicle crashes the leading cause of death in the United States.

Fortunately, America boasts one of the most effective emergency response systems globally, with the renowned 9-1-1 being used as a global symbol for emergencies. The country continues to invest funds into this system, leading to significant advancements such as the ability to call the line even without service and the network accommodating individuals with hearing or speech impairments or those unable to speak English (Why 9-1-1 Is the Emergency Number, 2023). This service has resulted in an average Emergency Medical Service (EMS) response time of 9 minutes, although this still leaves room for improvement considering the possibilities that could occur within that time frame (Byrne et al., 2019). Cutting this by half a minute could potentially save hundreds of lives.

1.2 Statement of the Problem

The United States is grappling with a serious issue concerning the staggering number of motor vehicle crashes that occur annually, and the number of fatalities that follow. This problem is exacerbated by the challenge of efficiently managing emergency resources, which are already spread thin across the nation due to a combination of relatively low population density and insufficient funding.

To address this, there is an urgent need to develop predictive models that can estimate the number of accidents likely to occur within specific time frames and particular geographic areas. By anticipating these incidents more accurately, emergency services can be better prepared, leading to a more effective allocation of resources. This proactive approach could enhance traffic regulations, increase public awareness, and ultimately improve overall public safety. Implementing such predictive measures could result in more strategically placed emergency personnel, quicker response times, and a reduction in the severity of crash outcomes. Moreover, better-informed traffic policies and targeted public safety campaigns could contribute to a decrease in accident rates and fatalities, ensuring safer roads for all.

1.3 Scope and Limitations

The primary focus of this study is to analyze the patterns of motor vehicle accidents in the United States with the "US Accidents" dataset, which covers incidents from January 2016 to March 2023 across 48 states. The study employs time series models, specifically autoregressive (AR), moving average (MA), and autoregressive moving average (ARMA) models, to forecast the number of monthly accidents. The project aims to offer a predictive framework that can aid in better resource allocation, policy-making, and public safety initiatives.

The scope of the study includes:

1. *Data Collection and Preprocessing*: Obtaining and cleaning traffic incident data to be further analyzed for patterns and trends.

2. *Exploratory Analysis*: Conducting tests for autocorrelation and stationarity to understand the underlying characteristics of the data.

3. *Model Development*: Building AR, MA, and ARMA models to capture the temporal dependencies in the car accidents data.

4. *Model Evaluation*: Using information criteria such as AIC, AICc, and BIC to select the most suitable model.

5. *Forecasting*: Predicting future accident occurrences and validating the model's performance through residual analysis.

It is to be noted that the data is limited to accidents occurring within the Contiguous United States, which refers to states that share common borders, thus excluding Hawaii and Alaska. Additionally, Washington D.C., the nation's capital, is referred to as a state in the dataset, resulting in 49 unique entries in the State column.

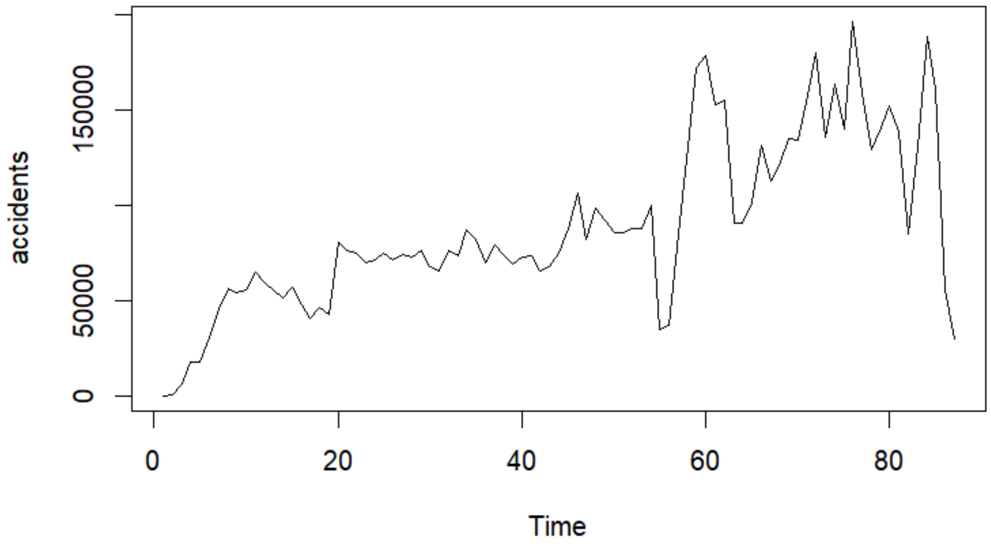

Figure 1: Monthly traffic accident rates in the US (January 2016 - March 2023).

Furthermore, the dataset may have missing data for certain days, likely due to network connectivity issues during data collection. These gaps can affect the continuity and reliability of the time series analysis and might lead to slight inaccuracies in modeling and forecasting.

2 Methods

2.1 Data Description and Cleaning

The “US Accidents” dataset is a countrywide car accident dataset that covers 48 states of the USA. The dataset comprises a comprehensive collection of traffic incident records spanning from January 2016 to March 2023. This dataset was assembled by aggregating real-time traffic incident data obtained from various sources via multiple Traffic APIs. The streaming data sources include inputs from state and federal departments of transportation, law enforcement agencies, traffic cameras, and road-based traffic sensors operating across the Contiguous United States.

As of the dataset’s latest compilation, it encompasses approximately 7.7 million accident records, each documenting a specific traffic-related event occurring within this timeframe. These events encompass a wide range of accident types, severity levels, and geographical locations across the contiguous United States.

The dataset’s primary focus is on traffic accidents and related incidents captured over a period of seven years, making it a valuable resource for understanding patterns, trends, and factors influencing traffic safety and incident occurrences.

To facilitate exploratory data analysis and subsequent modeling efforts, the initial preprocessing steps involved grouping the accident records by year and month. This preprocessing step enables a structured approach to analyzing and forecasting the monthly trends in accident occurrence.

The visualization in Figure 1 depicts the monthly distribution of traffic accidents.

2.2 Exploratory Analysis

Upon loading the US Accidents dataset, we initiated exploratory tests to better understand the underlying patterns and characteristics of the data. Two key analyses conducted during this phase were tests for autocorrelation and unit-root tests. For the statistical tests, we will adhere to a significance level of $\alpha = 0.05$.

To assess autocorrelation within the dataset, we employed the Ljung-Box Test, a statistical test used to determine whether autocorrelation is present in residuals from a time series analysis (Brockwell & Davis, 2016). The Ljung-Box Test is structured around the following hypotheses:

H_0: The data are independently distributed (i.e., no autocorrelation).

H_a: The data are not independently distributed; they exhibit serial correlation (autocorrelation).

The goal of this test is to fail to reject the null hypothesis, indicating that the data are not independent and there is autocorrelation.

Upon conducting the Ljung-Box Test on our dataset, we obtained a calculated p-value of $1.232e^{-14}$, which is significantly less than 0.05. Consequently, we reject the null hypothesis (H_0) and conclude that autocorrelation is present within the dataset.

In addition to autocorrelation assessment, we conducted a unit-root test using the augmented Dickey-Fuller Test (ADF Test) to evaluate the stationarity of the time series data within the US Accidents dataset.

The ADF Test is a statistical method used to determine whether a time series is stationary or exhibits a unit root, indicating non-stationarity (Brockwell & Davis, 2016). The test hypotheses are structured as follows:

H_0: The presence of a unit root; the time series data is non-stationary.

H_a: The absence of a unit root; the time series data is stationary.

Upon conducting the ADF Test on our dataset, the initial p-value obtained was 0.07705, which exceeds the significance level of 0.05. Therefore, we fail to reject the null hypothesis (H_0), suggesting that the time series data is non-stationary.

Given the non-stationarity of the dataset, preprocessing steps are essential to prepare the data for further analysis and modeling. As noted by Hyndman and Athanasopoulos (2018), two prevalent techniques employed to stabilize time series data are logarithmic transformation and differencing.

Logarithmic transformation serves as an effective method to stabilize the variance of a time series, especially when the data displays heteroscedasticity (unequal variance over time). On the other hand, differencing entails computing the discrepancies between consecutive observations in a time series. This approach aids in stabilizing the mean of the time series by eliminating alterations in the level, thus mitigating the influence of trend and seasonality effects.

Now, after performing logarithmic-differencing, the p-value obtained by performing ADF is calculated to be smaller than 0.05. Thus, we reject the null hypothesis and conclude that the data exhibits stationarity.

2.3 Models

In this data analysis project, we will leverage a range of time series models to analyze and forecast traffic accident occurrences within the US Accidents dataset. The primary models employed will include autoregressive (AR) models, moving average (MA) models, and autoregressive moving average (ARMA) models.

To understand the relevance of these models, it is essential to define what constitutes a time series. According to Brockwell and Davis (2016), a time series is a set of observations, each one being recorded at a specific time. That is, it is a collection of random variables indexed according to the order they are obtained in time.

Now, a time series model for the observed data $\{x_t\}$ is a specification of the joint distributions (or possibly only the means and covariances) of a sequence of random variables $\{X_t\}$ of which $\{x_t\}$ is postulated to be a realization.

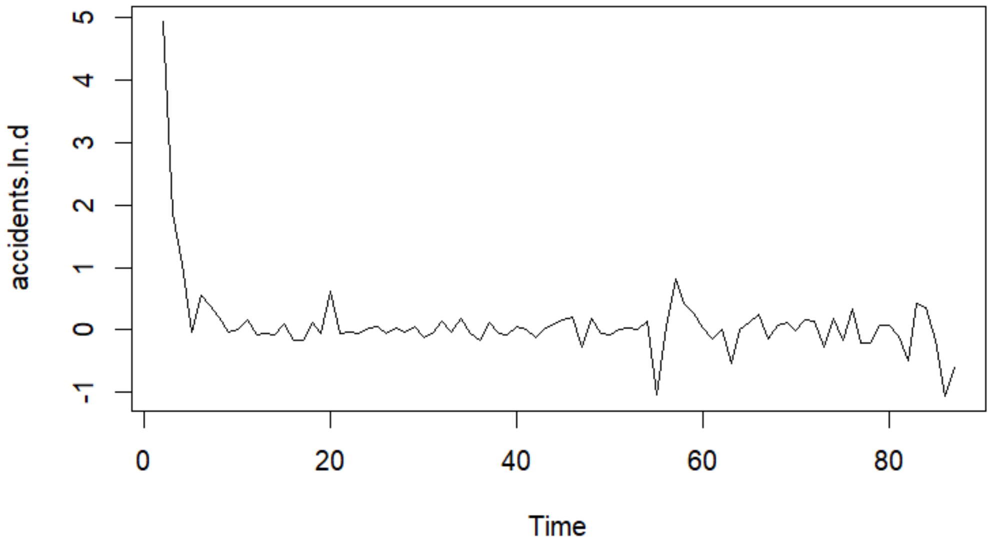

Figure 2: Monthly traffic accident rates in the US (January 2016 - March 2023) after applying logarithmic-differencing.

An example of a simple time series model is white noise, a collection of uncorrelated random variables, $\{w_t\}$, with zero mean $\mu = 0$ and finite variance σ_w^2. It is denoted as $w_t \sim WN(0, \sigma_w^2)$ (Brockwell & Davis, 2016).

Moreover, a time series $\{X_t\}$ is an autoregressive process of order p, if

$$X_t = \phi_0 + \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + Z_t, \tag{1}$$

where $\{Z_t\} \sim WN(0, \sigma^2)$ and Z_t is uncorrelated with X_s for each $s < t$ (Brockwell & Davis, 2016).

AR models predict future values based on past values in the time series. The key assumption is that the current value of the series can be expressed as a linear combination of its previous values (Hyndman & Athanasopoulos, 2018).

Additionally, a time series $\{X_t\}$ is a moving average process of order q, if

$$X_t = \mu + Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q}, \tag{2}$$

where $\{Z_t\} \sim WN(0, \sigma^2)$ (Brockwell & Davis, 2016).

MA models incorporate the influence of past forecast errors to improve predictions. They achieve this by expressing the current value of the series as a linear combination of weights applied to a specific number of past error terms. (Hyndman & Athanasopoulos, 2018).

Both AR and MA models (and their combinations) are foundational in time series forecasting, and their applicability depends on the characteristics of the data and the nature of the underlying processes generating the time series. Often, these models are combined to model and forecast time series data more effectively.

One of the most commonly used is Autoregressive Moving Average (ARMA). A time series $\{X_t\}$ is ARMA(p,q) if it is stationary and

Table 1: AIC, AICc, and BIC values of the different time series models.

	Time Series Model	AIC	AICc	BIC
1	AR(1)	132.46	134.75	141.82
2	AR(4)	135.47	138.53	152.19
3	MA(2)	135.73	138.22	147.55
4	ARMA(1,2)	134.09	136.84	148.37
5	ARMA(4,2)	130.24	134.11	151.87

$$X_t = \phi_1 X_{t-1} + \cdots + \phi_p X_{t-p} + +Z_t + \theta_1 Z_{t-1} + \cdots + \theta_q Z_{t-q}, \tag{3}$$

where $\phi_p \neq 0, \theta_q \neq 0$, and $\sigma^2 > 0$. The parameters p and q are called the autoregressive and moving average orders, respectively (Brockwell & Davis, 2016).

Usually, the term stationary means weakly stationary. Note that a time series is said to be (weakly) stationary if it satisfies the following conditions:

- $\mathbb{E}(X_t) = \mu_{X_t} = \mu < \infty$ (i.e., the expectation of X_t is finite and does not depend on t); and
- $\gamma(t+h, t) = \gamma_h$ (i.e., for each lag h, the autocovariance of random variables X_{t+h} and X_t is constant for for a given lag h.

Having established a foundational understanding of time series models and their theoretical underpinnings, we are now positioned to advance to the model building phase. In this phase, we will apply these methodologies to develop our models.

2.4 Model Building

With a focus on understanding and capturing the temporal dependencies inherent in the US Accidents dataset, we transitioned from exploratory analysis to time series modeling. The modeling process begins with lag order selection, where we identify the optimal lag values based on the autocorrelation function (ACF) and partial autocorrelation function (PACF).

To determine the appropriate lag order for our time series modeling, we performed an analysis of the ACF and PACF using statistical functions available in R. The ACF helps identify the presence of moving average (MA) components in a model by showing correlations between the time series and itself at different lags. The PACF isolates the correlation at a specific lag by controlling for the influence of intermediate lags, aiding in the identification of autoregressive (AR) components.

Based on the principle of parsimony—preferring simpler models without compromising model fit—we considered lag values up to $h = 5$ for both ACF and PACF analysis. The goal was to identify significant lag values that capture the key autocorrelation structure of the time series data.

In Figure 3, we observed that ACF values tended to cut off at lag $h = 2$, suggesting potential significant autocorrelation up to this point. The PACF values, as seen in Figure 4 indicated significant autocorrelation at lag $h = 1$ and with a noticeable drop-off beyond these points. However, to broaden our modeling options and account for potential variations, we also considered the second-best lag values identified from the analysis. This included exploring lag $h = 4$ based on the PACF results. As such, we test AR(1), AR(4), MA(2), ARMA(1,2), and ARMA(4,2).

Following the lag order determination and initial time series modeling, we proceeded to evaluate and select the most suitable model based on information criteria such as the Akaike Information Criterion (AIC), corrected Akaike Information Criterion (AICc), and Bayesian Information Criterion (BIC).

The Akaike Information Criterion (AIC) is a statistical measure used to assess the goodness of fit of a model while penalizing for model complexity (Brockwell & Davis, 2016). Lower AIC values indicate a better trade-off between model fit and complexity.

Upon evaluating different time series models, we identified that the ARMA(4, 2) model demonstrated the most favorable performance based on the AIC criterion (see Table 1). The ARMA(4, 2) model yielded an AIC value of 130.24, suggesting a strong fit to the data relative to other considered models.

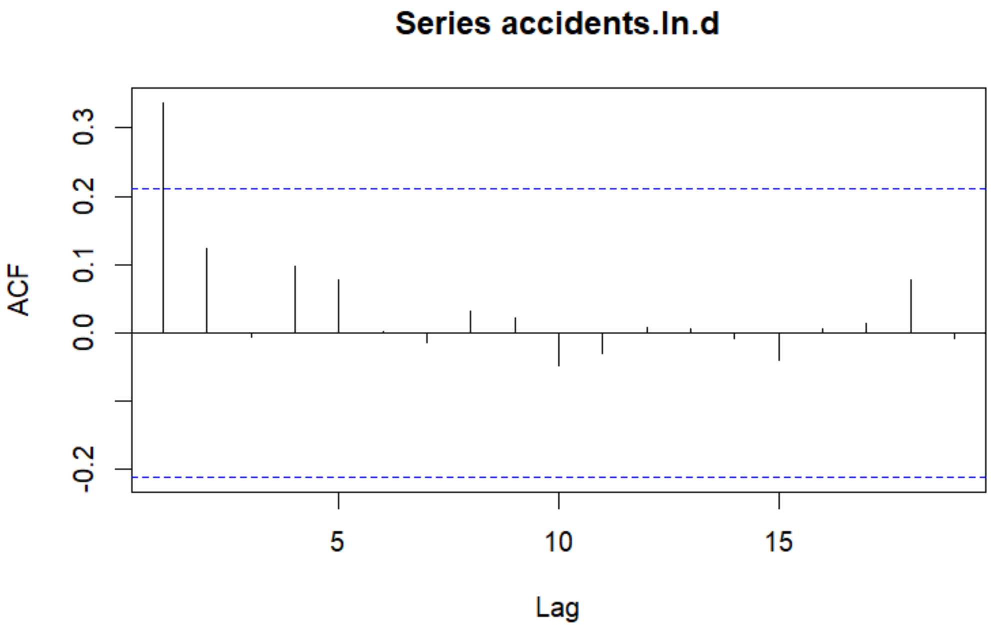

Figure 3: ACF plot.

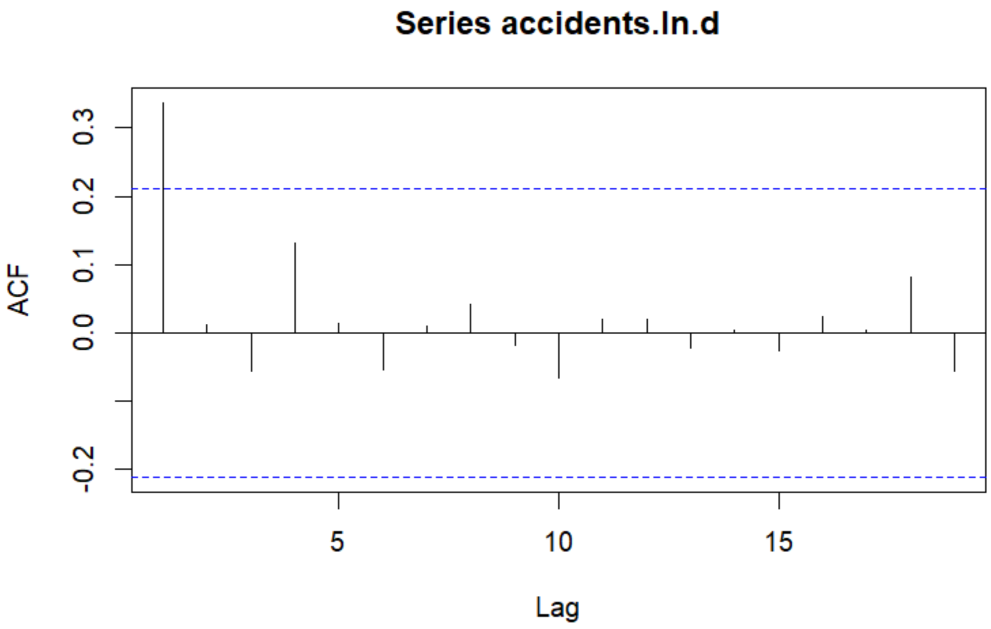

Figure 4: PACF plot.

In addition to the AIC, we also examined the Corrected Akaike Information Criterion (AICc) and Bayesian Information Criterion (BIC) to further validate our model selection. The AICc is a variant of AIC that provides a correction for small sample sizes, penalizing additional model parameters more heavily (Hastie, Tibshirani, & Friedman, 2009). For the ARMA(4, 2) model, we obtained an AICc value of 134.11.

The BIC, on the other hand, is another information criterion that balances model fit and complexity, with a greater penalty for additional parameters compared to AIC (Hastie, Tibshirani, & Friedman, 2009). The ARMA(4, 2) model yielded a BIC value of 151.87.

2.5 Parameter Estimates

Given that the data follows an ARMA(4, 2), its functional form is given by

$$\begin{aligned} X_t = 0.5076035X_{t-1} - 0.4258054X_{t-2} + 0.5372309X_{t-3} + 0.2576482X_{t-4} \\ + Z_t + 0.3954321Z_{t-1} + 0.9999999Z_{t-2} \end{aligned} \tag{4}$$

While the results yielded $\mu = 0.6661451$, it can be noticed that the confidence interval for μ is $0.666145 \pm 1.96e$. Since the interval contains 0, it can be argued that μ is not significantly different from 0.

3 Results and Discussion

The preference for the ARMA(4,2) model over purely AR or MA models reveals significant insights into the nature of monthly accident data in the US. The ARMA model's effectiveness suggests that the number of accidents each month is influenced by a combination of past values (autoregressive component) and past errors (moving average component). The autoregressive component (AR(4)) implies that recent months' accident counts directly influence the current month's count, potentially due to factors such as seasonal trends, ongoing road conditions, or policy effects. The moving average component (MA(2)) indicates that past errors or shocks also play a significant role, suggesting that unexpected events (e.g., sudden changes in weather, major incidents) from the recent past impact the current month's accident count.

Moreover, the sudden decrease in accidents observed during August and September 2020 aligns with the impact of the COVID-19 pandemic, where widespread lockdown measures and mobility restrictions led to reduced travel and traffic volume across the US (Ebrahim Shaik & Ahmed, 2022). This decline in accidents mirrors the global trend observed during the pandemic, where a sharp decrease in traffic volume resulted in fewer road traffic collisions and road deaths (Yasin et al., 2021). Additionally, the decrease in annual road accidents in the months of 2020 compared to those in 2019 further supports the notion that the pandemic-induced reduction in travel had a significant impact on road safety outcomes (Yasin et al., 2021). These events likely contributed to the model's ability to capture the effects of both systematic trends and abrupt disturbances in accident data, highlighting the importance of a comprehensive approach to modeling monthly accident trends.

Additionally, the sudden increase in variance observed after the pandemic reflects the uncertainty and changing dynamics of traffic patterns as regions gradually reopened and economic activities resumed. The relaxation of restrictions and resumption of travel may have led to more diverse and fluctuating traffic conditions, resulting in higher variability in accident counts. This increased variability underscores the need for adaptive modeling techniques that can capture the evolving nature of accident data and inform timely interventions to ensure road safety.

The ARMA(4,2) model's ability to incorporate both autoregressive and moving average components makes it well-suited to handle such dynamic and uncertain conditions, thereby explaining its superior performance in forecasting monthly accident trends amidst changing traffic patterns.

3.1 Model Validation

After fitting the ARMA(4,2) model, analyzing the residuals is crucial to ensure the model's adequacy. One key assumption of time series models is that residuals should be uncorrelated. This was checked

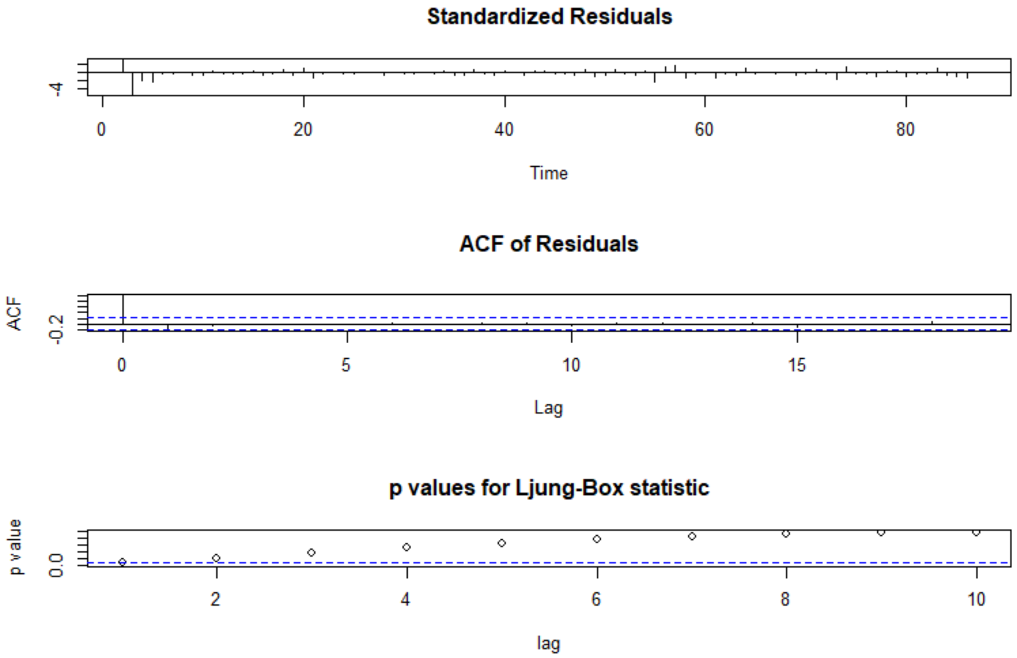

Figure 5: Plots of the standardized residuals, ACF plots of residuals, and p-values from the Ljung-Box test.

using the tsdiag() function, which provides plots of standardized residuals, ACF plots of residuals, and p-values from the Ljung-Box test. Additionally, the Ljung-Box test was performed using the residuals stored in arma42$residuals.

The standardized residual plot indicates that the residuals are centered around zero, suggesting that, on average, the ARMA(4,2) model adequately captures the mean structure of the data. This alignment with the zero line implies that the model's predictions are unbiased, with errors evenly distributed above and below zero. Furthermore, the absence of any discernible pattern or trend in the residuals suggests that the model accounts for the systematic variations in the data. In the ACF plot, no lags extend beyond the dashed lines, indicating that there is no significant autocorrelation in the residuals. This implies that the residuals are independent, meeting one of the key assumptions of time series modeling. The increasing trend in the p-values across lag intervals, with only the p-value at lag 1 touching the significance threshold, suggests that while there may be some minor autocorrelation in the first lag, subsequent lags exhibit increasingly less correlation. Overall, these observations collectively support the adequacy of the ARMA(4,2) model in capturing the underlying dynamics of the monthly accident data.

Furthermore, to assess the normality of residuals, the Jarque-Bera test was conducted. The test yielded a significant result ($X^2 = 333.16, df = 2, p - value < 2.2e^{-16}$), indicating deviations from normality. This suggests the presence of outliers or other non-linear patterns that may not be fully captured by the model.

3.2 Forecasting

After model validation, the ARMA(4,2) model was then utilized to predict the number of accidents for the next 10 months. Forecasting was carried out using both the predict() function for the ARIMA model and the forecast() function from the forecast library in R.

Forecasting for ARMA(p, q) with predict() involves specifying the ARIMA object (in this case, arma42) and the number of steps ahead, n.ahead=m. The forecast values are stored in arma42.p$pred, while the associated prediction errors are in arma42.p$se.

Table 2: Forecasted values of log differences of the number of accidents for the next 10 months.

	predict()	forecast()
1	-0.306427684	-0.306427684
2	-0.328521246	-0.328521246
3	-0.555711926	-0.555711926
4	-0.381715788	-0.381715788
5	-0.130426646	-0.130426646
6	-0.204706118	-0.204706118
7	-0.314470075	-0.314470075
8	-0.158727945	-0.158727945
9	-0.008095641	-0.008095641
10	-0.076056562	-0.076056562

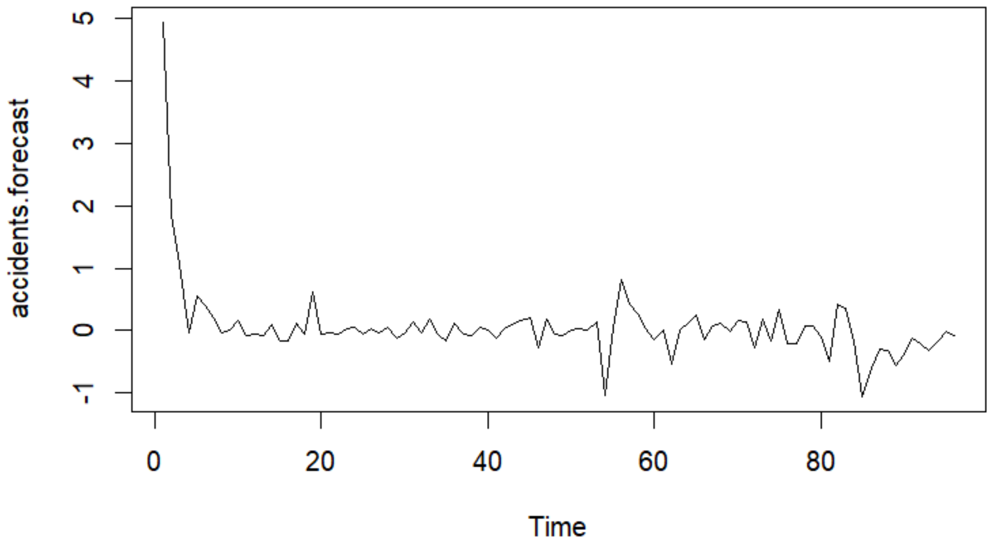

Figure 6: Monthly traffic accident rates, in log-difference, in the US appended with the forecasted values.

On the other hand, the Arima() function from the forecast library offers a comparable estimation and forecasting process, with the advantage of providing prediction intervals and plot compatibility. For Arima(), the parameter n.ahead=m is replaced with h=m, and forecast values are stored in arma42.p2$mean. Additionally, Arima() provides values for AIC, BIC, and AICc, unlike predict(), which only outputs AIC.

The forecasted values (in log differences) using both predict() and forecast() functions were identical as shown in Table 2. The forcasted values are then plotted along the previous values as shown in Figure 6.

3.3 Conclusion

Predictive modeling empowers proactive safety measures and interventions by anticipating the impact of recent events on future accident rates (Agyemang et al., 2023). Real-time data analysis combined with predictive techniques enables timely deployment of emergency services and road closures in response to adverse weather conditions or major traffic incidents (Agyemang et al., 2023). This proactive approach mitigates the severity and frequency of accidents, fostering safer road environments and minimizing potential hazards.

The insights derived from the ARMA(4,2) model also inform policy adjustments and targeted public safety campaigns. For instance, policymakers can use predictive forecasts to implement temporary traffic regulations or public awareness initiatives during periods of anticipated high accident rates, such as holiday seasons. Evidence-based policymaking guided by accurate forecasting leads to more effective strategies for reducing accident-related injuries and fatalities (Obasi & Benson, 2023).

In conclusion, the superior performance of the ARMA(4,2) model underscores the importance of integrating historical trends and recent disturbances in accident data forecasting. Leveraging predictive modeling techniques enables traffic authorities and policymakers to develop more effective traffic safety strategies and make informed decisions aimed at reducing accident rates and enhancing public safety. This integration of advanced analytical tools with real-time data monitoring and proactive interventions holds immense promise for improving road safety and ultimately saving lives.

References

[1] Why 9-1-1 is the emergency number and 10 interesting facts about america's emergency telephone number. `https://rb.gy/x1ppfd`. Accessed: 2024-5-18.

[2] The unique u.s. car culture. `https://www.aii.org/the-unique-u-s-car-culture/`, Nov. 2023. Accessed: 2024-5-18.

[3] E. F. Agyemang, J. A. Mensah, E. Ocran, E. Opoku, and E. N. N. Nortey. Time series based road traffic accidents forecasting via SARIMA and facebook prophet model with potential changepoints. *Heliyon*, 9(12):e22544, Dec. 2023.

[4] P. J. Brockwell and R. A. Davis. *Introduction to time series and forecasting*. Springer Texts in Statistics. Springer International Publishing, Basel, Switzerland, 3 edition, Aug. 2016.

[5] J. P. Byrne, N. C. Mann, M. Dai, S. A. Mason, P. Karanicolas, S. Rizoli, and A. B. Nathens. Association between emergency medical service response time and motor vehicle crash mortality in the united states. *JAMA Surg.*, 154(4):286–293, Apr. 2019.

[6] J. D. Christy Bieber. Car accident statistics for 2023. `https://www.forbes.com/advisor/legal/car-accident-statistics/`, Jan. 2023. Accessed: 2024-5-18.

[7] M. Ebrahim Shaik and S. Ahmed. An overview of the impact of COVID-19 on road traffic safety and travel behavior. *Transportation Engineering*, 9(100119):100119, Sept. 2022.

[8] R. Hastie, T. Tibshirani and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Texts in Statistics. Springer International Publishing, New York, NY, 2 edition, 2009.

[9] R. Hyndman and G. Athanasopoulos. *Forecasting: principles and practice*. OTexts, Melborne, Australia, 2 edition, May 2018.

[10] I. C. Obasi and C. Benson. Evaluating the effectiveness of machine learning techniques in forecasting the severity of traffic accidents. *Heliyon*, 9(8):e18812, Aug. 2023.

[11] A. Valentine. Car ownership statistics 2024. `https://www.forbes.com/advisor/car-insurance/car-ownership-statistics/`, Mar. 2024. Accessed: 2024-5-18.

[12] Y. J. Yasin, M. Grivna, and F. M. Abu-Zidan. Global impact of COVID-19 pandemic on road traffic collisions. *World J. Emerg. Surg.*, 16(1):51, Sept. 2021.

A Appendix

A.1 Data Cleaning in Python

```python
import pandas as pd
import numpy as np

accidents = pd.read_csv("US_Accidents_March23.csv")
```

```python

accidents.head()

accidents['Start_Time'] = pd.to_datetime(accidents['Start_Time'])

# Extract year and month from 'Start_Time'
accidents['Year_Month'] = accidents['Start_Time'].dt.to_period('M')

# Group by 'Year_Month' and count accidents
accidents_by_year_month =
accidents.groupby('Year_Month').size().reset_index(name='Accident_Count')

# Display the DataFrame with summed accidents by year and month
print(accidents_by_year_month)

accidents_by_year_month.to_csv('accidents.csv', index=False)
```

A.2 Time Series Analysis in R

```r
library("TSA")
library("tseries")

data = read.csv("accidents.csv", header = TRUE)

head(data)

accidents = data[,"Accident_Count"]

head(accidents)

accidents = ts(accidents)

plot(accidents)

#Test for stationarity
#H_0: Data is NOT stationary (so we want to reject this)
adf.test(accidents) #not stationary

# perform pre-processing
accidents.ln = log(accidents) # take log transformation
accidents.ln.d = diff(accidents.ln) # take first difference of log

plot(accidents.ln.d)

adf.test(accidents.ln.d) #not stationary

Box.test(accidents, type = "Ljung") # p-value = 1.232e-14 so slay

Box.test(accidents.ln.d, type = "Ljung") # p-value = 0.001474 so slay

#ACF
# Compute for ACF
```

```r
acf(accidents.ln.d) #MA(2) or MA(3)

# Compute for PACF
acf(accidents.ln.d, type="partial") #AR(1) or AR(4)

arima(accidents.ln.d, order=c(1,0,0)) # test AR(1)
arima(accidents.ln.d, order=c(4,0,0)) # test AR(4)

arima(accidents.ln.d, order=c(0,0,2)) # test MA(2)
arima(accidents.ln.d, order=c(0,0,3)) # test MA(3)

arima(accidents.ln.d, order=c(1,0,2)) # test ARMA(1,2)
arima(accidents.ln.d, order=c(1,0,3)) # test ARMA(1,3)
arima(accidents.ln.d, order=c(4,0,2)) # test ARMA(4,2)
arima(accidents.ln.d, order=c(4,0,3)) # test ARMA(4,3)

#We choose ARMA(4,2) as it has the lowest AIC

#install.packages('forecast')
library("forecast")

# from previous result, use ARMA (4, 2)
arma42 = arima(accidents.ln.d, order = c(4, 0, 2))
arma42$coef

# forecast 10-step ahead
arma42.p = predict(arma42, n.ahead =10)

arma42.p$pred #forecast values

arma42.2 = Arima(accidents.ln.d, order=c(4,0, 2))
arma42.2
arma10 = Arima(accidents.ln.d, order=c(1,0, 0))
arma10
arma40 = Arima(accidents.ln.d, order=c(4,0, 0))
arma40
arma02 = Arima(accidents.ln.d, order=c(0,0, 2))
arma02
arma12 = Arima(accidents.ln.d, order=c(1,0, 2))
arma12

arma42.2$coef
arma42.p2 = forecast(arma42.2, h=10)
arma42.p2$mean

arma42.2

#residual diagnostics
tsdiag(arma42)
jarque.bera.test(residuals(arma42))

accidents.forecast = c(accidents.ln.d, arma42.p2$mean)
accidents.forecast = ts(accidents.forecast)
plot(accidents.forecast)
```