FORECASTING MOTOR VEHICLE FATALITIES:

EXAMINING THE IMPACT OF DEMOGRAPHIC FACTORS

by

TADESSE HAILEYESUS

(Under the Direction of Daniel Jung)

ABSTRACT

Background

The Centers for Disease Control and Prevention (CDC) recommends sharing science and data faster. This study focuses on forecasting motor vehicle (MV) related deaths in the United States (US) to share forecasted data faster while examining how demographic factors—such as age, sex, race/ethnicity, and geographic location—affect these fatalities.

Methods

This dissertation consists of two studies. The first study utilizes new and improved forecasting procedures to identify highly accurate techniques that provide estimates, bridging the gap created by the lack of timely data. The second study employs two regression models to describe the associations between demographic factors and MV fatalities. The project utilizes the National Vital Statistics System (NVSS) motor vehicle mortality data from July 2020 to June 2024 (N = 175,600 deaths) in the US.

Results

The study achieves a high average forecast accuracy of 98% and provides predictions for 12 months (July 2024-June 2025) based on demographic factors. Among subgroups, the analysis reveals significantly higher reported MV deaths among males, by race white individuals, by urbanization in large central metro areas, by age within the age group of 25-44 years, and during August to October months.

Conclusions

The high forecast accuracy achieved through the Exponential Smoothing Model (ESM) demonstrates its potential to enhance the timeliness of the provisional data (by eliminating a one-year data lag) and help inform effective traffic safety strategies. The findings highlight the urgent need for targeted interventions to reduce MV fatalities and safeguard public health and safety. Targeted interventions should focus on young adult males aged 25–44, mainly white individuals in large central metro areas, with enhanced enforcement, education, and prevention campaigns intensified during the high-risk months of August to October, also improved road signage, better lighting, and changes in road layout.

INDEX WORDS: Motor vehicle accident; forecast; predict; trend; time series; demographic factors; negative binomial regression.

FORECASTING MOTOR VEHICLE FATALITIES:
EXAMINING THE IMPACT OF DEMOGRAPHIC FACTORS

by

TADESSE HAILEYESUS

M.S., University of Georgia, 2002

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PUBLIC HEALTH

ATHENS, GEORGIA

2025

FORECASTING MOTOR VEHICLE FATALITIES: EXAMINING THE IMPACT OF DEMOGRAPHIC FACTORS

by

TADESSE HAILEYESUS

Major Professor:	Daniel Jung
Committee:	Jessica Smith Carlos Siordia

Electronic Version Approved:

Ron Walcott
Vice Provost for Graduate Education and Dean of the Graduate School
The University of Georgia
May 2025

Dedication

To my loving parents, Tsegemariam and Haileyesus — thank you for planting in me the love of learning and the faith that continues to guide my path. Your wisdom and prayers have shaped who I am. To my caring wife, Saba (and the Focolare, her prayer group), and our incredible children, Adam, Omega, Nathan, Aman, and Liya — your love has been my anchor, your sacrifices, my strength, and your belief in me, my constant motivation. This successful journey is ours.

Acknowledgments

I am grateful to God, my parents, and my family for their unconditional love and support. I am also thankful to my dissertation committee, Daniel Jung, PhD (Chair); Jessica Cathey Smith, PhD; and Carlos Siordia, PhD, for their tireless guidance and passionate encouragement. I thank the University System of Georgia's DrPH faculty/staff, including Zhuo (Adam) Chen, PhD, MS, MMgt; Charles Crisman MLS., M.Ed., and the 2021 cohort students for their assistance and encouragement. I am also grateful to the many Colleagues: Judith Qualters, PhD, MPH; Peter Edwards, MPH, MPA; Amy Wolkin, DrPH; Marcie-jo Kresnow, MS; Mark Stevens, MS; Scott Kegler, PhD; Lee Annest, PhD; Daniel Whitaker, PhD; Alexander Crosby, MD, MPH; Tesfaye Bayleyegn, MD, MPH; and countless others; the Injury Prevention Center; NCHS; Applied Statistics Team, Data Science Team, and Transportation Softy Team for their technical assistance and support.

TABLE OF CONTENTS

Chapter 1: Background and Introduction

Motor vehicle (MV) deaths are a leading cause of death in the United States (US), killing over 100 people every day, the Centers for Disease Control and Prevention (CDC) (2025a). This study will explore forecasting techniques to produce estimates based on demographic factors (sex, age group, race, ethnicity, and urbanization level). Identifying the highest-performing forecasting technique can inform public health prevention efforts seeking to mediate MV fatality prevalence. The analysis explores how demographic factors were associated with the distribution of MV fatalities (Beck et al., 2017) to help inform prevention efforts in the US. Hence, this study analyzes historical data, using improved statistical models and forecasting techniques to predict future crashes (number of crashes by month) while considering the impact of various demographic factors specific to the US.

Problem Statement

This study assesses the following two interconnected problems. First, historically, publicly available CDC injury datasets are available for the public 2 years after data collection in the US (Ehlman et al., 2021). Similarly, final injury data are currently approximately two years old on average; WONDER (Wide-ranging Online Data for Epidemiologic Research) typically publishes tabulations on MV fatality deaths with a two-year time lag. The absence of near-real-time information presents significant obstacles to effective public health responses and limits the ability of public health officials to address urgent trends and emerging risk factors.

Second, traffic fatalities continue to pose a significant public health problem, with disparities observed across various demographic subgroups. For example, a study by Raifman and Choma (2022) revealed that the risk of being exposed to traffic fatalities varies among different racial and ethnic groups, as well as different modes of transportation. They also found that these observed disparities persist even considering only urban areas and seem to worsen in different conditions. Another study on sex-based disparities in crash risk echoes these findings (Cullen et al., 2021).

Predictive analytics based on historical data offer a way to estimate potential injury cases that may otherwise remain unreported for years. These advanced forecasting techniques enable the estimation of potential injury cases in near-real-time, facilitating proactive public health surveillance and allowing for earlier intervention.

Predicting the number of potential injuries accurately from a public health surveillance system plays a crucial role in various aspects of public health (German et al., 2001). Reliable and accurate estimates of injuries can help produce actionable insights to help guide prevention efforts, measure disease burden and identify populations at high risk, monitor trends including epidemics and pandemics, guide program planning and evaluation, detect changes in health practices, prioritize resource allocation, describe the clinical course of diseases, and provide a foundation for epidemiologic research (German et al., 2001). By analyzing population injury estimates, public health officials can make informed decisions to prevent and control diseases, protect vulnerable populations, allocate resources effectively, and contribute to scientific research to understand better and manage public health issues (German et al., 2001). For example, public awareness campaigns can educate the public about high-risk periods or areas, encouraging safer driving behaviors. By utilizing predictive analytics, stakeholders can proactively address potential issues before they result in fatal accidents, ultimately saving lives. Thus, by collaborating across both public and private sectors, the CDC can integrate tools, resources, and innovation, enhancing the science of disease forecasting to protect Americans' health, safety, and security (CDC, 2024a).

Scientific knowledge of MV-related fatalities has several limitations. For example, there are limited insights into how MV fatal prevalence varies between populations. This knowledge gap limits stakeholders' ability to develop prevention strategies seeking to reduce the frequency and severity of MV fatal deaths. Thus, the second study assesses disparities in crashes among different subgroups to enhance injury prevention in the US (particularly considering the potential impact of disruptions in healthcare delivery caused by the COVID-19 pandemic). The study analyzes the underlying contributing factors of MV fatalities and identifies geographic areas or population groups more prone to accidents and injuries (NHTSA, 2020; Raifman & Choma, 2022). Raifman and Choma (2022) analyzed traffic fatality disparities by race/ethnicity, including fatality rates during darkness and in urban areas.

In summary, this study utilizes a forecasting tool to bridge the gap caused by the lack of timely injury data. Investigations will enable researchers to forecast MV-related fatal deaths and understand how they contribute to fatal traffic accidents. Additionally, it helps describe significant factors contributing to the unequal distribution of fatalities, allowing for targeted interventions and potential solutions to be developed and assessed.

Background

MV deaths pose significant public health concerns in the US and worldwide. MV crashes are responsible for over 100 deaths daily, which are a leading cause of death in the US (CDC, 2024a; CDC, 2024b). In 2021, MV crashes resulted in more than 45,000 fatalities, led to 2.3 million emergency department visits for injuries, and incurred a total cost of $480 billion from crash-related deaths, including medical spending and the value of lost life years (CDC, 2024a);

additionally, deaths from crashes in 2022 resulted almost 44,000 people and in over $470 billion (CDC, 2024b). Similarly, road traffic fatalities remain a significant global public health concern; the WHO recorded approximately 1.19 million road traffic deaths in 2023 (WHO, 2024).

Context 1: Lack of Timely Injury Data

The delay in injury data presents a significant barrier to effective and practical intervention efforts for public health. Data collection poses significant challenges due to the need for extensive human effort/involvement, competing priorities, considerable time investment, and the growing demand for up-to-date information (Sohail et al., 2023).

The CDC published mortality data for the proposed study, originating from death certificates filled out by over 2,000 local coroners and medical examiner offices nationwide (Brooks, 2021). Due to the decentralization of vital records, the assembly of provisional mortality statistics requires approximately 7 to 8 months, with final national motor vehicle mortality statistics experiencing delays of over a year (Swedo et al., 2023).

Currently, publicly available CDC injury data are two years old on average. Historically, injury datasets are available for the public 3–4 years after data collection in the US (Ehlman et al., 2021). However, the world is moving fast with rapid technological advancements and fast-changing lifestyles. The CDC, as a public data steward (not its owner), recently conducted a "CDC Moving Forward" review, and one of the key recommendations was to *share science and data faster* (CDC, 2024a). CDC Moving Forward is an initiative that aims to enhance the CDC by improving its structure, processes, and systems to safeguard the nation's health, safety, and security. CDC promotes accountability, collaboration, communication, and timeliness. In response to emerging diseases like H1N1 and COVID-19 and ongoing public health issues such as cancer and injuries, the CDC launched a review in April 2022. The goal was to identify areas where systems and processes could be modernized for more effective sharing of scientific findings and data. One of the key areas for improvement identified under CDC Moving Forward was the need for faster sharing of scientific findings and data (CDC, 2024a).

The gap in the availability of the most recent injury final data is created due to the time it takes to collect, process, and release it because of thorough quality assurance that must go through before releasing statistical data and results to the public (Ahmad & Cisewski, 2023; Ehlman et al., 2021). Thus, the CDC is exploring expediting public data release by adding *provisional estimates* to improve wait times for data by further shortening the lag periods by approximately 6 months. According to the CDC, provisional data refers to preliminary estimates derived from death certificates received, processed, and coded but not finalized (Ahmad & Cisewski, 2023; Arias et al., 2021). Hence, the CDC releases provisional vital statistics data and the final data for conducting public health surveillance for annual national natality and mortality statistics.

Furthermore, Spencer and Ahmad (2017) noted that the availability of death certificate data for analysis varies depending on the cause. For instance, the delays in reporting deaths due to injuries were even more prolonged than those for deaths not related to injuries. When conducting mortality surveillance, it is crucial to consider the delay between the occurrence/date of a death and the availability of data from the death certificate for analysis.

Limited efforts and success have been made to tackle the delays in public data related to crashes. Therefore, the objectives of this study were to build upon the achievements of past efforts by exploring trends of time series data from several sources (Agyemang et al., 2023; Hao & Liu, 2024; Swedo et al., 2023; Zhang et al., 2007) to identify accurate forecast to share timely "predicted" data, which is currently unavailable, and bridging this gap. The first study assesses and identifies reliable statistical forecasting techniques from different perspectives. It also explores how this approach can effectively address the issue of data delays in public health prevention.

Context 2: Disparities in Traffic Fatalities

Studies in different countries have found disparities in traffic fatalities based on race, ethnicity, sex, and age across different modes of transportation and geographical areas (Cullen et al., 2021; Raifman & Choma, 2022). For example, a study by Cullen et al. (2021) acknowledged that young men are more affected by road crashes and injuries than young women and older drivers. However, there is limited understanding of how these sex differences change over time as individuals gain more driving experience. To address this, the researchers conducted a detailed analysis of over 20,000 young drivers in Australia. The researchers followed them for up to 13 years after they obtained their car driver's license to explore sex differences in crash and crash-related injury rates. They found that men had significantly higher rates of any-crash (1.25 times), single-vehicle crashes (2.07 times), and crashes in wet conditions (1.59 times) compared to women. However, men were less likely than women to be involved in crashes, resulting in hospitalization (0.73 times).

In summary, multiple studies have found disparities in traffic fatalities based on race, ethnicity, sex, and age across different modes of transportation and geographical areas, highlighting the need for further research to understand how these differences evolve.

Further, studies reflect similar patterns of inequality. From 1999 to 2019, the differences in motor vehicle traffic death rates among males by age group have decreased from 14.7 per 100,000 to 11.1 (Spencer et al., 2021). In 1999, males aged 15-24 had the highest rate (35.0), but by 2019, their rate was lower than that of males aged 25-64 (20.6) and 65 and over (21.8). Over the years, MV traffic death rates among females decreased across all age groups. Overall, MV death rates were increased from 2010 through 2019 from 10.7 to 11.1. The authors reported changes in MV traffic death rates among various demographic groups (i.e., by sex, age group, and type of road user). According to Raifman and Choma (2022), initial analyses indicate disparities in traffic fatality rates based on race/ethnicity. They reported that all-mode traffic fatalities were higher among American Indian and Black Americans compared to White Americans, while Asian Americans had the lowest fatality rates.

In summary, studies found patterns of inequality in MV traffic death rates, with differences decreasing among males by age group and overall rates increasing. At the same time, disparities based on race/ethnicity indicate higher fatality rates among American Indian and Black Americans compared to White Americans.

In early investigations, Storie (1977) explored the differences in driving behavior between male and female drivers. The author found significant differences in speed, skill, and attitude between the two sexes. Females tended to drive at lower speeds and overtake more cautiously. At the same time, males were generally more skilled, capable of performing difficult maneuvers, and more likely to take risks, such as driving under the influence of alcohol. However, no significant difference was found previously between male and female drivers in terms of being at fault in accidents. Similarly, age labels can encourage individuals to seek out age-stereotypic information about others. Nevertheless, there were mixed findings; on the other hand, Ng et al. (1991), in their study in New Zealand, showed that the search for age-stereotypic information across the lifespan was independent of the sex of drivers involved in a car accident or the sex of study participants and the participant's age.

In conclusion, early investigations have found disparities in driving behavior between sexes, with females tending to drive at lower speeds and take fewer risks, while males were generally more skilled and prone to risky behaviors.

Beck and coauthors (2017) highlighted the role of differing seatbelt usage rates between rural and urban areas, partially explaining the higher fatality rates in rural regions. In 2015, 50% of fatally injured occupants on rural roads were unrestrained, while the percentage was slightly lower at 46% for fatally injured occupants on urban roads. This geographic disparity underscores the importance of region-specific safety interventions and policies to mitigate risk factors in urban and rural settings.

Crashes at intersections account for the highest percentage of fatalities among older drivers (Samuel et al., 2016). Various reasons have been suggested for their increased risk at intersections, and it has been discovered that older drivers are significantly less likely to check for potential hazards after entering an intersection than middle-aged drivers. Due to the aging population, the increasing number of older drivers in the US (Lombardi et al., 2017) necessitates a better understanding of the factors contributing to their involvement in vehicle crashes. It is crucial to develop strategies and measures to decrease the frequency and severity of these crashes.

The above paragraph highlights that crashes at intersections pose a significant risk for older drivers, who are less likely to check for potential hazards after entering an intersection compared to middle-aged drivers, highlighting the need for a better understanding of the factors contributing to their involvement in vehicle crashes and the development of strategies to reduce their frequency and severity.

Hence, the second study examines the disparity in US fatal motor fatalities between population groups. It will analyze factors such as sex, age, race/ethnicity, and urbanization level to identify sub-populations with disproportionate risk rates. Traffic safety experts and other stakeholders can develop effective prevention strategies by identifying the variables strongly associated with motor vehicle accidents.

In summary, this study aims to explore and continue to build upon the strengths identified in the existing and novel forecasting methods to provide timely motor vehicle crash data. It achieves this by comparing the effectiveness of new and improved procedures to deliver more precise and reliable results. Moreover, this study describes how those contributing factors might affect motor vehicle-related fatalities, identifying areas and population groups more susceptible to accidents and injuries.

Study Purpose

The first study aims to identify highly accurate forecasting techniques to provide predictions based on demographic factors (sex, age group, race, ethnicity, and urbanization level). It describes new and improved forecasting procedures that bridge the gap caused by the lack of timely data. The second study describes how demographic factors are associated with the distribution of MV fatalities to improve population health in the US. Accordingly, this original research aims to contribute to the injury data release practice by identifying new and innovative perspectives and novel forecasting methods. It aligns with the CDC's dedication to public health, as it strives to "protect health and improve lives." This research assists in developing specific safety interventions and policies to reduce fatalities and address disparities among different population groups.

Delivering accurate "forecast data" offers valuable insights for the CDC to anticipate future trends, aiding decision-making and forward planning. The study will analyze and propose strategies to address data management challenges by implementing innovative analytical methods. MV crash prediction helps reduce data barriers and prevention efforts. Thus, the study aims to showcase the efficiency of this novel method in analyzing and disseminating data more promptly.

Specifically, study one used the most recent finalized and provisional fatal MV data (i.e., from July 2020 to June 2024) and forecasted crashes for the next 12 months (July 2024 to June 2025) in the US (Appendix 1). This study aims to add a significant piece that fills the gap specifically to the lack of these timely data by identifying more accurate forecasting techniques (i.e., by delivering 12 months of future data). Accurate forecasting enables the faster sharing of predicted data, which is currently unavailable. It provides new insights into little-understood phenomena. The findings will improve the lack of timeliness by expediting the release of injury data for prevention and decision-making.

Consequently, the second study seeks to improve injury prevention in the US by assessing crashes among various subgroups. It examines factors contributing to the unequal distribution of motor vehicle-related fatalities, analyzing underlying contributing factors and identifying areas or population groups more susceptible to accidents and injuries (NHTSA, 2020; Raifman & Choma, 2022). The NHTSA (2020) report includes a special analysis of fatalities and fatality rates per 100 million MV crashes in various sub-categories. For example, the NHTSA (2020) analysis includes sub-categories such as rural/urban interstates, weekend crashes, and fatalities by sex and age groups. Raifman and Choma (2022) conducted a study that developed exposure-based estimates of fatalities per mile traveled for pedestrians, cyclists, vehicle occupants, and by race/ethnicity, including an analysis of rates during darkness and in urban areas.

Hence, this study seeks to describe further and expand upon the strengths identified in previous studies by identifying factors such as sex, age, race/ethnicity, and urbanization level. Consequently, this analysis will consider these demographic factors as covariates when describing risks for subpopulations.

Significance of the Study

The significance of the study lies in its potential to enhance understanding of health status and inform timely forecasts, which can lead to targeted safety interventions that positively impact society. It specifically addresses the problem of delays in obtaining crash data, which can hinder timely interventions to prevent these fatalities. Thus, the first study aims to improve current methods of releasing estimates by exploring more accurate forecasting techniques and facilitating the sharing of predicted data.

Next, in the long run, the study seeks to add promising predictive models by incorporating additional injury mechanisms like falls, suicide, self-harm, and drowning into injury data-sharing methods. Finally, the study findings can provide a baseline to validate/verify outputs and measurements by comparing forecasts generated using SAS software with other software, programming languages, methods, and technologies, e.g., the new generative artificial intelligence (AI-generated) technology. In advancing public health, the CDC started harnessing the power of AI to forecast disease outbreaks. Advances continue to make conducting prediction more efficient, and the CDC uses data from the dashboard to drive data-informed decision approaches to improve prevention and prioritize actions (e.g., to create new content, such as text or codes, by learning from existing data).

Similarly, the significance of the second study lies in its potential to enhance understanding of health status and quantify the relationship between the study variables. By analyzing these variables, the study seeks to identify the key predictive factors that influence motor vehicle deaths. This understanding can help stakeholders develop effective prevention strategies by targeting specific populations at higher risk.

Timely and accurate injury data aids in framing effective and impactful prevention measures. It contributes to the CDC’s commitment to protect health and improve lives. Effective prevention requires awareness and foresight. To prevent diseases, injuries, or other health challenges, one must first understand and anticipate them. Finding solutions tailored to the organization's goals involves monitoring trends, conducting research, or using data to identify risks early on, such as predicting potential health threats or recognizing patterns that indicate emerging risks. For example, crash detection systems played a crucial role in minimizing casualties resulting from vehicle accidents by reducing emergency service response time (Thompson et al., 2010).

Accordingly, this study reflects various perspectives needed to offer alternative and more efficient tools to meet faster data-sharing challenges. The data and insights quickly obtained from the prediction model can be utilized as follows:

1. To increase public service announcements by state and local transportation departments to prevent a predicted increase in crashes before they occur (e.g., to raise public awareness and encourage safer driving behaviors), i.e., enabling organizations and individuals to take proactive measures to prevent or mitigate these issues before they occur/escalate.

2. **To improve road safety** by identifying high-risk areas, which allows for targeted interventions, such as improved road signage, better lighting, and changes in road layout.
3. **Communicate injury forecasts** to health service providers to assist in planning for emergency medical services preparedness, such as optimizing their resources and response times and prioritizing interventions for optimal allocation of resources. Moreover, it contributes to research communities and societies by providing the latest estimate data (Table 3).

The CDC informs public health decision-makers on the potential consequences of deploying control measures, collaborates with public, private, and government partners to realize solutions, and supports innovation to improve the science of outbreak analytics and modeling continuously. Releasing data that reflects occurrences in communities nationwide empowers individuals, families, and communities to make informed decisions on the best ways to safeguard themselves and others.

Consequently, the end goals or outcomes expected to result from the proposed study include:

- *Share data faster* by implementing the optimal predictive methods identified.
- Identify the disparities and significant demographic factors/variables that lead to motor vehicle deaths to improve population health.

Furthermore, the study provides the CDC with new, additional hands-on forecasting experience to contribute technical support and assistance. The study will pave the way for collaboration with the CDC/transportation team and others to investigate (the most effective model for making predictions and sharing science and data swiftly, as well as decision-making). For instance, as noted by Swedo et al. (2023), injury forecasts can expedite the response of public health practitioners and other stakeholders to unforeseen changes.

Hence, by accurately forecasting where the highest risks occur, stakeholders can design more impactful interventions, allocate resources more efficiently, and save lives by tailoring safety measures accordingly. Overall, forecasts help decision-makers implement safety programs to reduce MV fatalities (they also generate actionable insights). Specifically, timely data will support the CDC and Injury Center's goals to prevent injuries and reduce their consequences. Effective prevention strategies for motor vehicle crashes focus on addressing the risk and protective factors associated with these incidents. By doing so, the study aims to improve population health (quality of life).

Delimitations

The delimitations or boundaries of this study include:

- Data Source: National Center for Health Statistics (NCHS), National Vital Statistics System (NVSS), CDC/WONDER (Wide-ranging Online Data for Epidemiologic Research) mortality files linked with death certificate, Jul 2020-Dec 2022 final data, Jan 2023-Jun 2024 provisional data (Appendix 1).
- The study population contains mortality counts for all US county residents by sex, age group, race/ethnicity, and urbanization level.
- It includes Injury Intent: Unintentional Injuries
- Injury Mechanism and All Other Leading Causes: Motor Vehicle Traffic

Assumptions

The following were the assumptions of this study:

- The MV death certificate data reflects the mortality counts for all US counties.
- The analysis variables in the death certificate capture major factors.
- The measurements in the historical data were coded/measured correctly.

Definition of Terms

This subsection defines relevant terms to clarify their meaning in this study's context.

Motor Vehicle Fatalities refers to deaths resulting from motor vehicle accidents.

Demographic factors refer to a subgroup of people within the US population (e.g., by age group, sex, race/ethnicity, urbanization level).

A time series is a sequence of data points for analyzing trends or forecasting (e.g., monthly counts).

Temporal refers to the time dimension in which data points are organized to study patterns, trends, and dependencies over time.

A trend is an overall direction of time series data over a period, which can be classified as increasing (upward), decreasing (downward), or horizontal (stationary).

Forecasting Models are statistical procedures used to predict data based on historical data.

A stochastic process is a statistical method that describes an evolving sequence of random variables over time.

Auto-Regressive Integrated Moving Average (ARIMA) is a classical statistical modeling technique for forecasting time series data.

The Exponential Smoothing Model (ESM) is a procedure that generates forecasts using optimized smoothing weights for time series.

Mean Absolute Percentage Error (MAPE) measures forecasting accuracy where a lower MAPE is desirable (e.g., 0% MAPE means a perfect fit with no error).

Poisson Regression is a statistical method used to model "count data" by assuming that the counts follow a Poisson distribution (has a positive mean and equals the variance).

Negative Binomial Regression is a statistical method used to model "count data" that allows for overdispersion (where the variance exceeds the mean) by extending the Poisson regression model, incorporating an additional parameter to account for the count variability.

Forecasting vs. Prediction

The key distinction between forecasting and prediction lies in their focus and methodology: forecasting estimates explicitly future values using historical time series data, where each observation is sequentially dependent, while prediction encompasses estimating outcomes for unseen data without temporal constraints. Prediction involves fitting a model to a training dataset to create an estimator for new samples. In contrast, forecasting explicitly predicts future events based on the influence of historical data points. Thus, forecasting can be viewed as a subset of prediction that emphasizes the sequential nature of time-related data (Hyndman & Athanasopoulos, 2018).

Consequently, this study investigates two primary research questions:

Research Question 1: Which forecasting model predicts motor vehicle fatalities in the US with higher accuracy for effective prevention and decision-making?

Research Question 2: Is there a difference in motor vehicle fatalities among population groups in the US?

Outline of the Dissertation Proposal

The remainder of this dissertation proposal is organized into four chapters and appendices: Chapter 2 presents a literature review that introduces the topic and examines relevant literature on evolving trends in injury forecasting practices while highlighting demographic disparities; Chapter 3, the first study, introduces new and improved statistical forecasting methods, detailing the research question, methods, results, human subjects review, limitations, conclusions, and recommendations for future research; Chapter 4 focuses on the second study which describes the impacts of demographic factors in crashes through detailed regression analysis and summarizes the discussions and conclusions; finally, Chapter 5 provides a combined discussion to address both research questions presented in the previous chapters, concluding with the summary of findings, limitations faced during the study, and recommendations for future research.

Chapter 2: Literature Review

Introduction

Quality data are at the center of everything a scientist does. In 2022, the CDC conducted a "CDC Moving Forward" review, and one of the key recommendations was to share scientific findings and data faster (CDC, 2024a). It enables public, private, and government partners and the CDC to fulfill their commitment to "protect health and improve lives." Thus, the first study addresses this need for timely injury data using novel forecasting methods. The second study examines disparity by demographic factors, such as age, sex, race/ethnicity, and geographical location in MV injury in the US. This study examines how demographic factors contribute to higher injury risk rates and identifies key predictive factors influencing car crashes. Understanding these variables strongly correlated with MV accidents enables researchers and stakeholders to develop effective preventions.

This literature review provides a background for the essential concepts and discusses existing studies relevant to the proposed study. It will cover major areas such as determinism vs. randomness of crashes, causes of road accidents, driver behaviors, and factors influencing crashes. It will draw from in-depth research reviews in the field, journal articles, and scientific presentations previously published. The final section of the study focuses on developing forecasting models, discussing the various factors that influence car crashes, and comparing them to other related published research.

Overview of Published Research

"Motor Vehicle Fatalities" pertains to deaths resulting from motor vehicle accidents. Despite advancements made in recent decades, traffic fatalities continue to pose a significant public health challenge (Raifman & Choma, 2022). In a recent study by Raifman and Choma (2022), exposure-based estimates of fatalities per mile traveled were developed for pedestrians, cyclists, and light-duty vehicle occupants in America.

From 2006 to 2010, MV traffic death rates gradually declined from 14.5 to 10.7 per 100,000 population, followed by noticeable increases in 2010 and 2019 (11.1) in the US (Spencer et al., 2021), and distracted driving may have contributed to these changes (NHTSA, 2024). Despite some progress in the past decades, road traffic fatalities remain a significant global public health concern.

Men have higher rates of car crashes than women, except for hospitalization cases, where men were less likely to be hospitalized compared to women (0.73), according to a study analyzing over 20,000 young drivers in Australia (Cullen et al., 2021). Older adult drivers account for the highest annualized fatal intersection crash rates (Lombardi et al., 2017; West & Naumann, 2013), while American Indians/Alaska Natives, as well as older adults, have had the highest rates of traffic-related pedestrian deaths (Naumann & Beck, 2013). Intersections also had the highest proportion of fatalities among older drivers compared to other types of crashes (Samuel et al., 2016).

Moreover, a study in Brazil revealed that more men, especially young individuals, were exposed to and died in traffic accidents compared to women (Medeiros & Nadanovsky, 2016); likewise, cell phone use while driving among adolescents, particularly newly licensed drivers poses a public health threat as it is a common behavior often occurring at high speeds

(McDonald et al., 2019). Furthermore, factors related to crashes differed significantly between younger and older drivers, including time of day, weather conditions, roadway type, and driver fault (Cullen et al., 2021; Lombardi et al., 2017).

The impact of urbanization on road safety is mixed, as urban college settings can pose risks due to factors like lack of signage and faded crosswalks, as well as jaywalking (Pollack et al., 2014). On the other hand, Beck et al. (2017) found evidence linking rural areas in the US to higher crash-related death rates and lower seat belt usage.

Shaw et al. (2022) also found higher crash rates among non-Hispanic Black and American Indian or Alaska Native (AI/AN) children in rural areas, while Asian or Pacific Islander children had the lowest rates, especially in the most rural counties in the US. A study examined the disparity of Latinos being overrepresented in alcohol-related crashes concluded that all drivers face similar crash risks based on blood-alcohol levels, indicating that factors such as consumption patterns, driving exposure, awareness of driving rules, and socioeconomics may explain the higher representation of Latino drivers in alcohol-related crashes (Orres et al., 2014).

Randomness vs Determinism of Crashes

The debate between randomness versus determinism in MV crashes revolves around the unpredictable nature of accidents vs. the belief that every event results from preceding or *historical random* causes and is theoretically predictable (e.g., utilizing statistical models for forecasting time series data). On the other hand, *deterministic* models are mathematical models that yield a predetermined outcome based on specific inputs, disregarding randomness/chance. Deterministic models are commonly employed for systems with clearly defined dynamics and predictable behavior. Deterministic models may appear to be as effective as forecasting models but turn out not to be. A random sequence can be directly calculated from the observed data if the parameters are precisely identified (Box & Pierce, 1970; Box et al., 2015).

A study by Fridstrøm and colleagues (1995) analyzed monthly road accidents by randomness, exposure, weather, daylight, or changing reporting routines and speed limits in Denmark, Finland, Norway, and Sweden. They found that randomness and exposure explained 80% to 90% of the observable variation in fatal accidents. Where the number of accidents was low, they found that the randomness or chance events (e.g., unusual weather conditions) can cause more changes in the data from one period to another. However, traffic volume was considered a significant predictor of crashes. They concluded that significant decreases in accidents could not be realized without reducing traffic volume, the most crucial systematic factor. In other words, higher traffic volumes increase the likelihood of vehicle interaction, thereby raising the probability of collisions (i.e., crashes are predictable). For example, in the summer, traffic volumes and motor vehicle fatalities rise (Faust et al., 2021; Teymuri et al., 2014; DOT, 2024a).

On the contrary, a study by Gao and Davis (2017) stated that driver "Distraction Duration" was "the primary direct cause of the increase in reaction time, with other factors having indirect effects." The study was based on a small sample (n = 103 events, five crashes, and 98 "near-crash") and suggested road accidents as deterministic occurrences. However, the critiques of this study and its limitations include its small sample size and the exact mechanism of how distraction affects reaction time, which still requires future efforts.

Causes of Road Traffic Accidents

Several studies attributed the causes of road traffic accidents to major factors such as over-speeding, distracted driving, drunk driving, improper lane changing, and the poor condition of roads and vehicles (Ko, 2019; Tasca, 2000; NHTSA, 2020; DOT, 2024b; Wegman et al., 2017). According to a study conducted by NHTSA (2020), approximately 37 individuals lose their lives in drunk-driving accidents every day in America. Furthermore, a recent systematic review found notable causes of crashes and road safety issues, such as risky road surface conditions, risky or reckless road users (drivers, pedestrians), and traffic conditions (Sohail et al., 2023). Zsifkovits and Pham (2017) provided an overview of the techniques used to identify pedestrian behavior in public spaces, while Ridel et al. (2018) concentrated on predicting pedestrian behavior in urban settings.

In addition to the impact of traffic flow and congestion on road safety, several research papers have also argued the importance of studying and predicting traffic flow and congestion. For example, Akhtar and Moridpour (2021) specifically explored research on congestion prediction using artificial intelligence methods. Hossain et al. (2019) systematically reviewed real-time crash prediction methodologies. Additionally, Nagy and Simon (2018) reviewed data-driven techniques for predicting traffic flow in smart cities.

Driver Behavior

Driving behavior refers to the habits or actions of motorists while driving, such as turning the steering wheel, accelerating, or decelerating (Higgs & Abbas, 2014; Miyajima et al., 2007), and the amount of space they maintain between their car and the vehicle ahead to ensure a safe driving distance (Miyajima et al., 2007). Similarly, some external factors may also influence motorist's behavior, such as driver's attitude, mood, level of fatigue, demographic background (e.g., age and sex), and environmental parameters, e.g., visibility, weather, and traffic (Martinez et al., 2017).

Beyond driver behavior, some road safety implications are related to other road users, such as pedestrians, autonomous vehicles, and vehicles/bicycles (Haileyesus et al., 2007; Jahangiri et al., 2015; Papadoulis et al., 2019).

Several studies divided driving behavior into two broader categories: (1) inattentive driving and (2) aggressive or reckless driving (Chan et al., 2019; Iio et al., 2021; Tasca, 2000; Jahangiri et al., 2016). Distracted driving occurs when drivers shift their attention away from driving to other activities, and this behavior has been widely associated with fatal crashes (Alnawmasi & Mannering, 2022; NHTSA, 2024).

Correspondingly, aggressive and reckless driving habits include speeding, improper lane changing, and tailgating. Studies have highlighted driving behavior as a primary cause of traffic crashes e.g., the US Department of Transportation (DOT) reported that "driving while drowsy results in a four- to six-times higher near-crash/crash risk relative to alert drivers" (Klauer et al., 2006, p. 1).

Demographic Factors

Demographic factors related to crashes for the second study include age, sex, race/ethnicity, and urbanization. These factors have been found to influence the likelihood of being involved in a crash or experiencing more severe outcomes (Cullen et al., 2021; Raifman & Choma, 2022). For example, younger drivers are often associated with higher crashes due to inexperience and risk-taking behaviors. Sex differences may also play a role, with males typically having higher crash involvement rates compared to females. Additionally, disparities in crash rates have been observed among different racial/ethnic groups and urbanization. The CDC Multiple Cause of Death database contains mortality and population counts for all US counties (CDC, 2025a). These data are based on death certificates for US residents by a single underlying cause of death and demographic data. The data provides a number of crash deaths by age, race categories, sex, month, year, and urbanization categories.

Development of Forecasting Models

Several studies have used different forecasting models and approaches to predict the occurrence of crashes worldwide. This subsection provides a background for the study variables and concepts involved and discusses existing studies relevant to the proposed research in detail. The discussion was organized chronologically, aiming to track the application and development of forecasting models over time in various countries. It also highlights the main criticisms raised, what is currently known, and what remains unknown.

In the early 1970s, Box and Jenkins' landmark research introduced autoregressive integrated moving average (ARIMA) models, which are a class of statistical procedures used for forecasting equally spaced time series data (Bheemanna & Budihal, 2023; Box & Jenkins, 1970; Box & Pierce, 1970; Box et al., 2015). Since then, the ARIMA models, also known as the Box-Jenkins methods, a classic and influential text in the field, have identified the most suitable fit for a time series model based on past time series values.

Soon after, in the mid-1970s, the Statistical Analysis System (SAS) included ARIMA and auto-regressive moving average (ARMA) modeling in SAS software. Subsequently, in 1979, Dickey and Fuller introduced the Dickey-Fuller formal diagnostic tests to determine whether a time series has a unit root, indicating it is nonstationary and possesses a trend (Dickey & Fuller, 1979; Ihueze & Onwurah, 2018; Nassiri et al., 2023; Ren et al., 2013).

Over the years, ARIMA models have been refined and are widely used in time series, econometrics, and other analyses to make future predictions based on past patterns. Since the early 2000s, researchers have employed ARIMA models in several time series analyses (Andreoni & Postorino, 2006; Zhang et al., 2007; Wei, 2013; Wei, 2018).

A cross-sectional study in Tehran using data between 2010 and 2011 examined traffic accidents and forecasted daily accidents (Teymuri et al., 2014). The authors used Minitab software and an ARMA time-series model for their forecast. They reported that the number of accidents in Tehran displayed a seasonal trend, with more accidents occurring in the summer. However, this paper studied seasonal trends using only two years of daily data, which was challenging because it limits the ability to capture and analyze recurring patterns that may vary year by year.

On the other hand, Doğan and Akgüngör (2013) conducted a study on forecasting highway casualties in Turkey under the railway development policy. They utilized annual data from 1980 to 2006 (n = 27 annual data points). The first 23 years of data were used for training, and the forecast was made until 2020 (Doğan & Akgüngör, 2013). The training data was used to forecast future death counts and assess model performance by comparing forecasts with actual counts. The study employed nonlinear multiple regression (NLMR) and artificial neural network (ANN) approaches to predict highway casualties. Road accidents and injuries were forecasted until 2020 using two scenarios: maintaining current transport trends or shifting passenger traffic from highways to railways while keeping air traffic unchanged. One of the main criticisms of studies such as Doğan and Akgüngör (2013) was that they use a relatively small number of data points to make predictions for distant future periods, as this approach increases the risk of inaccuracies and uncertainties.

In another study for a non-transportation domain by Ren et al. (2013), coauthors aimed to forecast the incidence of hepatitis E in Shanghai, China. The model was trained with monthly data from 2000 to 2011 and validated with data from 2012. To predict the monthly data for 2013, the researchers used SPSS and MATLAB software. They reported a time series analysis that revealed a seasonal pattern that allowed for accurate predictions.

Similarly, in another non-transportation domain, Jerrhag et al. (2017) conducted a study on forearm fractures in Skåne, Sweden, between 1999 and 2010, focusing on analyzing fractures by sex and projecting future injury rates. The researchers found that anticipated demographic changes, including a projected population increase of 25%, could result in a significant 38% rise in forearm fractures by the year 2050 compared to 2017.

Dickey (2019) demonstrated regression using the SAS procedures AUTOREG and ARIMA time series for forecasting purposes. The author examined issues related to utilizing these procedures and compared their features. The emphasis was placed on understanding when to use each procedure, interpreting the results, and employing diagnostics to enhance the model. The conclusion highlighted that AUTOREG and ARIMA offer time series practitioners the ability to incorporate predictor variables and autocorrelation for forecasting purposes (Dickey, 2019). The author discussed that autocorrelation was influential in the short term, but the input data determined forecasts further into the future. Thus, deterministic predictors have the advantage of known future values. Furthermore, advantages of ARIMA include incorporating moving average terms, differencing, and outlier detection capabilities (Brocklebank et al., 2018; Dickey, 2019).

A recent study by Inada et al. (2020) conducted a secondary analysis of country-level data from 1990 to 2017 to predict deaths caused by road traffic injuries worldwide until 2030. The study utilized forecasting models and projected a higher number of deaths (5.1% higher) than previous estimates, highlighting the pressing need for global efforts to prevent accidents and decrease fatalities to meet road safety objectives.

In a study conducted by Pandey et al. (2020), disease patterns and projections were analyzed to aid in health intervention planning in Nepal. The researchers utilized public data from the Global Burden of Disease database from 1990 to 2016, projecting mortality rates and disease risk factors until 2040. Similarly, Lazarus et al. (2022) conducted a recent study focusing on Spain's mortality, risk factors, and disease burden. Their analysis included updated estimates from 1990 to 2019 and projections up to 2030. Both studies emphasized the importance of appropriate responses and monitoring trends, particularly considering the ongoing COVID-19 pandemic, to effectively address the issue and reduce future disease burden.

Hao and Liu (2024) discussed the limitations of Transformer-based approaches for long-term multivariate time series forecasting, highlighting their struggle to capture complex interdependencies and utilize temporal features effectively. To address this, they developed a seasonal-trend decomposition-based 2-dimensional temporal convolution dense network that outperformed existing methods in forecasting accuracy. The model's efficacy was validated through experiments on six datasets, showing an average reduction in mean square error (MSE) of 3.2% to 13.8%. However, the authors acknowledged challenges in multivariate time series forecasting and noted limitations in their model's decomposition process.

In recent studies by Bheemanna and Budihal (2023) and Deretić et al. (2022), ARIMA models were employed for forecasting traffic crashes, focusing on the sustainability of traffic systems, and predicting accidents. Deretić et al. specifically utilized a Seasonal Autoregressive Integrated Moving Average (SARIMA) model to forecast traffic accidents in Belgrade, achieving accurate predictions with a high Mean Absolute Percentage Error (MAPE) of 5.22%

(i.e., accuracy of 94.78%). Likewise, several recent studies have employed the SARIMA model in their analyses (Agyemang et al., 2023; Gultekin & Acik Kemaloglu, 2023; Nassiri et al., 2023; Swedo et al., 2023).

A study by Bheemanna and Budihal (2023) focused on forecasting three demographic variables in India: Life Expectancy at Birth, Infant Mortality Rate (IMR), and Death Rate. The study utilized annual data from 1971 to 2020 and predicted 2021 to 2030. By employing the ARIMA model, the researchers identified the best-fitting models for each variable. The authors used Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC) to find the best-fitting ARIMA model, and they compared the variables based on Mean Absolute Percentage Error (MAPE) values. The study highlighted the robustness of ARIMA in forecasting demographic trends. It revealed improvements in life expectancy, declining death rates, and IMR in India, indicating advancements in public health and medical care. The findings emphasized the significance of these forecasts for healthcare providers and researchers in planning public health programs. The high predictive accuracy of the ARIMA model confirmed its usefulness as a tool for informed decision-making regarding future healthcare needs and societal changes due to demographic shifts. However, it is essential to note that one limitation of the study was its reliance on annual data rather than monthly data.

Summary of the Literature Review

This review synthesized existing studies on forecasting models in time series analysis and their potential application in predicting motor vehicle fatalities. The research aims to contribute to injury data release practices by identifying accurate forecasting models and investigating disparities in crash forecasting. Additionally, the study intends to identify the influence of demographic factors such as age, sex, race, and urbanization/geographical location on the occurrence of these fatalities.

In summary, the study emphasizes the urgent need for action to prevent crashes and reduce fatalities. Thus, this leads to research questions that help identify accurate forecasting models for motor vehicle fatalities and identify demographic factors in the US for effective prevention and decision-making.

Hence, the first study focuses on exploring and comparing robust statistical forecasting models to address the objectives of accurately predicting motor vehicle fatalities. In contrast, the second study describes disparities in fatal motor vehicle crashes among different population groups and offers recommendations for reducing these incidents.

In conclusion, the conceptual framework serves as a pivotal foundation for developing the forecasting model by elucidating the hypothesized relationships between the three main stages of the study and their anticipated outcomes. By systematically mapping out these connections, we can better understand how each stage influences the overall forecasting process. This framework guides the methodological approach and informs the selection of variables and metrics essential for accurate predictions. As we transition into the final section, we will delve deeper into how this theoretical underpinning translates into practical applications within our forecasting model, highlighting specific methodologies employed to operationalize these relationships and ultimately enhance predictive accuracy.

Conceptual Framework

The first conceptual/theoretical framework illustrates the hypothesized relationships between the three main steps/stages and the outcomes of the first study (Figure 1). *Model Identification:* The initial step in constructing a model involves assessing the stationarity of the time series data using a formal test and visual inspections. *Step 2. Model Estimation and Validation (Diagnostic Checking):* The next step was to check the diagnostic statistics to see if the model was adequate. The additive and multiplicative models were the two main models for the relationship between the observed series and unobserved components (Brocklebank et al., 2018), and their forecasts were compared next. *Step 3. Model Applying (Forecasting Stage)*: The best models were identified and used to generate forecasts. The accuracy of each forecast was then compared using the Mean Absolute Percentage Errors (MAPEs). If the error is small, generate a forecast; otherwise, repeat the process. In conclusion, the expected outcomes were health status crash forecasts by demographic variables generated by identifying a robust forecasting procedure, as depicted in Figure 1, Conceptual framework.

Study 1. Conceptual framework.

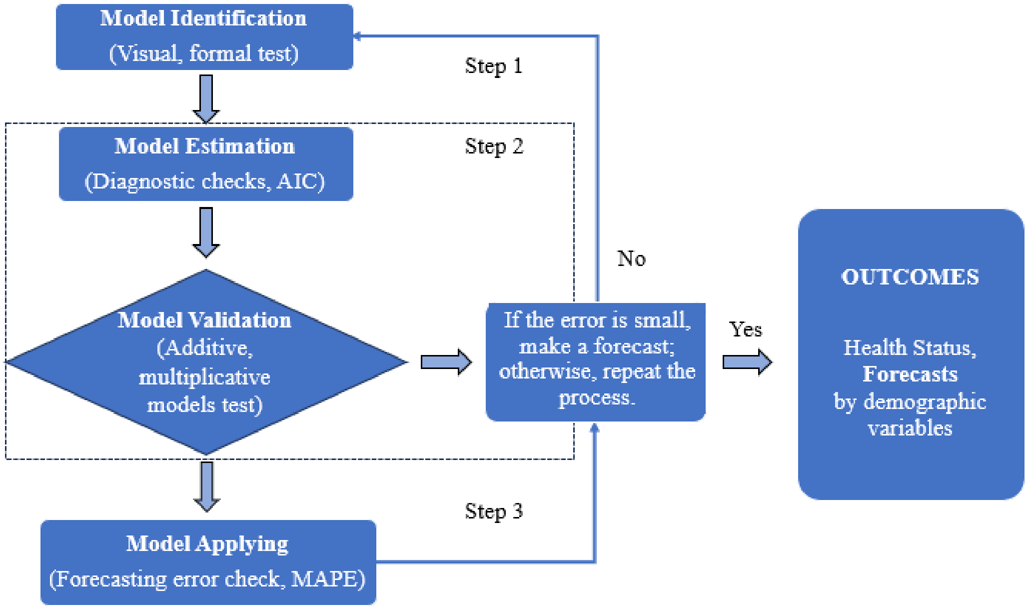


Figure 1: Conceptual framework for the first study adapted from Ismail et al. (2021).

Abbreviations: Akaike Information Criterion (AIC), Mean Absolute Percentage Error (MAPE).

The second study employs regression analysis and aims to describe and quantify the relationship between dependent and independent variables. Dependent (responses or outcome) variables are the variables that are being predicted or explained by the independent variables. Meanwhile, independent (predictor, exposure, input, or explanatory) variables are the variables that are hypothesized to affect the dependent variable and are used to predict or explain its variation in a regression analysis. Potential confounding will be examined by including candidate confounders one at a time in the model.

Dependent Variable: Number of crashes by month.

Independent Variables: sex (Male, Female), age (10 Age groups/1 to 90 years), race (American Indian or Alaska Native, Asian, Black or African American, Native Hawaiian or Other Pacific Islander, White, More than one race), ethnicity (Hispanic or Latino, Not Hispanic or Latino), and urbanization level (Large Central Metro, Large Fringe Metro, Medium Metro, Small Metro, Micropolitan/Non-metro (Micro), Non-core/Non-metro).

The rationale for focusing specifically on these factors in the second study was based on their repeated prioritization over others in examining fatal crash outcomes in several literature sources, as discussed previously. Additionally, the availability of the most up-to-date data through the CDC further supports the decision to include these variables. Hence, the second study analyzes how age, sex, race, and urbanization affect the number of crashes by month in the population over time, and the expected outcomes were health status and improved population health, as depicted in Figure 2, Conceptual framework.

Study 2. Conceptual framework.

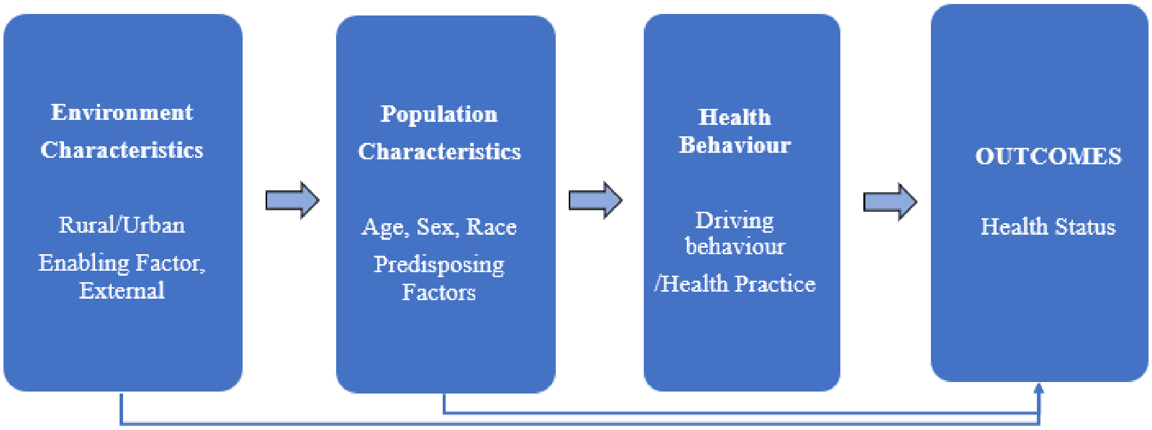


Figure 2: Conceptual framework for the second study adapted from Andersen's behavior model (Guilcher et al., 2012; Neupane et al., 2021).

Chapter 3: Forecasting Motor Vehicle Fatalities in the United States

Introduction

Motor vehicle (MV) crashes remain a major public health issue globally and in the US. According to the Centers for Disease Control and Prevention (CDC) (2025a), MV crashes have claimed the lives of over 100 people daily and are ranked as a leading cause of death in the US. The CDC collects mortality data nationwide and recommends sharing science and data more rapidly; however, the decentralized nature of vital records results in delays, with provisional statistics taking about 7 to 8 months and final national statistics potentially exceeding a year (Brooks, 2021; Swedo et al., 2023). Additionally, reporting delays for injury-related deaths exceed those for non-injury deaths (Spencer & Ahmad, 2017; Ehlman et al., 2021). These reporting delays hinder timely decision-making, limiting public health officials' ability to implement evidence-based interventions promptly.

Given the urgent need for timely insights into traffic-related fatalities, researchers have explored advanced statistical models for forecasting trends and mitigating data lags. Among these methodologies, the autoregressive integrated moving average (ARIMA) modeling has been recognized as one of the robust techniques for univariate forecasting (Brocklebank et al., 2018). However, a recent study indicates that the Exponential Smoothing Models (ESM) offer improved functionality compared to ARIMA (Brocklebank et al., 2018). These evolving methodological approaches underscore the importance of leveraging more accurate and responsive forecasting techniques to understand better and address traffic fatalities. Like ARIMA, the ESM is a forecasting technique that predicts future values by analyzing past data. However, ESM assigns greater importance to more recent observations, allowing the model to "smooth" the data and reveal underlying trends and patterns. For example, suppose one estimates park attendance for next week. In that case, they consider recent visitor numbers more heavily than those from earlier weeks, leading to a more informed prediction based on current behavior.

Despite ongoing efforts to address delays in public data related to traffic crashes, significant gaps remain in our understanding of how these incidents vary across different demographic subgroups. Existing literature highlights disparities in traffic fatalities based on race, ethnicity, sex, and age (Cullen et al., 2021; Raifman & Choma, 2022), yet there is a need for more in-depth analyses that consider how these factors interact over time. For instance, while studies have established that young men are disproportionately affected by road crashes compared to their female counterparts (Cullen et al., 2021), the dynamics of these differences as individuals gain driving experience remain underexplored.

Addressing these persistent gaps in publicly available data on traffic crashes is crucial for understanding the complex interplay of demographic factors over time. This understanding can ultimately inform more effective safety interventions and strategies. Targeted safety programs have shown promise in addressing specific behaviors, such as seat belt usage among construction workers and specific communities (Beck et al., 2019; Boal et al., 2016). However, the effectiveness of these programs could be significantly enhanced with more granular, real-time data that accounts for demographic variations. With timely and accurate predictions stratified by subgroups, stakeholders will be better equipped to implement effective strategies tailored to high-risk populations.

Therefore, this study focuses on bridging this gap by leveraging time series data to develop reliable statistical forecasting techniques. Doing so, it aims to provide accurate predictions stratified by subgroups that enhance our understanding of traffic crash trends and inform better prevention strategies.

Methods

Study design: This study is both predictive and descriptive. Hence, a quantitative correlational research approach is presented to examine relationships between variables without manipulating them, thereby gaining insights into the strength and nature of these relationships. The research design investigates the research question related to identifying a precise predictive model for motor vehicle deaths. Consequently, the study results can inform road safety measures and prevention strategies.

Data sources: The study utilized the CDC’s National Vital Statistics System (NVSS) multiple cause-of-death motor vehicle mortality data. The CDC collects death certificate data from vital statistics offices in all 50 states and the District of Columbia. These public data are provided by the National Center for Health Statistics (NCHS) CDC/WONDER (Wide-ranging Online Data for Epidemiologic Research). The query parameters included underlying cause of death (UCD) - Injury Intent: “Unintentional”; UCD – “Injury Mechanism & All Other Leading Causes”; Motor Vehicle Traffic. The study data were from July 2020 to December 2022 (finalized data) and from January 2023 to June 2024 (provisional data), a total of 48 monthly data (CDC, 2025a). CDC/WISQARS (Web-based Injury Statistics Query and Reporting System) also began providing these provisional data (CDC, 2025b). The study utilized monthly summary motor vehicle death count data rather than individual-level data (i.e., microdata) to leverage the most up-to-date information available information. These data are nationally comprehensive (not a sample) (CDC, 2025a; Xu et al., 2021), ensuring that the study's findings can be generalizable to the entire US population.

Dependent Variable

The dependent variable is the predicted monthly number of motor vehicle deaths, derived from a robust forecasting procedure and stratified by demographic factors.

Independent Variables

The independent variables include several demographic factors. First, sex is categorized into two groups: male and female. Race is classified into six distinct categories: American Indian or Alaska Native, Asian, Black or African American, Native Hawaiian or Other Pacific Islander, White, and individuals identifying as more than one race. Additionally, Hispanic origin or ethnicity is divided into two groups: Hispanic or Latino and Not Hispanic or Latino; a small amount of missing data (0.2%) exists in the "Not Stated" group. Urbanization levels are defined using a six-tier classification based on the updated 2013 urbanization scheme, which includes Large Central Metro, Large Fringe Metro, Medium Metro, Small Metro, Micropolitan/Non-metro, and non-core/non-metro areas. Age is categorized into ten groups: less than 4 years; 5–14 years; 15–24 years; 25–34 years; 35–44 years; 45–54 years; 55–64 years; 65–74 years; 75–84 years; and those aged 85 years and older. The age groups for less than 1 year and 1–4 years are combined due to small counts in specific demographic subgroups. Finally, the total monthly count is summarized as an overall count.

Based on the CDC WONDER reporting application (CDC, 2025a), death counts of less than ten (and associated statistics) in tables, graphs, and figures were not reported due to statistical reliability and data confidentiality concerns.

Data analysis procedure: First, a new “average date” variable has been created to examine the potential effects of the varying number of days (ranging from 28 to 31 each month), which serves as the outcome variable for modeling purposes (Output 2). Next, the study compares the forecasting capabilities of robust ARIMA and new ESM procedures.

Since its introduction by Box and Jenkins (1976), the ARIMA analysis procedure has been widely adopted by researchers and is structured into three main stages, as depicted above in the Conceptual Framework and discussed below.

Step 1. Model Identification: The initial step in constructing an ARIMA model involves assessing the stationarity of the time series trend to identify the best ARIMA model, which will be compared with the ESM models in Step 3 below. A trend is an overall direction of time series data over a period, which can be classified as increasing (upward), decreasing (downward), or horizontal (stationary). A stationary time series exhibits consistent statistical properties, such as a constant mean, variance, and autocorrelation over time. Autocorrelation function plots and the Augmented Dickey-Fuller (ADF) test (Brocklebank et al., 2018; Dickey & Fuller, 1979) were used to determine the stationarity of the crash time series data utilized in this study (Appendix 1, Output 3).

A formal test and visual inspection of the autocorrelation function (ACF) plots indicate that the crash data were nonstationary since the ACF decays very slowly (Output 3). Additionally, a strongly significant chi-square statistic ($p < .0001$) indicates that the model does not fit well (Output 3; Brocklebank et al., 2018). Likewise, the “Augmented Dickey-Fuller Unit Root Tests” for Single Mean, Tau = -3.00 ($p = .0424$) was < .05, we reject the null hypothesis test stating that the data has a unit root (i.e., that the series is nonstationary) (Output 3; Brocklebank et al., 2018, pp. 85-86). Furthermore, per Brocklebank et al. (2018), the adjustment for the lagged differences was motivated by large sample theory, and $n = 48$ months was not particularly large and required a realistic model for our data.

The summary statistics (Appendix 1, Output 3) show that the period of differencing was set at (1, 12), meaning that 13 observations were removed through this operation. A chi-square test (p = .1544) > .05 indicates that this model does fit better than previous iterations. The differencing operation was necessary to transform a nonstationary time series into a stationary one, making it easier to model and predict future values.

Differencing (1) removes upward or downward trends by subtracting the value from the previous time step from the current value, (2) it subtracts values from the same season in prior cycles to handle periodic fluctuations, (3) it decreases heteroscedasticity, creating a more consistent variance. However, the limitation of the differencing operation is that it reduces the number of observations.

Next, the partial and inverse autocorrelation function plots and the check for white noise were used to aid in identifying appropriate ARIMA models for the series. Thus, based on Brocklebank and coauthors (2018, p. 83), we used a “Moving Average MA (1) model” to fit this process (Appendix 1, Output 4). MA (1) refers to a Moving Average model where the current value was based on the mean of the previous period's value and a random error term.

Data analysis procedure: Step 2. Model Estimation and Validation (Diagnostic Checking): The next step involved checking the diagnostic statistics to determine if the MA (1) model was adequate (Output 4). Other candidate models included low-order mixed ARMA models. This part utilized autocorrelation checks for white noise and chi-square statistics, which aided in comparing this model to others, leading to identifying the most suitable model. The additive and multiplicative models were the main approaches used to describe the relationship between the observed series and unobserved components (Brocklebank et al., 2018, pp. 285-286). Subsequently, ARIMA forecasts using a seasonal multiplicative moving average model were evaluated. However, several warnings from SAS software were issued: estimates may not have converged, and the model defined by these new estimates was unstable, resulting in the termination of the iteration process.

Step 3. Model Applying (Forecasting Stage): In the final stage of the analysis, we identified the best models and used them to generate forecasts. The accuracy of each forecast was then compared using Mean Absolute Percentage Errors (MAPEs), a commonly used metric that expresses errors as a percentage of actual data. This comparison utilized training data from July 2020 to December 2022 after removing holdout data (i.e., testing data from July 2023 to June 2024) for forecasting purposes—predicting outcomes over 12 months served as an indication of how well each model could predict death counts.

As a final step, we determined four highly accurate forecast models based on MAPE results (Appendix 1, Output 7), with average accuracy percentages as follows:

1. Exponential Smoothing Model - Seasonal Method	(98.5%)
2. Exponential Smoothing Model - Additive Winters Method	(98.5%)
3. Exponential Smoothing Model - Multiplicative Winters Method	(97.8%)
4. ARIMA - Moving Average Model - MA (1)	(97.7%)

(1) The Seasonal Method focuses only on additive seasonality without accounting for trends. (2) The Additive Winters Method incorporates both additive seasonality and trends but assumes constant seasonal effects. (3) The Multiplicative Winters Method also includes both trends and seasonality but allows for proportional changes in seasonal effects relative to the level of the series. For this study, both ESM seasonal methods produced a notably high overall average accuracy percentage of 98%. All data processing and analyses were conducted using SAS software version 9.4 (SAS, 2012).

Ethical consideration: This research utilized de-identified publicly available secondary data from the CDC/National Center for Health Statistics and did not require Institutional Review Board (IRB) approval (Common Rule 45 CFR§46; IRB/UGA Policy 3.2, 2015).

Results

During the study period from July 2020 to June 2024, there were a total of N = 175,600 reported deaths in the US attributed to motor vehicle accidents (Appendix 1, Output 1). This averages out to approximately 3,658 deaths per month. Based on the diagnostic tests and high accuracy percentages, the Exponential Smoothing Model (ESM) was selected to produce the final forecasts by month and demographic factors, as discussed previously in the Methods section and detailed below.

This study found that in the 12 months from July 2023 to June 2024, there were 42,206 reported deaths resulting from motor vehicle crashes. Consequently, the predicted number of deaths during this period was slightly higher at 42,612, yielding an average forecast accuracy of 98% as measured by Mean Absolute Percentage Error (MAPE) (Table 1). In the subsequent 12 months (July 2024-June 2025), projections indicate there will be approximately 41,829 fatalities due to crashes, translating to an average of around 3,500 monthly deaths (Table 1). We found that average forecast accuracy increases as crash counts increase in agreement with the large sample theory (Scott, 2021).

Males constituted a significant majority of reported deaths, totaling 126,471 (72%) of all crashes. Similarly, predictions for the upcoming 12 months suggest that out of an estimated total of deaths, approximately 30,087 (72%) will involve male individuals (Table 2). The average forecast accuracy is 98% for males and 97% for females.

In absolute numbers (not rates), among crash predictions, White individuals represented over three-quarters at 76.5% (with an average forecast accuracy of 97%), followed by Black or African American at 17.8% (accuracy of 96%), Asian at 2.3% (accuracy of 91%), and AI/AN just at 1.8% (accuracy of 84%), (Table 3).

When examining crash predictions based on urbanization levels in absolute numbers, Large Central Metro accounts for 27.7% (accuracy of 97%) of the total, followed by Medium Metro at 24.7% (accuracy of 98%). Large Fringe Metro is projected to represent 17.6% (accuracy of 97%). Additionally, Non-metro Micropolitan is estimated at 9.7%, and Non-Core (Non-metro) stands at 9.3% (each with an accuracy of 95%) (Table 5).

Examining the data on ethnicity and Hispanic origin presented in Table 4, forecasts for the upcoming year suggest that a significant majority—33,618 individuals, or approximately 80.5%—will identify as non-Hispanic or Latino. Conversely, 8,142 crashes are estimated to involve Hispanic or Latino individuals, accounting for about 19.5% (with both average forecast accuracies of over 97%).

Among the projected crash estimates, nearly half of those involved is estimated to be between the ages of 15 and 44 years (49.6%), while individuals aged 45 to 64 years will represent 26.6% of the total (each accuracy of over 95%) (see Table 6).

Figure 3 illustrates actual (i.e., reported) motor vehicle fatalities alongside predicted death counts represented by dashed lines (males) and solid lines (females). The dotted trend lines indicate the 95% confidence intervals surrounding these forecasts. The predicted number of male fatalities is expected to peak between July 2024 and October 2024, exceeding 2,600 deaths each month. Following this peak, crashes are projected to decline in February 2025 before rising again in May and June of that year, with estimated deaths of approximately 2,375 and 2,403, respectively (Figure 3).

Limitations

The findings in this study are subject to at least six limitations. First, the robustness of the predictions might have been improved further with more time series data. However, to address the impact of the COVID-19 pandemic (which caused a sudden jump in crashes), data from the pre-pandemic period (before June 2020) were excluded, reducing the available data for accuracy testing. Second, unforeseen events or the use of provisional death counts in this study (due to a lack of final data) may result in minor changes (approximately 0.06%) to the predictions. Third, other risk factors for crash-related deaths—such as vehicle characteristics, health behaviors (e.g., alcohol and drug use), and roadway conditions (e.g., traffic density)—were not available from the underlying cause of death – Injury Intent and Mechanism. Fourth, the NVSS defines urbanization level using the county of a person’s legal residence rather than the county where the crash occurred. Fifth, some racial misclassification is likely in death certificate data and could result in undercounts for specific sub-populations (Rosenberg, 1999). Finally, it remains unclear whether individuals involved in fatalities were drivers themselves or victims of accidents caused by other drivers, limiting our understanding of responsibility and risk factors.

Discussion and Conclusion

This research showcases high forecast accuracy and highlights the significant influence of nuanced demographic factors on traffic safety outcomes. By integrating these dimensions into forecasting models, we can more effectively design traffic safety campaigns and strategies to reduce motor vehicle fatalities across the US, enhancing public health outcomes. The findings of this study underscore the importance of adopting optimal predictive methods to develop accurate forecasting models for motor vehicle fatalities in the US. By stratifying data by demographic factors, we can enhance our understanding of the dynamics surrounding motor vehicle crashes and improve the precision of our predictions. This research aligns with previous studies that have successfully utilized time series data from various sources to generate near-real-time projections of crash-related deaths and injuries (Agyemang et al., 2023; Hao & Liu, 2024).

The implications of these findings are significant for bolstering initiatives led by the CDC aimed at reducing motor vehicle crash deaths and injuries. Providing timely crash projections offers critical insights to inform effective strategies and interventions tailored to specific demographic groups. For instance, recognizing that males account for a substantial majority (72%) of reported deaths allows targeted outreach efforts to address this high-risk group more effectively. Similarly, understanding that White individuals represent over three-quarters (76.5%) of forecasted fatalities—slightly above their population representation (75.3%)—and that Black individuals account for 17.8% of forecasted fatalities, exceeding their population percentage of 13.7% highlights the need for culturally sensitive prevention strategies.

Moreover, the stratification by age reveals that nearly half (49.6%) of those involved in projected crashes fall within the 15 to 44-year age range, indicating a crucial demographic for targeted educational campaigns and interventions. The CDC can leverage these insights to create actionable resources to evaluate, plan, and implement improved guidelines designed to decrease fatalities among these vulnerable populations.

Urbanization levels also play a critical role in shaping traffic safety strategies. With Large Central Metro areas accounting for 27.7% of total crash predictions followed closely by Medium Metro areas at 24.7%, it is evident that urban environments require distinct approaches compared to rural settings. Tailoring interventions based on urbanization levels can enhance their effectiveness in reducing crash incidents.

We found that average forecast accuracy increases as crash counts increase in agreement with the large sample theory (Scott, 2021). The highest forecast accuracy, ranging from 97% to 98%, was observed among males and females, both Hispanic or Latino and Not Hispanic or Latino, as well as Whites, Medium Metro, Large Central Metro, and Large Fringe Metro areas, age groups of 15-24 and 35-44 years. The lowest forecast accuracy, ranging from 82% to 87%, was found among individuals aged 14 years and younger and those identifying as AI/AN.

Data sparsity poses a significant challenge in forecasting crash occurrences, particularly in areas with fewer incidents where limited data can lead to unreliable predictions. Small sample sizes are more vulnerable to outliers and variability, complicating the ability of models to identify consistent trends (Scott, 2021). Additionally, behavioral differences among various demographic or geographic groups can further complicate forecasts, as distinct driving behaviors—such as variations in risk-taking, compliance with traffic laws, and exposure levels—are often inadequately captured by existing models. Lastly, external factors such as infrastructure disparities—including road quality, traffic density, and distinctions between urban and rural settings—also play a crucial role in influencing crash patterns; regions with poor infrastructure tend to exhibit greater unpredictability in crash occurrences compared to well-maintained areas.

Refining predictive models to account for specific local conditions and behaviors can improve resource allocation and facilitate targeted interventions to enhance traffic safety in high-risk or underreported areas. By implementing tailored strategies, such as community-specific safety campaigns or infrastructure improvements, stakeholders can address unique reporting challenges and behavioral issues, ultimately leading to better traffic safety outcomes in those communities with lower prediction accuracy.

Understanding the reasons behind data sparsity and underreporting can inform strategies for improving data collection methods, leading to enhanced reporting systems in areas where crashes are frequently underreported. Additionally, insights from improved predictive models can guide resource allocation—such as law enforcement presence or educational programs—toward areas most in need based on their distinct crash patterns. Engaging local communities in discussions about their specific traffic safety challenges fosters collaboration between residents and stakeholders, resulting in more effective solutions tailored to local needs. By integrating these insights into prevention frameworks, stakeholders can address disparities across different demographic groups while significantly improving overall traffic safety outcomes.

In summary, this research not only demonstrates a high forecast accuracy but also emphasizes how nuanced demographic factors—can significantly influence traffic safety outcomes. By integrating these dimensions into forecasting models, we can better shape traffic safety campaigns and strategies to reduce motor vehicle fatalities across the US, ultimately contributing to enhanced public health outcomes.

Conclusion: This study emphasizes the importance of stratified data analysis in improving predictive models for motor vehicle fatalities in the US. By considering demographic factors, we can create targeted interventions for high-risk populations. The findings reveal a notable forecast accuracy of 98%, highlighting the potential of these models to guide effective traffic safety strategies. Utilizing these insights will empower stakeholders to implement tailored initiatives to reduce motor vehicle crash deaths and enhance public health outcomes across communities. The combination of reported crash case data and timely estimates supports the development of effective prevention measures. Future research should expand to include additional injury mechanisms—such as falls, suicide, self-harm, and drowning—while integrating various forecasting methods to enhance predictive capabilities and improve accuracy in future projections.

Chapter 4: The Impact of Demographic Factors in Motor Vehicle Fatalities

Introduction

Motor vehicle (MV) fatalities remain a critical public health challenge in the United States, shaped by demographic factors that influence safety behaviors and outcomes. Multiple studies conducted across different countries have identified disparities in traffic fatalities based on race, ethnicity, sex, and age across various modes of transportation and geographical areas (Beck et al., 2017; Beck et al., 2019; Cullen et al., 2021; Raifman & Choma, 2022). For instance, Cullen et al. (2021) found that young men are disproportionately affected by road crashes and injuries compared to their female counterparts and older drivers. Despite these findings, there remains a limited understanding of how these sex differences evolve with increased driving experience.

Further research underscores persistent inequalities in MV traffic death rates among different demographics. Between 1999 and 2019, disparities among males by age group decreased; specifically, while males aged 15-24 had the highest fatality rate in 1999, by 2019, their rate fell below that of males aged 25-64 and those aged 65 and older (Spencer et al., 2021). During this same period, MV traffic death rates among females decreased across all age groups; however, overall rates began to rise from 2010 through 2019. Additionally, early investigations indicate higher all-mode traffic fatalities among American Indian and Black Americans compared to White Americans, while Asian Americans experience the lowest fatality rates (Raifman & Choma, 2022).

Driving behavior also exhibits significant sex differences. Storie (1977) found that females tend to drive at lower speeds and engage in fewer risky behaviors than males, who generally demonstrate greater skill but also a higher propensity for dangerous practices such as driving under the influence of alcohol. However, mixed findings exist regarding fault in accidents; Ng et al. (1991) reported no significant difference based on driver sex concerning accident involvement or information-seeking behavior across ages.

Geographic context further complicates the landscape of motor vehicle fatalities. Beck et al. (2017) noted that lower seatbelt usage in rural areas contributes to higher fatality rates—50% of fatally injured occupants on rural roads were unrestrained in 2015 compared to only 46% in urban areas. Moreover, intersection crashes pose a serious risk for older drivers, underscoring the need for targeted strategies as the aging population increases (Samuel et al., 2016; Lombardi et al., 2017).

Understanding these variables is essential for crafting effective prevention strategies that enhance community health and safety (Beck et al., 2017; CDC, 2024a). The existing literature highlights a pressing need for targeted safety programs aimed at promoting consistent practices within specific populations.

Thus, this study seeks to comprehensively address disparities in motor vehicle fatalities across various population groups in the US by analyzing key demographic factors. By identifying sub-populations disproportionately at risk for fatal accidents—particularly focusing on how these variables correlate with motor vehicle incidents—traffic safety experts can develop effective prevention strategies tailored to specific demographics. The study employs two regression models to explore relationships between demographic factors and motor vehicle fatalities over time (after the beginning of the COVID-19 pandemic). Ultimately, this research intends to identify vulnerable areas and populations susceptible to accidents and injuries while facilitating the development of targeted interventions designed to enhance road safety for all users—thereby reducing disability and mortality related to motor vehicle incidents nationwide. This approach addresses a research gap regarding assessing disparities in MV deaths among different subgroups over time—a necessary step toward improving public health outcomes related to road safety.

Methods

Study design: This descriptive study utilizes a quantitative correlational research approach that examines relationships between variables without manipulating them to gain insights into the strength and nature of these relationships. The study describes the research question related to analyzing the impact of demographic factors on crash count data.

Data sources: We utilized data from the CDC's National Vital Statistics System (NVSS) multiple cause-of-death motor vehicle mortality data. The CDC collects death certificate nationally comprehensive data from vital statistics offices in all 50 states and the District of Columbia. The National Center for Health Statistics (NCHS) CDC/WONDER (Wide-ranging Online Data for Epidemiologic Research) provides these public data. Crash counts were tabulated for the query parameters in the CDC WONDER reporting application (CDC, 2025a). The query parameters included underlying cause of death (UCD): "Unintentional"; UCD – "Injury Mechanism & All Other Leading Causes"; Motor Vehicle Traffic, grouped by Month, Sex, Urbanization, and Ten-Year Age Groups. The data were collected from July 2020 to December 2022 (finalized data) and from January 2023 to June 2024 (provisional data), totaling 48 months of data (CDC, 2025a).

Dependent Variable

The dependent variable is a monthly number of motor vehicle deaths stratified by demographic variables.

Independent Variables

The independent variable sex is categorized into two groups: male and female. Urbanization is assessed from most urban to least urban, with classifications encompassing large central metro areas, large fringe metro areas, medium metro areas, small metro areas, micropolitan areas (referring to towns or cities that are non-metro), and the non-core regions (representing rural or non-metro locations). Age is divided into ten distinct groups: less than 4 years old, 5 to 14, 15 to 24, 25 to 34, 35 to 44, 45 to 54, 55 to 64, 65 to 74, 75 to 84, and those aged 85 years and older. The age categories for "< 1 year" and "1-4 years" are combined into a single group labeled "< 4" due to the small number of deaths (n = 127) in the "< 1-year" group. Lastly, the total monthly count is summarized as an overall count.

Other variables included Period (July 2020 to June 2022 versus July 2022 to June 2024, to compare early and later COVID-19 periods, Table 7), Year, and Season.

According to CDC, the urban-rural updated Residence 2013 six-level classification scheme for counties includes: 1) large "central" metro (akin to inner cities), part of a metropolitan statistical area (MSA) with ≥1 million population and covers a principal city; 2) large "fringe" metro (akin to suburbs), part of a metropolitan statistical area with ≥1 million population but does not cover a principal city; 3) medium metro: part of a metropolitan statistical area with ≥250,000 but <1 million population; 4) small metro: part of a metropolitan statistical area with <250,000 population; 5) micropolitan (non-metro): part of a micropolitan statistical area (has an urban cluster of ≥10,000 but <50,000 population); and 6) non-core (non-metro): not part of a metropolitan or micropolitan statistical area.

Death counts of fewer than ten were not reported due to data confidentiality concerns. The study utilized monthly summary crash data rather than individual-level data to leverage the most up-to-date information. Furthermore, to conduct regression analyses, the number of crashes was summarized for each combination of 2 sexes, 6 race categories, 6 urbanization levels, and 10 age groups across 12 months over 4 years (34,560 combinations). This level of combination in the monthly summary data resulted in several data losses (16.6%), which were suppressed due to small counts for some subgroups. Thus, summarizing the data while excluding a race category resulted in better lower combinations and only a loss of 4.3%, leading to a total of N = 168,370 crash records available for regression analyses (Table 8).

Data analysis procedure

First, descriptive statistics are computed for demographic variables to summarize the frequency and proportion of the regression analysis dataset (Table 7). Next, the study examines a histogram of the data, which includes a superimposed normal distribution and a nonparametric kernel density curve fit to assess whether the normal distribution is an appropriate model (Figure 5). Furthermore, summary statistics from the univariate procedure (SAS, 2012) and goodness-of-fit tests for normality are also evaluated.

This study employs two regression methods suitable for analyzing count data: Poisson regression and negative binomial regression (Agresti, 2012; Cameron & Trivedi, 2013; SAS, 2012). The Poisson model is recognized as the foundational modeling for count data (Agresti, 2012; Cameron & Trivedi, 2013). In contrast, the negative binomial method is utilized to address potential overdispersion in the count data—defined as a scenario where observed variability in counts (in this case, motor vehicle traffic deaths) exceeds what would be anticipated under a Poisson distribution.

Overdispersion may arise due to clustering within populations such as families, households, neighborhoods, or urban areas (SAS, 2012). Unlike the Poisson distribution—which assumes that mean and variance are equal—the negative binomial distribution incorporates a dispersion parameter that can either be estimated or set to a fixed value. The variance function for the negative binomial distribution is expressed as:

$$\text{Variance} = \mu + k(\mu^2)$$

Where μ represents the mean and k denotes the dispersion parameter. This structure allows variance to exceed mean values and effectively accommodates overdispersion (SAS, 2012). When k equals zero, it aligns with the assumptions of a Poisson regression model where variance equals mean.

The analysis employs a log link function due to its prevalence in count data modeling compared to an identity link function (Agresti, 2012). Although modeling rates of occurrence may be more relevant when outcomes are measured over time or space—such as monthly crash data—an appropriate denominator for calculating rates (e.g., per mile driven) is not available.

Consequently, based on model diagnostics indicating better fit and flexibility compared to the Poisson model, a negative binomial regression model is selected. This model analyzes monthly motor vehicle mortality counts while controlling for selected factors (Cameron & Trivedi, 2013; SAS, 2012).

Results

Descriptive Analysis

During the study period from July 2020 to June 2024, a total of 168,370 deaths in the US were reported as being attributed to motor vehicle traffic crashes, averaging over 115 fatalities per day. Males accounted for 73.3% of these reported deaths, which is more than three times the percentage of females (Table 7; Figure 4). Analysis indicates no significant difference in deaths by sex between the early period (July 2020 to June 2022) and the later period (July 2022 to June 2024), with a chi-square statistic ($p = 0.46$).

In absolute numbers, motor vehicle fatalities were greater in Large Central Metro areas than in non-core (non-metro) areas. This difference can be largely attributed to the significantly larger populations and increased vehicle traffic in metropolitan areas, leading to a greater number of deaths. However, it is noteworthy that the fatality rate per mile driven tends to be lower in urban settings than in rural areas (Beck, 2017; NHTSA, 2020). Large central and medium metro areas accounted for approximately half of all crashes (47.8%), followed by large fringe metro areas at 20.9% (Table 7). By age group, individuals aged 25 to 44 years alone accounted for over one-third (35.4%) of crashes, while individuals aged 45 to 64 years constituted 28.3% (28.3%) (Table 7).

Traffic-related death counts varied by race; White individuals represented a significant majority at 79%, followed by Black or African American individuals at 18% (see Table 7). This disparity is even more pronounced among females, with crashes accounting for 82.2% of White individuals and 17.5% of Black or African American individuals. Moreover, over one-third of the crashes occurred between July and October (36.9%).

Regression Analysis

The univariate skewness statistic was calculated at 1.43, indicating a positive skew since it exceeded zero; furthermore, the mean crash count value was found to be greater than the median (146.9 vs. 114.0), suggesting a positively skewed distribution, as illustrated in Figure 5. The variance was determined to be much larger than the mean value (14,901 vs. 146.9). Moreover, goodness-of-fit tests for normal distribution ($p < .01$) suggest that the death counts may not follow a normal distribution (Output 8; Figure 5).

Collinearity diagnostics (Output 8) revealed that the largest Variance Inflation Factor (VIF) was recorded at 1.00—well below the threshold of ten—while Condition Index values also remained below ten during collinearity diagnostics assessments, thus indicating that "multicollinearity does not appear to be a problem" (SAS, 2012, pp. 4-57). Multicollinearity refers to situations in regression analysis where high correlations among predictors obscure their individual contributions, explaining variability in the dependent variable (Agresti, 2012).

Significant predictors and interaction terms were identified as anticipated contributors toward explaining variability within each predictor variable. Interactions occur when the impact of one independent variable on the dependent variable varies based on the level of another independent variable. Thus, simple effects tests were conducted rather than relying solely on interpretations of the main effects.

The criteria used for assessing goodness-of-fit when fitting a Poisson model indicated (Output 10) Pearson Chi-Square values equating to 3.5 —potentially indicating data overdispersion commonly encountered within Poisson regression (SAS, 2012), which could lead towards underestimating standard errors while overestimating test of statistics.

Output 10 statistics presented additional diagnostic measures, including the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), both serving as indicators balancing model fit against simplicity—with smaller values representing better model fit overall alongside Chi-Square tests assessing significance across demographic parameters provided herein. As anticipated, fit statistics for the negative binomial were better compared to those of the Poisson model (Output 9).

The assessment of the goodness of fit for the negative binomial model (Output 11) yielded a Pearson Chi-Square value of 1.2. A value closer to 1 suggests overdispersion in the data (SAS, 2012, pp. 5-34). The results from the negative binomial model, which accounted for overdispersion, are presented in Output 11. The dispersion parameter was estimated using maximum likelihood at 0.0043 and was significantly different from zero, as indicated by the 95% confidence interval (0.0033, 0.0055), which does not include zero.

In the "GENMOD" procedure (Output 11) utilizing the negative binomial model, all predictor variables (sex, urbanization level, age, and month) were found to be strongly significant at an alpha level of .05. The likelihood ratio (LR) statistics for Type 3 Analysis indicated $p < .0001$. The significance of each predictor variable was assessed by comparing the fit of the "full model" with that of a "reduced model," which excluded each predictor variable; this demonstrated that their inclusion improved the model's explanatory power. Thus, the final full model retained all four predictor variables.

In absolute numbers, the high crash count group (designated as a high-risk group) was selected as the "Reference" group (i.e., Males, Large Central Metro areas, Age group 25-44 years, and October) (Table 8). The "Analysis of Maximum Likelihood Parameter Estimates" (Table 8; Output 11) reveals a significant difference in crash counts between sexes while controlling for selected factors; males experienced higher fatalities indicated by a negative coefficient for females (-1.101, p < .0001).

Additionally, individuals in the age group of 25-44 years exhibited significantly higher crash counts compared to those in the age groups of 5-24 years (-0.794, p < .0001), 45-64 years (-0.299, p < .0001), and those aged 65+ years (-0.949, p < .0001) (Table 8). Furthermore, October recorded significantly higher crashes than any other month, except August, which showed no significant difference (-0.027, p = .139) when compared to October (notably August having the second-highest monthly reported crashes) (Table 8).

Regarding geographic disparities, there were notable differences in absolute numbers among urbanization levels; significantly higher crash counts were observed in Large Central Metro areas compared to Large Fringe Metro areas (-0.265, p < .0001), Medium Metro areas (-0.176, p < .0001), Micropolitan (Non-metro) (-0.958, p < .0001), and Small Metro areas (-0.992, p < .0001) (Table 8).

In summary, the analysis indicates that males had higher fatalities in crashes, individuals aged 25-44 experienced more crashes than those in the 5-24, 45-64, and 65+ age groups, October saw a spike in crash counts, and Large Central Metro areas reported significantly more crashes than other urbanization levels.

Limitations

The study has at least five potential shortcomings. First, more time series data could have enhanced the regression analysis. However, to mitigate the effects of the COVID-19 pandemic (which caused a sudden increase in crashes), data from before June 2020 were excluded, leaving only 48 months of available data. Second, the use of provisional death counts (due to the unavailability of final data) may lead to small discrepancies. Third, suppressed crash summary data for counts under ten (< 4.3%) were not available for analysis due to confidentiality constraints such as AI/AN, Native Hawaiian or Other Pacific Islander. Fourth, the National Vital Statistics System categorizes urbanization based on an individual's legal residence rather than the location of the crash itself. Lastly, there is a likelihood of racial misclassification in death certificate data that could lead to underreporting for some minority subpopulations (Rosenberg, 1999).

Discussion

This analysis highlights the value of demographic factors in modeling and describing monthly crash death counts. The analysis of motor vehicle accident fatalities in the US reveals the need for urgent attention from stakeholders. With a total of 168,370 reported deaths—an average of over 115 fatalities each day—the data underscore the ongoing public health problem posed by road traffic incidents.

A notable finding is the predominance of male victims, who accounted for 73.3% of the fatalities. This aligns with existing literature suggesting that males are generally at higher risk for severe outcomes in traffic accidents, potentially due to factors such as risk-taking behavior and greater exposure to driving situations (Cullen et al., 2021; Raifman & Choma, 2022). The implications of this demographic trend call for targeted interventions explicitly aimed at male drivers (ages 25-44 years), who represent a significant portion (36.2%) of those involved in fatal crashes.

Urbanization patterns further complicate the landscape of motor vehicle safety. The data indicate that approximately half of all crashes occurred in the Large Central Metro and Medium Metro areas. This suggests that urban environments may present unique challenges related to traffic density, road design, and driver behavior (Pollack et al., 2014). Stakeholders should consider implementing tailored strategies for these regions, such as improved traffic management systems and enhanced public transportation options to reduce congestion and promote safer travel alternatives.

The age distribution of crash victims also highlights critical areas for intervention. The highest percentage of fatalities was among individuals aged 25 to 44 years, indicating a need for focused educational campaigns about safe driving practices targeted at younger drivers, in line with the research conducted by Cullen et al. (2021) and Raifman & Choma (2022). Additionally, the findings reveal significant seasonal variations in crash rates, with over one-third of deaths occurring between July and October, aligning with the research conducted by Faust et al. (2021), Teymuri et al. (2014), and the Department of Transportation (2024a). This temporal pattern suggests that those months may require heightened awareness campaigns or enforcement efforts during peak travel times.

In conclusion, addressing motor vehicle accident fatalities requires a multifaceted approach considering demographic vulnerabilities, urbanization effects, and seasonal trends. Stakeholders should prioritize evidence-based interventions for high-risk populations while enhancing road safety in urban environments.

Chapter 5: Discussion and Conclusion

Introduction

Motor vehicle (MV) crashes continue to pose a significant public health challenge in the United States, resulting in over 100 deaths daily (CDC, 2025a). Despite ongoing efforts to enhance data collection and reporting on injuries, substantial delays persist in obtaining timely final injury-related data. This lag impedes monitoring trends and responding to emerging risks effectively. Furthermore, existing studies highlight disparities in MV crash outcomes based on sex, race, ethnicity, and age (Cullen et al., 2021; Raifman & Choma, 2022), highlighting the urgent need for more advanced forecasting methods.

Hence, to address this gap, the first study employs time series data and an advanced statistical forecasting technique known as Exponential Smoothing Models (ESM). The primary objective is to generate reliable predictions stratified by demographic subgroups, thereby enhancing the understanding of how traffic crash trends vary over time. By providing these analyses, the study aims to inform more effective prevention strategies tailored to populations at higher risk of traffic fatalities.

The second study focuses on investigating how various demographic factors influence motor vehicle crashes. Previous research has identified significant disparities in motor vehicle safety behaviors and outcomes among different demographic groups, emphasizing the necessity for targeted prevention strategies that enhance community health and safety (Boal et al., 2016; Beck et al., 2019). Through a comprehensive analysis of these relationships, this study seeks to provide valuable insights for stakeholders such as public health professionals, researchers, and clinicians.

Thus, to achieve this goal, the second study employs two regression models to assess the associations between demographic factors and motor vehicle fatalities. This analysis seeks to uncover how specific variables impact fatalities, enabling stakeholders to develop tailored interventions that address the unique risks faced by diverse populations. Ultimately, this study aspires to make a significant contribution toward reducing disability and mortality associated with motor vehicle accidents across the US by equipping decision-makers with data-driven strategies for effective prevention.

First Study - Chapter 3 Summary

The first study's findings highlight the critical need for optimal predictive methods to develop accurate forecasting models for motor vehicle fatalities in the United States. By stratifying data by sex, age, race, and urbanization level, we can deepen our understanding of the dynamics surrounding motor vehicle crashes. This research aligns with previous studies that have effectively utilized time series data from diverse sources to generate near-real-time projections of crash-related deaths and injuries (Agyemang et al., 2023; Hao & Liu, 2024).

Key Messages

Forecasting significantly enhances the timeliness of provisional data through high forecast accuracy achieved via predictive modeling, effectively eliminating a one-year data lag and allowing for more effective traffic safety strategies. The study underscores the necessity for targeted interventions based on demographic insights, particularly highlighting the elevated risk among males and White individuals, with nearly half of projected crash involvement occurring in those aged 15 to 44. Large Central Metro and Medium Metro regions also account for substantial crash predictions, making tailored strategies essential for improving overall intervention effectiveness. Additionally, the findings create opportunities for collaboration among stakeholders to explore these predictive models further while facilitating the timely sharing of scientific findings across various sectors.

The study achieved a notable average forecast accuracy of 98%, measured by Mean Absolute Percentage Error (MAPE). This highlights how nuanced demographic factors—such as sex distribution, racial composition, age brackets, and urbanization levels—can significantly influence traffic safety outcomes.

Future Improvements Needed: Building on these findings, future research should expand its scope to include additional injury mechanisms such as falls, suicide, self-harm, and drowning while integrating various forecasting methods to enhance predictive capabilities further.

Broader Data Integration: Incorporating more comprehensive datasets could improve model accuracy and reliability.

Continuous Monitoring: Establishing systems for ongoing data collection and analysis will allow real-time adjustments to strategies based on emerging trends.

Enhanced Community Engagement: Engaging communities in developing tailored initiatives will ensure that interventions are culturally relevant and effective.

Therefore, this study emphasizes the vital role of stratified data analysis in improving predictive models for motor vehicle fatalities. By leveraging these insights into targeted interventions for high-risk populations, stakeholders can implement effective strategies to reduce motor vehicle crash deaths and enhance public health outcomes across the US.

Second Study - Chapter 4 Summary

The regression analysis underscores the critical role of demographic factors in modeling and understanding monthly crash death counts. While the study does not assert a causal relationship, it reveals clear associations between these factors and motor vehicle traffic fatalities. The examination of motor vehicle accident fatalities in the US highlights an urgent need for action from stakeholders, as evidenced by a staggering total of 168,370 reported deaths—an average of over 115 fatalities each day—emphasizing the ongoing public health crisis posed by road traffic incidents.

A significant finding is that utilizing a negative binomial model yielded more reliable results by effectively accounting for data variability. All predictor variables—including sex, urbanization level, age group, and month—were statistically significant at an alpha level of .05. Notably, male victims constituted 73.3% of fatalities, aligning with existing literature that indicates males are generally at higher risk for severe outcomes in traffic accidents due to factors such as risk-taking behavior and increased exposure to driving situations. This demographic trend necessitates targeted interventions aimed specifically at male drivers (25-44 years), who represent a substantial portion (36.2%) of those involved in fatal crashes.

Urbanization patterns further complicate the landscape of motor vehicle safety; approximately half of all crashes occurred in Large Central Metro and Medium Metro areas. This suggests that urban environments may present unique challenges related to traffic density, road design, and driver behavior. Stakeholders should implement tailored strategies for these regions—such as improved traffic management systems and enhanced public transportation options—to alleviate congestion and promote safer travel alternatives.

The age distribution among crash victims also points to critical areas for intervention; individuals aged 25 to 44 years accounted for the highest percentage of fatalities. This indicates a pressing need for focused educational campaigns about safe driving practices directed at younger drivers. Additionally, significant seasonal variations in crash rates were observed, with over one-third of deaths occurring between July and October (36.9%). This temporal pattern suggests that heightened awareness campaigns or enforcement efforts during peak travel months could be beneficial.

Conclusion

This comprehensive analysis reveals vital trends and demographic patterns regarding motor vehicle accident fatalities in the US, spanning from July 2020 to June 2024. The alarming total of 168,370 reported deaths emphasizes a significant public health concern requiring immediate attention. Therefore, addressing motor vehicle accident fatalities requires a multifaceted approach that considers demographic vulnerabilities, urbanization effects, and seasonal trends. To reduce this persistent public health issue, stakeholders must prioritize evidence-based interventions while enhancing road safety measures across various contexts.

Limitations

The findings in this study are subject to several limitations. First, the robustness of the predictions could be enhanced with more time series data; however, to mitigate the effects of the COVID-19 pandemic—which caused a sudden increase in crashes—data from the pre-pandemic period (before June 2020) were excluded, leaving only 48 months of available data for accuracy testing. Second, using provisional death counts due to the unavailability of final data may lead to minor discrepancies (approximately 0.06%) in predictions. Additionally, suppressed crash summary data for counts under ten (< 4.3%) are not available for analysis due to confidentiality constraints; specific demographic categories with counts fewer than ten—including American Indian or Alaska Native, Native Hawaiian or Other Pacific Islander, and individuals identifying as more than one race—are also unavailable. Third, other risk factors for crash-related deaths—such as vehicle characteristics, health behaviors (e.g., alcohol and drug use), and roadway conditions (e.g., traffic density)—are not included in the underlying cause of death (UCD) – Injury Intent and Mechanism. Fourth, the National Vital Statistics System defines urbanization levels based on an individual's legal residence rather than where the crash occurred. Lastly, there is a likelihood of racial misclassification in death certificate data that could result in undercounts for specific subpopulations. Despite these limitations, the study provides a more effective means of describing and estimating MV death counts.

Recommendations and Implications

The findings from both studies highlight the urgent need for targeted interventions to reduce motor vehicle fatalities in the US and to improve data collection and sharing methods. Data collection can be enhanced by leveraging advanced technologies, such as mobile applications and real-time reporting systems, to increase traffic incident data's accuracy and comprehensiveness while engaging communities in the reporting process. The high average forecast accuracy achieved through predictive modeling demonstrates its potential to share data faster and can inform effective traffic safety strategies. Finally, collaboration among the CDC's transportation team, local governments, community organizations, and law enforcement is crucial for effective data sharing and informed decision-making based on predictive models.

Areas for Improvement: Future Research Directions

Future research needs to expand the focus beyond motor vehicle accidents to include other injury mechanisms such as falls, suicide, self-harm, and drowning. Additionally, integrating diverse forecasting methods alongside Exponential Smoothing Models will enhance predictive capabilities; this can be achieved through Python implementation and leveraging new generative artificial intelligence (AI) technologies. Conducting longitudinal studies will help identify trends related to demographic factors influencing motor vehicle fatalities over time. Furthermore, evaluating the effectiveness of interventions aimed at high-risk groups identified in this research is crucial while investigating the underlying behavioral factors contributing to these incidents.

Public Health Implications

Considering the alarming statistic of reported crashes over four years, focusing on targeted educational campaigns aimed at young male drivers by enhancing driver education programs emphasizing safe driving practices is crucial. Additionally, prioritizing traffic safety improvements in urban planning is essential; this includes redesigning roadways to ensure safer pedestrian crossings and implementing traffic calming measures. Seasonal safety campaigns should be launched during peak travel months, particularly around holidays or events known for increased travel. Furthermore, fostering community engagement through outreach programs will promote a collective responsibility toward road safety, encouraging all community members to contribute to safer driving environments.

By addressing these recommendations while focusing on areas needing improvement, stakeholders can work collaboratively toward significantly reducing motor vehicle fatalities, ultimately addressing the public health mission and outcomes of safeguarding the nation’s health, safety, and security. Targeted interventions should focus on young adult males aged 25–44, mainly white individuals in large central metro areas, with enhanced enforcement, education, and prevention campaigns intensified during the high-risk months of August to October, also improved road signage, better lighting, and changes in road layout.

Funding source

This dissertation did not receive any specific grant from funding agencies in the commercial or not-for-profit sectors, except for the CDC partial tuition assistance for Public Health courses.

Declaration of Interest

None.

Disclaimer

References

Agyemang, E. F., Mensah, J. A., Ocran, E., Opoku, E., & Nortey, E. N. (2023). Time series based road traffic accidents forecasting via SARIMA and Facebook Prophet model with potential changepoints. *Heliyon*, *9*(12). https://doi.org/10.1016/j.heliyon.2023.e22544

Agresti, A. (2012). *Categorical data analysis* (Vol. 792). John Wiley & Sons. https://www.google.com/books/edition/Categorical_Data_Analysis/

Ahmad, F. B., & Cisewski, J. A. (2023). Quarterly provisional estimates for selected indicators of mortality, 2022-Quarter 1, 2023. *National Center for Health Statistics. National Vital Statistics System, Vital Statistics Rapid Release Program*. Last Reviewed: 17 October 2023. https://www.cdc.gov/nchs/nvss/vsrr/mortality-dashboard.htm#

Akhtar, M., & Moridpour, S. (2021). A review of traffic congestion prediction using artificial intelligence. *Journal of Advanced Transportation*, *2021*, 1-18. https://pdfs.semanticscholar.org/7222/2ce8644fd1af7bd298ab7efb35438aa457d8.pdf

Alizadeh, M., Rahimi, S., & Ma, J. (2022). A hybrid ARIMA–WNN approach to model vehicle operating behavior and detect unhealthy states. *Expert Systems with Applications*, *194*, 116515. https://doi.org/10.1016/j.eswa.2022.116515

Alnawmasi, N., & Mannering, F. (2022). A temporal assessment of distracted driving injury severities using alternate unobserved-heterogeneity modeling approaches. *Analytic methods in accident research*, *34*, 100216. https://doi.org/10.1016/j.amar.2022.100216

Andreoni, A., & Postorino, M. N. (2006). A multivariate ARIMA model to forecast air transport demand. *Proceedings of the Association for European Transport and Contributors*, 1-14. https://citeseerx.ist.psu.edu/document

Arias, E., Tejada-Vera, B., & Ahmad, F. (2021). Provisional life expectancy estimates for January through June 2020. https://dx.doi.org/10.15620/cdc:100392

Beck, L. F., Kresnow, M. J., & Bergen, G. (2019). Belief about seat belt use and seat belt wearing behavior among front and rear seat passengers in the United States. *Journal of safety research*, *68*, 81-88. https://doi.org/10.1016/j.jsr.2018.12.007

Beck, L.F., Downs, J., Stevens, M.R., & Sauber-Schatz, E.K. (2017). Rural and urban differences in passenger-vehicle–occupant deaths and seat belt use among adults—United States, 2014. *MMWR. Surveillance Summaries*, *66*. https://www.cdc.gov/mmwr/volumes/66/ss/ss6617a1.htm

Bheemanna, M. N., & Budihal, H. H. (2023). Forecast of demographic variables using the ARIMA Model in India. https://doi.org/10.22271/maths.2023.v8.i5Sn.1322

Boal WL, Li J, Rodriguez-Acosta RL (2016). Seat Belt Use Among Adult Workers — 21 States, 2013. MMWR Morb Mortal Wkly Rep 2016;65:593–597. DOI: http://dx.doi.org/10.15585/mmwr.mm6523a1

Box, G. E. P., & Jenkins, G. M. (1970). *Time series Analysis.* Holden-Day, San Francisco, *Calif*. https://doi.org/10.1111/jtsa.12194

Box, G. E., & Jenkins, G. M. (1976). Time series analysis: Forecasting and control. *San Francisco, CA: Holden Day*, *3226*(3228), 10. https://www.scirp.org/reference/referencespapers?referenceid=1969833

Box, G. E., & Pierce, D. A. (1970). Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. *Journal of the American Statistical Association, 65*(332), 1509-1526. https://www.tandfonline.com/doi/abs/10.1080/01621459.1970.10481180

Box, G. E., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time series analysis: forecasting and control*. John Wiley & Sons. https://www.google.com/books/edition/Time_Series_Analysis/

Brocklebank, J. C., Dickey, D. A., Choi, B. (2018). *SAS for forecasting time series*, Third Edition. United States: SAS Institute. https://support.sas.com/content/dam/SAS/support/en/books/sas-for-forecasting-time-series-third-edition/68469_excerpt.pdf

Brooks, C. (2021). Medical examiner and coroner offices, 2018. *Bureau of Justice Statistics*. https://bjs.ojp.gov/sites/g/files/xyckuh236/files/media/document/meco18.pdf

Cameron, A. C., & Trivedi, P. K. (2013). *Regression analysis of count data* (No. 53). Cambridge university press. https://books.google.com/books/Regression

Centers for Disease Control and Prevention (CDC). (2024a). *CDC moving forward*. https://intranet.cdc.gov/cdc-moving-forward/index.html

Centers for Disease Control and Prevention (CDC). (2024b). *Transportation safety*. https://www.cdc.gov/transportationsafety/

Centers for Disease Control and Prevention (CDC). (2025a). *Wide-ranging online data for epidemiologic research (WONDER)*. https://wonder.cdc.gov/mcd-icd10-provisional.html

Centers for Disease Control and Prevention (CDC). (2025b). *Web-based Injury Statistics Query and Reporting System (WISQARS)*. https://wisqars.cdc.gov

Chan, T. K., Chin, C. S., Chen, H., & Zhong, X. (2019). A comprehensive review of driver behavior analysis utilizing smartphones. *IEEE Transactions on Intelligent Transportation Systems*, *21*(10), 4444-4475. https://doi.org/10.1109/TITS.2019.2940481

Cullen, P., Möller, H., Woodward, M., Senserrick, T., Boufous, S., Rogers, K., ... & Ivers, R. (2021). Are there sex differences in crash and crash-related injury between men and women? A 13-year cohort study of young drivers in Australia. *SSM-population health*, *14*, 100816. https://www.sciencedirect.com/science/article/pii/S2352827321000914

Dadashova, B., Arenas-Ramírez, B., Mira-McWilliams, J., & Aparicio-Izquierdo, F. (2016). Methodological development for selection of significant predictors explaining fatal road accidents. *Accident Analysis & Prevention*, *90*, 82-94. https://doi.org/10.1016/j.aap.2016.02.003

Deretić, N., Stanimirović, D., Awadh, M. A., Vujanović, N., & Djukić, A. (2022). SARIMA modeling approach for forecasting of traffic accidents. *Sustainability*, *14*(8), 4403. https://doi.org/10.3390/su14084403

Dickey, D. (2019). *The best of both worlds: Forecasting using time series with inputs*, NC State University. https://www.sas.com/content/dam/SAS/support/en/sas-global-forum-proceedings/2019/3559-2019.pdf

Dickey, D. A., & Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, *74*(366a), 427-431. https://doi.org/10.1080/01621459.1979.10482531

Doğan, E., & Akgüngör, A. P. (2013). Forecasting highway casualties under the effect of railway development policy in Turkey using artificial neural networks. *Neural Computing and Applications*, *22*, 869-877. https://link.springer.com/article/10.1007/s00521-011-0778-0

Eberhardt, M. S., Ingram, D. D., & Makuc, D. M. (2001). Urban and rural health chartbook: Health, United States, 2001. https://stacks.cdc.gov/view/cdc/5363

Ehlman, D. C., Haileyesus, T., Lee, R., Ballesteros, M. F., & Yard, E. (2021). Evaluation of the National Electronic Injury Surveillance System–All injury program's self-directed violence data, United States, 2018. *Journal of Safety Research*, *76*, 327-331. https://www.sciencedirect.com/science/article/pii/S0022437520301420?via%3Dihub

Faust, J. S., Du, C., Mayes, K. D., Li, S. X., Lin, Z., Barnett, M. L., & Krumholz, H. M. (2021). Mortality from drug overdoses, homicides, unintentional injuries, motor vehicle crashes, and suicides during the pandemic, March-August 2020. *Jama*, *326*(1), 84-86. https://jamanetwork.com/journals/jama/fullarticle/2780436.

Fliss, M. D., Baumgartner, F., Delamater, P., Marshall, S., Poole, C., & Robinson, W. (2020). Re-prioritizing traffic stops to reduce motor vehicle crash outcomes and racial disparities. *Injury epidemiology*, *7*, 1-15. https://link.springer.com/article/10.1186/s40621-019-0227-6

Fridstrøm, L., Ifver, J., Ingebrigtsen, S., Kulmala, R., & Thomsen, L. K. (1995). Measuring the contribution of randomness, exposure, weather, and daylight to the variation in road accident counts. *Accident Analysis & Prevention*, *27*(1), 1-20. https://doi.org/10.1016/0001-4575(94)E0023-EGet rights and content

Gao, J., & Davis, G. A. (2017). Using naturalistic driving study data to investigate the impact of driver distraction on driver's brake reaction time in freeway rear-end events in car-following situations. *Journal of safety research*, *63*, 195-204. https://doi.org/10.1016/j.jsr.2017.10.012Get rights and content

German, R. R., Horan, J. M., Lee, L. M., Milstein, B., & Pertowski, C. A. (2001). Updated guidelines for evaluating public health surveillance systems; recommendations from the Guidelines Working Group. https://stacks.cdc.gov/view/cdc/13376

Guilcher, S. J., Craven, B. C., McColl, M. A., Lemieux-Charles, L., Casciaro, T., & Jaglal, S. B. (2012). Application of the Andersen's health care utilization framework to secondary complications of spinal cord injury: a scoping review. *Disability and rehabilitation*, *34*(7), 531-541. https://doi.org/10.3109/09638288.2011.608150

Gultekin, N., & Acik Kemaloglu, S. (2023). Evaluation of the impact of COVID-19 on air traffic volume in Turkish airspace using artificial neural networks and time series. *Scientific reports*, *13*(1), 6551. https://doi.org/10.1038/s41598-023-33784-x

Haileyesus, T., Annest, J. L., & Dellinger, A. M. (2007). Cyclists injured while sharing the road with motor vehicles. *Injury prevention*, *13*(3), 202-206. https://doi.org/10.1136/ip.2006.014019

Hao, J., & Liu, F. (2024). Improving long-term multivariate time series forecasting with a seasonal-trend decomposition-based 2-dimensional temporal convolution dense network. *Scientific Reports*, *14*(1), 1689. https://doi.org/10.1038/s41598-024-52240-y

Higgs, B., & Abbas, M. (2014). Segmentation and clustering of car-following behavior: Recognition of driving patterns. *IEEE Transactions on Intelligent Transportation Systems*, *16*(1), 81-90. https://ieeexplore.ieee.org/abstract/document/6832510

Hossain, M., Abdel-Aty, M., Quddus, M. A., Muromachi, Y., & Sadeek, S. N. (2019). Real-time crash prediction models: State-of-the-art, design pathways and ubiquitous requirements. *Accident Analysis & Prevention*, *124*, 66-84. https://doi.org/10.1016/j.aap.2018.12.022

Hourdos, J. N. (2005). *Crash prone traffic flow dynamics: Identification and real-time detection*. University of Minnesota. https://www.proquest.com/dissertations-theses/crash-prone-traffic-flow-dynamics-identification/docview/305431650/se-2?accountid=26724

Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: principles and practice*. OTexts. https://www.google.com/books/edition/Forecasting_principles_and_practice

Ihueze, C. C., & Onwurah, U. O. (2018). Road traffic accidents prediction modelling: An analysis of Anambra State, Nigeria. *Accident analysis & prevention*, *112*, 21-29. https://doi.org/10.1016/j.aap.2017.12.016

Iio, K., Guo, X., & Lord, D. (2021). Examining driver distraction in the context of driving speed: An observational study using disruptive technology and naturalistic data. *Accident Analysis & Prevention*, *153*, 105983. https://doi.org/10.1016/j.aap.2021.105983Get rights and content

Inada, H., Li, Q., Bachani, A., & Hyder, A. A. (2020). Forecasting global road traffic injury mortality for 2030. *Injury prevention*, *26*(4), 339-343. http://dx.doi.org/10.1136/injuryprev-2019-043336

Institutional Review Board (IRB). (2015). *Determination of human subject's research*. University of Georgia UGA). https://research.uga.edu/docs/policies/compliance/hso/IRB-Determination-of-Human-Subject-Research.pdf

Ismail, M. T., Shah, N. Z. A., & Karim, S. A. A. (2021). Modeling solar radiation in peninsular Malaysia using ARIMA model. *Clean Energy Opportunities in Tropical Countries*, 53-71. https://link.springer.com/chapter/10.1007/978-981-15-9140-2_3/figures/2

Jahangiri, A., Rakha, H. A., & Dingus, T. A. (2015). Developing a system architecture for cyclist violation prediction models incorporating naturalistic cycling data. *Procedia Manufacturing*, *3*, 5543-5550. https://doi.org/10.1016/j.promfg.2015.07.724

Jahangiri, A., Rakha, H., & Dingus, T. A. (2016). Red-light running violation prediction using observational and simulator data. *Accident Analysis & Prevention*, *96*, 316-328. https://doi.org/10.1016/j.aap.2016.06.009

Jerrhag, D., Englund, M., Karlsson, M. K., & Rosengren, B. E. (2017). Epidemiology and time trends of distal forearm fractures in adults - a study of 11.2 million person-years in Sweden. *BMC Musculoskeletal Disorders*, *18*, 1-8. https://link.springer.com/article/10.1186/s12891-017-1596-z

Kheirabadi, G. R., & Bolhari, J. (2012). Role of human factors in road accidents. https://www.sid.ir/paper/106557/en

Klauer, C., Dingus, T. A., Neale, V. L., Sudweeks, J. D., & Ramsey, D. J. (2006). The impact of driver inattention on near-crash/crash risk: An analysis using the 100-car naturalistic driving study data. https://vtechworks.lib.vt.edu/items/0d2fb484-ddf4-4e63-82b2-7110afc984a9

Ko, T. *Analysis and forecasting of road traffic accidents in Yangon municipal area.* (2019). (Doctoral dissertation, MERAL Portal). https://meral.edu.mm/record/9344/files/Thiri%20Ko,%20M.Econ.%20Stats%20Roll.1,%202019.pdf

Lazarus, J. V., Ortiz, A., Tyrovolas, S., Fernández, E., Guy, D., White, T. M., ... & Soriano, J. B. (2022). A GBD 2019 study of health and Sustainable Development Goal gains and forecasts to 2030 in Spain. *Scientific Reports*, *12*(1), 21154. https://doi.org/10.1038/s41598-022-24719-z

Lombardi, D. A., Horrey, W. J., & Courtney, T. K. (2017). Age-related differences in fatal intersection crashes in the United States. *Accident Analysis & Prevention*, *99*, 20-29. https://www.sciencedirect.com/science/article/pii/S0001457516303918

Luo, H., Lau, K. K., Wong, G. H., Chan, W. C., Mak, H. K., Zhang, Q., Knapp, M., & Wong, I. C. (2020). Predicting dementia diagnosis from cognitive footprints in electronic health records: A case-control study protocol. https://doi.org/10.1136/bmjopen-2020-043487

McDonald, C. C., Ward, K., Huang, Y., Wiebe, D. J., & Delgado, M. K. (2019). Novel smartphone-based measures of cell phone use while driving in a sample of newly licensed adolescent drivers. *Health Education & Behavior*, *46*(1), 10-14. https://journals.sagepub.com/doi/full/10.1177/1090198118788612

Medeiros, A. L. D. S., & Nadanovsky, P. (2016). Car and motorcycle deaths: an evolutionary perspective. *Ciencia & saude coletiva*, *21*, 3691-3702. https://www.scielosp.org/article/csc/2016.v21n12/3691-3702/

Miyajima, C., Nishiwaki, Y., Ozawa, K., Wakita, T., Itou, K., Takeda, K., & Itakura, F. (2007). Driver modeling based on driving behavior and its evaluation in driver identification. *Proceedings of the IEEE*, *95*(2), 427-437. https://doi.org/10.1109/JPROC.2006.888405

Nagy, A. M., & Simon, V. (2018). Survey on traffic prediction in smart cities. *Pervasive and Mobile Computing*, *50*, 148-163. https://doi.org/10.1016/j.pmcj.2018.07.004

Nassiri, H., Mohammadpour, S. I., & Dahaghin, M. (2023). Forecasting time trends of fatal motor vehicle crashes in Iran using an ensemble learning algorithm. *Traffic injury prevention*, *24*(1), 44-49. https://doi.org/10.1080/15389588.2022.2130279

National Highway Traffic Safety Administration (NHTSA). (2020). *Early estimates of motor vehicle traffic fatalities and fatality rate by sub-categories through June 2020*. https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/813581

National Highway Traffic Safety Administration (NHTSA). (2024). *NHTSA Launches Put the Phone Away or Pay Campaign; Releases 2023 Fatality Early Estimates*. https://www.nhtsa.gov/press-releases/2022-traffic-deaths-2023-early-estimates

Naumann, R. B., & Beck, L. F. (2013). Motor vehicle traffic-related pedestrian deaths—United States, 2001–2010. *Morbidity and Mortality Weekly Report*, *62*(15), 277. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4604973/

Neupane, B., Rijal, S., Gc, S., & Basnet, T. B. (2021). A multilevel analysis to determine the factors associated with institutional delivery in Nepal: further analysis of Nepal demographic and health survey 2016. *Health Services Insights*, *14*, 11786329211024810. https://www.researchgate.net/figure/Conceptual-framework-of-Andersens-behavior-model-Abbreviation-ANC-antenatal-care_fig1_352395637

Ng, S. H., Giles, H., & Moody, J. (1991). Information-seeking triggered by age. *The International Journal of Aging and Human Development*, *33*(4), 269-277. https://journals.sagepub.com/doi/abs/10.2190/1uv0-uyql-7aqv-v6un

Pandey, A. R., Chalise, B., Shrestha, N., Ojha, B., Maskey, J., Sharma, D., ... & Aryal, K. K. (2020). Mortality and risk factors of disease in Nepal: Trend and projections from 1990 to 2040. *PloS one*, *15*(12), e0243055. https://doi.org/10.1371/journal.pone.0243055

Papadoulis, A., Quddus, M., & Imprialou, M. (2019). Evaluating the safety impact of connected and autonomous vehicles on motorways. *Accident Analysis & Prevention*, *124*, 12-22. https://doi.org/10.1016/j.aap.2018.12.019

Pollack, K. M., Gielen, A. C., Mohd Ismail, M. N., Mitzner, M., Wu, M., & Links, J. M. (2014). Investigating and improving pedestrian safety in an urban environment. *Injury epidemiology*, *1*, 1-9. https://link.springer.com/article/10.1186/2197-1714-1-11

Raifman, M. A., & Choma, E. F. (2022). Disparities in activity and traffic fatalities by race/ethnicity. American journal of preventive medicine, 63(2), 160-167. https://doi.org/10.1016/j.amepre.2022.03.012

Ren, H., Li, J., Yuan, Z. A., Hu, J. Y., Yu, Y., & Lu, Y. H. (2013). The development of a combined mathematical model to forecast the incidence of hepatitis E in Shanghai, China. *BMC Infectious Diseases, 13, 1-6*. https://link.springer.com/article/10.1186/1471-2334-13-421

Ridel, D., Rehder, E., Lauer, M., Stiller, C., & Wolf, D. (2018, November). A literature review on the prediction of pedestrian behavior in urban scenarios. *In 2018 21st International Conference on Intelligent Transportation Systems (ITSC)* (pp. 3105-3112). IEEE. https://doi.org/10.1109/ITSC.2018.8569415

Rosenberg, H. M. (1999). *Quality of death rates by race and Hispanic-origin: a summary of current research, 1999* (Vol. 128). National Center for Health Statistics. https://www.google.com/books/edition/Quality_of_Death_Rates_by_Race_and_Hispa/

Samuel, S., Yamani, Y., & Fisher, D. L. (2016). Large reductions are possible in older driver crashes at intersections. *Clinical and experimental optometry*, *99*(5), 419-424. https://www.tandfonline.com/doi/abs/10.1111/cxo.12443

SAS Institute Inc. (2012). Statistics 2: ANOVA and Regression. Cary, NC: Course Notes. https://learn.sas.com/course/view.php?name=STAT2

SAS Institute Inc. (2018). SAS/ETS® 15.1 User's Guide. Cary, NC: The ARIMA Procedure. https://support.sas.com/documentation/onlinedoc/ets/151/arima.pdf

Scott, J. (2021). Data Science in R: A Gentle Introduction. *The R Journal*, *1*, 155-187. https://bookdown.org/jgscott/DSGI/large-sample-inference.html

Shaw, K. M., West, B., Kendi, S., Zonfrillo, M. R., & Sauber-Schatz, E. (2022). Urban and rural child deaths from motor vehicle crashes: United States, 2015-2019. *The Journal of Pediatrics*, *250*, 93-99. https://www.sciencedirect.com/science/article/pii/S0022347622006205

Sohail, A., Cheema, M. A., Ali, M. E., Toosi, A. N., & Rakha, H. A. (2023). Data-driven approaches for road safety: A comprehensive systematic literature review. *Safety Science*, *158*, 105949. https://doi.org/10.1016/j.ssci.2022.105949

Spencer, M. R., Hedegaard, H., & Garnet, M. (2021). Motor vehicle traffic death rates by sex, age group, and road-user type: United States, 1999-2019. https://www.cdc.gov/nchs/products/databriefs/db400.htm

Spencer, M., & Ahmad, F. (2017). Timeliness of death certificate data for mortality surveillance and provisional estimates. National Center for Health Statistics. January 2017. https://www.cdc.gov/nchs/data/vsrr/report001.pdf

Storie, V. J. (1977). *Male and female car drivers: differences observed in accidents* (No. TRRL Lab Rept. 761). https://trid.trb.org/View/55988

Swedo, E. A., Alic, A., Law, R. K., Sumner, S. A., Chen, M. S., Zwald, M. L., ... & Mercy, J. A. (2023). Development of a machine learning model to estimate US firearm homicides in near real time. *JAMA Network Open*, *6*(3), e233413-e233413. https://jamanetwork.com/journals/jamanetworkopen/article-abstract/2802555

Tasca, L. (2000). A review of the literature on aggressive driving research. https://www.stopandgo.org/research/aggressive/tasca.pdf

Teymuri, G. H., Bahmani, R., Asghari, M., Madrese, E., Rahmani, A., Abbasinia, M., ... & Samavati, M. (2014). Survey of accidents in suburban Tehran and the prediction of future events based on a time-series model. *Electronic Physician*, *6*(1), 754. https://doi.org/10.14661%2F2014.754-758

Thompson, C., White, J., Dougherty, B., Albright, A., & Schmidt, D. C. (2010). Using smartphones to detect car accidents and provide situational awareness to emergency responders. *In Mobile Wireless Middleware, Operating Systems, and Applications: Third International Conference, Mobilware 2010, Chicago, IL, USA, June 30-July 2, 2010. Revised Selected Papers 3* (pp. 29-42). Springer Berlin Heidelberg. https://link.springer.com/chapter/10.1007/978-3-642-17758-3_3

Torres, P., Romano, E., Voas, R. B., De La Rosa, M., & Lacey, J. H. (2014). The relative risk of involvement in fatal crashes as a function of race/ethnicity and blood alcohol concentration. *Journal of safety research*, *48*, 95-101. https://www.sciencedirect.com/science/article/pii/S0022437513001746

United States Department of Transportation (DOT). (2024a). Federal Highway Administration. *Travel monitoring*. Accessed 6 June 2024. https://www.fhwa.dot.gov/policyinformation/travel_monitoring/tvt.cfm

United States Department of Transportation (DOT). (2024b). National Highway Traffic Safety Administration. *Drunk driving*. Accessed 6 June 2024. https://www.nhtsa.gov/risky-driving/drunk-driving

Wegman, F., Watson, B., Wong, S. V., Job, S., & Segui-Gomez, M. (2017). Road safety in Myanmar. *Recommendations of an Expert Mission invited by the Government of Myanmar and supported by the Suu Foundation., Paris, FIA*. https://www.fia.com/sites/default/files/myanmar_report.pdf

Wei, W. W. (2013). Time series analysis. https://doi.org/10.1093/oxfordhb/9780199934898.013.0022

Wei, W. W. (2018). *Multivariate time series analysis and applications*. John Wiley & Sons. https://www.google.com/books/edition/Multivariate_Time_Series

West, B. A., & Naumann, R. B. (2013). Motor Vehicle–Related Deaths—United States, 2005 and 2009. *CDC Health Disparities and Inequalities Report—United States, 2013*, *62*(3), 176.https://www.cdc.gov/mmwr/pdf/other/su6203.pdf?utm_source=guelphtoday.com&utm_campaign=guelphtoday.com%3A%20outbound&utm_medium=referral#page=178

World Health Organization (WHO). (2020). *WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020*. Accessed 6 June 2024. https://www.who.int/director-general

World Health Organization (WHO). (2024). *Global status report on road safety 2023*. https://www.who.int/publications/i/item/9789240086517

Xu, J., Murphy, S. L., Kochanek, K. D., & Arias, E. (2021). Deaths: Final data for 2019. https://dx.doi.org/10.15620/cdc:106058

Zhang, J., Liu, X. M., He, Y. L., & Chen, Y. S. (2007). Application of ARIMA model in forecasting traffic accidents. *Journal of Beijing University of Technology*, *33*(12), 1295-1299. https://www.researchgate.net/publication/291128888_Application_of_ARIMA_model_in_forecasting_traffic_accidents

Zsifkovits, M., & Pham, T. S. (2017). Modeling and parameterizing pedestrian behaviour in public places: A review. *International Journal of Simulation Modelling*, *16*(4), 630-643. 630-643 https://doi.org/10.2507/IJSIMM16(4)6.399

Figures

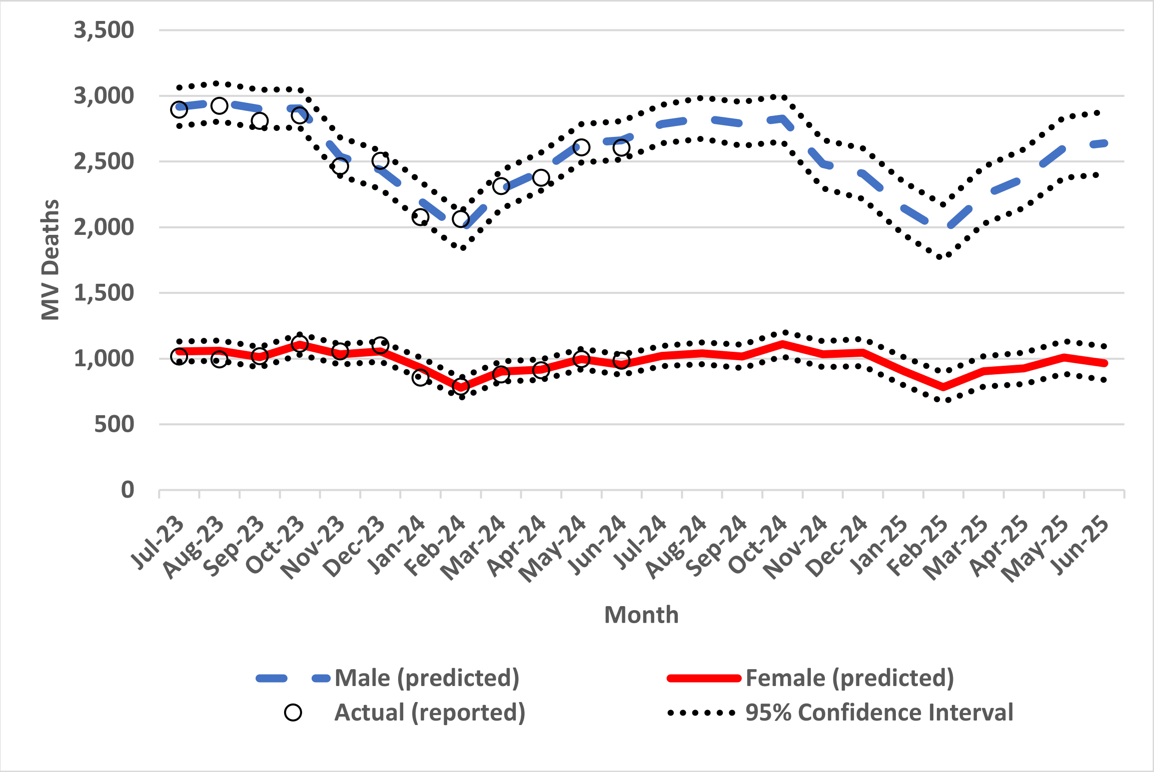


Figure 3: Motor Vehicle Deaths and Forecasts by Sex until June 2025, United States

Figure 3 depicts the actual reported MV deaths (circled) and the predicted death counts (Male dashed lines and female solid lines). The dotted trend lines represent the 95% confidence limits around the forecast. The predicted crash counts were derived from the actual historical data. The forecasts appear reasonable, given that "forecast intervals are fairly narrow" (Brocklebank et al., 2018, p. 227).

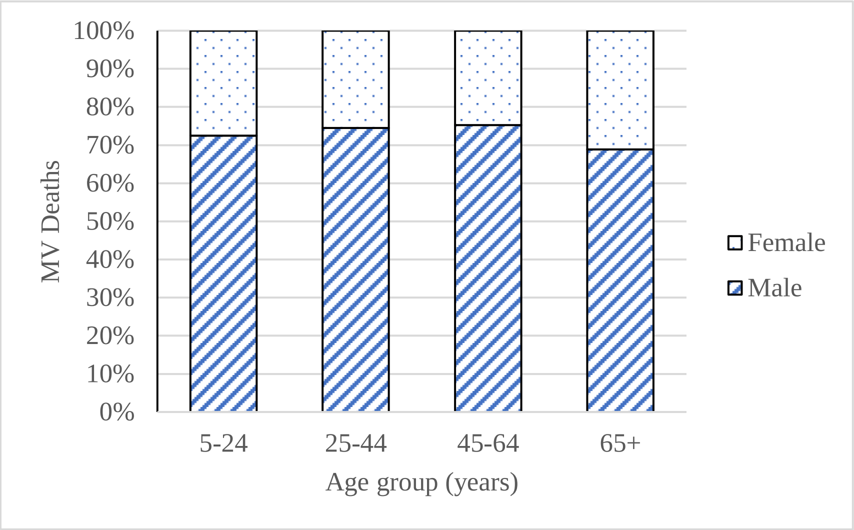


Figure 4: Motor Vehicle Deaths by Sex and Age, United States, Jul 2020-Jun 2024

Figure 4 shows that males consistently have a higher number of reported deaths, with the gap slightly decreasing as age increases to the age 65+ years group (i.e., from 75.2% to 68.9%)

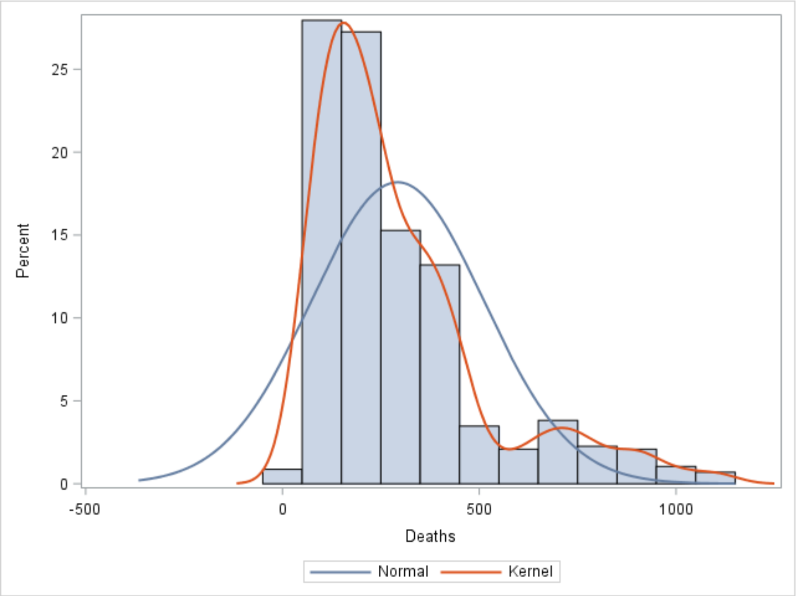


Figure 5: Histogram of Motor Vehicle Deaths, United States, Jul 2020-Jun 2024

Figure 5 shows that the crash data were skewed to the right; the kernel and normal density superimposed on the histogram indicate that the death counts do not follow a normal distribution (i.e., misaligned curves).

Tables (Study 1)

Table 1. Exponential Smoothing Model for Motor Vehicle Deaths, United States, July 2023-June 2025

Reported Deaths and Forecasts by Month

Month	**Reported (y)**	**Predicted**	**Error (e)**	**|e|**	**|e/y|**
Jul-23	3,912	3,969.4	-57.4	57.4	0.015
Aug-23	3,920	4,010.8	-90.8	90.8	0.023
Sep-23	3,829	3,907.8	-78.8	78.8	0.021
Oct-23	3,964	4,008.6	-44.6	44.6	0.011
Nov-23	3,520	3,567.4	-47.4	47.4	0.013
Dec-23	3,605	3,488.0	117.0	117.0	0.032
Jan-24	2,933	3,134.7	-201.7	201.7	0.069
Feb-24	2,850	2,746.0	104.0	104.0	0.037
Mar-24	3,192	3,188.5	3.5	3.5	0.001
Apr-24	3,287	3,340.9	-53.9	53.9	0.016
May-24	3,604	3,635.9	-31.9	31.9	0.009
Jun-24	3,590	3,614.4	-24.4	24.4	0.007
Total	42,206	42,612.3	-406.3	855.5	0.2540

Forecasts until June 2025

Month	**Predicted**	**LL 95% CI**	**UL 95% CI**	**SE**
Jul-24	3,804	3,603	4,005	102.5903
Aug-24	3,868	3,651	4,086	110.7995
Sep-24	3,802	3,570	4,035	118.4410
Oct-24	3,935	3,689	4,182	125.6186
Nov-24	3,512	3,253	3,772	132.4076
Dec-24	3,452	3,180	3,725	138.8651
Jan-25	3,051	2,767	3,335	145.0354
Feb-25	2,745	2,449	3,041	150.9537
Mar-25	3,145	2,838	3,452	156.6485
Apr-25	3,296	2,978	3,614	162.1435
May-25	3,613	3,285	3,941	167.4582
Jun-25	3,604	3,266	3,943	172.6094
Total	**41,829**	**39,366**	**44,292**	**1,256.77**

Mean Absolute Percentage Error (MAPE) = Sum |e/y|/12 i.e., average forecast accuracy = *98%*.

Abbreviations: SE = Standard Error; CI = confidence interval; LL = lower limit; UL = upper limit.

Table 2. Forecasting Motor Vehicle Deaths by Sex and Month, United States, July 2024-June 2025 Using Exponential Smoothing Model and July 2020-June 2024 data (48 months)

Sex: Female, Mean Absolute Percentage Error (MAPE), i.e., average forecast accuracy = 97%

Month	**Predicted**	**LL 95% CI**	**UL 95% CI**	**SE**
Jul-24	1,020	943	1,097	39.3813
Aug-24	1,041	958	1,124	42.3493
Sep-24	1,017	929	1,105	45.1224
Oct-24	1,110	1,017	1,204	47.7347
Nov-24	1,034	935	1,132	50.2113
Dec-24	1,046	943	1,149	52.5714
Jan-25	906	798	1,013	54.8300
Feb-25	783	671	894	56.9991
Mar-25	903	788	1,019	59.0887
Apr-25	927	807	1,046	61.1068
May-25	1,009	885	1,132	63.0604
Jun-25	966	839	1,093	64.9553
Total	**11,761**	**10,837**	**12,685**	**471.503**

Sex: Male, (MAPE), i.e., average forecast accuracy = 98%

Month	**Predicted**	**LL 95% CI**	**UL 95% CI**	**SE**
Jul-24	2,785	2,639	2,931	74.4628
Aug-24	2,829	2,672	2,985	79.8288
Sep-24	2,787	2,621	2,953	84.8562
Oct-24	2,827	2,651	3,002	89.6019
Nov-24	2,480	2,296	2,665	94.1086
Dec-24	2,408	2,215	2,601	98.4092
Jan-25	2,147	1,946	2,348	102.5295
Feb-25	1,964	1,755	2,173	106.4905
Mar-25	2,243	2,027	2,459	110.3094
Apr-25	2,371	2,148	2,594	114.0004
May-25	2,606	2,375	2,836	117.5756
Jun-25	2,640	2,403	2,877	121.0453
Total	**30,087**	**28,369**	**31,805**	**876.522**

Abbreviations: SE = Standard Error; CI = confidence interval; LL = lower limit; UL = upper limit.

Table 3. Forecasting Motor Vehicle Deaths by Race and Month, United States, July 2024-June 2025 Using Exponential Smoothing Model and July 2020-June 2024 data (48 months)

Race: American Indian or Alaska Native (MAPE), i.e., average forecast accuracy = 84%

Month	**Predicted**	**LL 95% CI**	**UL 95% CI**	**SE**
Jul-24	70	51	89	9.6416
Aug-24	69	50	89	9.8813
Sep-24	69	49	89	10.1153
Oct-24	72	52	92	10.3440
Nov-24	57	37	78	10.5678
Dec-24	59	38	80	10.7870
Jan-25	55	33	76	11.0017
Feb-25	48	26	70	11.2124
Mar-25	58	35	80	11.4192
Apr-25	55	32	78	11.6222
May-25	66	43	89	11.8219
Jun-25	59	35	83	12.0181
Total	**737**	**582**	**892**	**78.972**

Race: Asian (MAPE), i.e., average forecast accuracy = 91%

Month	**Predicted**	**LL 95% CI**	**UL 95% CI**	**SE**
Jul-24	75	58	92	8.5804
Aug-24	83	66	100	8.7706
Sep-24	86	69	104	8.9568
Oct-24	85	67	103	9.1391
Nov-24	91	72	109	9.3179
Dec-24	92	74	111	9.4934
Jan-25	73	54	92	9.6656
Feb-25	67	48	86	9.8349
Mar-25	79	59	98	10.0012
Apr-25	73	53	93	10.1649
May-25	77	56	97	10.3259
Jun-25	73	53	94	10.4845
Total	**954**	**821**	**1,087**	**67.904**

Abbreviations: SE = Standard Error; CI = confidence interval; LL = lower limit; UL = upper limit.

Table 3. Forecasting Motor Vehicle Deaths by Race and Month, United States, July 2024-June 2025 Using Exponential Smoothing Model and July 2020-June 2024 data (48 months)

Race: Black or African American (MAPE), i.e., average forecast accuracy = 96%

Month	**Predicted**	**LL 95% CI**	**UL 95% CI**	**SE**
Jul-24	676	610	742	33.6341
Aug-24	678	601	756	39.5821
Sep-24	674	587	762	44.7463
Oct-24	683	586	779	49.3733
Nov-24	620	515	725	53.6023
Dec-24	635	522	747	57.5213
Jan-25	582	462	702	61.1898
Feb-25	502	375	629	64.6504
Mar-25	573	440	706	67.9350
Apr-25	600	461	739	71.0679
May-25	657	511	802	74.0685
Jun-25	593	442	744	76.9521
Total	**7,474**	**6,353**	**8,594**	**571.473**

Race: White (MAPE), i.e., average forecast accuracy = 97%

Month	**Predicted**	**LL 95% CI**	**UL 95% CI**	**SE**
Jul-24	2,930	2,765	3,095	84.0771
Aug-24	2,990	2,817	3,164	88.7295
Sep-24	2,927	2,744	3,109	93.1498
Oct-24	3,046	2,856	3,237	97.3696
Nov-24	2,703	2,504	2,902	101.4139
Dec-24	2,627	2,420	2,833	105.3031
Jan-25	2,303	2,090	2,517	109.0537
Feb-25	2,095	1,875	2,316	112.6795
Mar-25	2,397	2,169	2,625	116.1921
Apr-25	2,528	2,293	2,762	119.6017
May-25	2,760	2,519	3,001	122.9168
Jun-25	2,836	2,588	3,083	126.1447
Total	**32,142**	**30,382**	**33,902**	**897.947**

Abbreviations: SE = Standard Error; CI = confidence interval; LL = lower limit; UL = upper limit.

Table 3. Forecasting Motor Vehicle Deaths by Race and Month, United States, July 2024-June 2025 Using Exponential Smoothing Model and July 2020-June 2024 data (48 months)

Race: More than one race (MAPE), i.e., average forecast accuracy = 93%

Month	**Predicted**	**LL 95% CI**	**UL 95% CI**	**SE**
Jul-24	55	45	64	4.8025
Aug-24	47	37	56	4.8709
Sep-24	49	39	58	4.9383
Oct-24	52	43	62	5.0049
Nov-24	47	37	57	5.0705
Dec-24	45	35	55	5.1354
Jan-25	43	33	53	5.1994
Feb-25	38	28	48	5.2626
Mar-25	41	30	51	5.3251
Apr-25	44	33	54	5.3868
May-25	58	47	68	5.4479
Jun-25	48	38	59	5.5083
Total	**566**	**500**	**632**	**33.573**

Abbreviations: SE = Standard Error; CI = confidence interval; LL = lower limit; UL = upper limit.

Race: Native Hawaiian or Other Pacific*

*****Death counts of less than ten were not reported because of statistical reliability and data confidentiality.

Table 4. Forecasting Motor Vehicle Deaths by Hispanic Origin and Month, United States, July 2024-June 2025 Using Exponential Smoothing Model and July 2020-June 2024 data (48 months)

Hispanic Origin: Hispanic or Latino (MAPE), i.e., average forecast accuracy = 97%

Month	**Predicted**	**LL 95% CI**	**UL 95% CI**	**SE**
Jul-24	691	632	750	30.1524
Aug-24	706	645	767	31.1407
Sep-24	686	623	749	32.0985
Oct-24	757	693	822	33.0285
Nov-24	700	634	767	33.9331
Dec-24	689	620	757	34.8142
Jan-25	644	574	714	35.6735
Feb-25	591	519	662	36.5126
Mar-25	620	547	694	37.3328
Apr-25	658	583	732	38.1354
May-25	708	632	784	38.9215
Jun-25	692	614	770	39.6920
Total	**8,142**	**7,615**	**8,670**	**269.32**

Hispanic Origin: Not Hispanic or Latino (MAPE), i.e., average forecast accuracy = 98%

Month	**Predicted**	**LL 95% CI**	**UL 95% CI**	**SE**
Jul-24	3,107	2,937	3,278	87.1187
Aug-24	3,155	2,970	3,341	94.7215
Sep-24	3,110	2,910	3,309	101.7579
Oct-24	3,169	2,957	3,382	108.3381
Nov-24	2,805	2,580	3,029	114.5410
Dec-24	2,759	2,523	2,995	120.4248
Jan-25	2,401	2,154	2,648	126.0343
Feb-25	2,150	1,893	2,408	131.4044
Mar-25	2,517	2,249	2,784	136.5636
Apr-25	2,635	2,357	2,912	141.5348
May-25	2,900	2,613	3,187	146.3373
Jun-25	2,908	2,612	3,204	150.9871
Total	**33,618**	**31,454**	**35,781**	**1,103.72**

Abbreviations: SE = Standard Error; CI = confidence interval; LL = lower limit; UL = upper limit.

Table 5. Forecasting Motor Vehicle Deaths by Urbanization and Month, United States, July 2024-June 2025 Using Exponential Smoothing Model and July 2020-June 2024 data (48 months)

Urbanization: Large Central Metro (MAPE), i.e., average forecast accuracy = 97%

Month	**Predicted**	**LL 95% CI**	**UL 95% CI**	**SE**
Jul-24	1,001	924	1,077	39.0212
Aug-24	1,038	958	1,117	40.7021
Sep-24	1,054	971	1,137	42.3163
Oct-24	1,095	1,009	1,181	43.8712
Nov-24	996	907	1,084	45.3728
Dec-24	979	887	1,071	46.8263
Jan-25	912	817	1,007	48.2360
Feb-25	809	712	906	49.6056
Mar-25	910	810	1,010	50.9385
Apr-25	920	817	1,022	52.2373
May-25	974	869	1,079	53.5046
Jun-25	975	867	1,082	54.7426
Total	**11,661**	**10,913**	**12,409**	**381.716**

Urbanization: Large Fringe Metro (MAPE), i.e., average forecast accuracy = 97%

Month	**Predicted**	**LL 95% CI**	**UL 95% CI**	**SE**
Jul-24	670	619	722	26.3716
Aug-24	667	613	720	27.4803
Sep-24	670	614	726	28.5459
Oct-24	694	636	752	29.5732
Nov-24	622	562	682	30.5659
Dec-24	617	556	679	31.5274
Jan-25	553	490	617	32.4605
Feb-25	498	433	564	33.3674
Mar-25	558	491	625	34.2504
Apr-25	587	518	655	35.1112
May-25	629	559	700	35.9513
Jun-25	632	560	704	36.7723
Total	**7,397**	**6,896**	**7,899**	**255.838**

Abbreviations: SE = Standard Error; CI = confidence interval; LL = lower limit; UL = upper limit.

Table 5. Forecasting Motor Vehicle Deaths by Urbanization and Month, United States, July 2024-June 2025 Using Exponential Smoothing Model and July 2020-June 2024 data (48 months)

Urbanization: Medium Metro (MAPE), i.e., average forecast accuracy = 98%

Month	Predicted	LL 95% CI	UL 95% CI	SE
Jul-24	944	871	1018	37.3545
Aug-24	953	876	1031	39.4393
Sep-24	909	827	990	41.4194
Oct-24	963	878	1048	43.3090
Nov-24	893	804	981	45.1195
Dec-24	863	771	955	46.8601
Jan-25	747	652	842	48.5384
Feb-25	670	572	769	50.1605
Mar-25	808	706	909	51.7318
Apr-25	845	741	949	53.2568
May-25	891	784	998	54.7393
Jun-25	879	769	989	56.1827
Total	**10,365**	**9,581**	**1,1149**	**400.183**

Urbanization: Micropolitan (Non-metro) (MAPE), i.e., average forecast accuracy = 95%

Month	Predicted	LL 95% CI	UL 95% CI	SE
Jul-24	396	361	431	17.8632
Aug-24	401	365	438	18.6455
Sep-24	376	338	414	19.3963
Oct-24	382	343	422	20.1191
Nov-24	335	295	376	20.8168
Dec-24	328	286	370	21.4919
Jan-25	280	237	324	22.1464
Feb-25	241	196	286	22.7821
Mar-25	289	243	335	23.4005
Apr-25	315	268	362	24.0031
May-25	362	314	410	24.5908
Jun-25	360	311	409	25.1648
Total	**4,066**	**3,722**	**4,411**	**175.735**

Abbreviations: SE = Standard Error; CI = confidence interval; LL = lower limit; UL = upper limit.

Table 5. Forecasting Motor Vehicle Deaths by Urbanization and Month, United States, July 2024-June 2025 Using Exponential Smoothing Model and July 2020-June 2024 data (48 months)

Urbanization: Non-Core (Non-metro) (MAPE), i.e., average forecast accuracy = 95%

Month	**Predicted**	**LL 95% CI**	**UL 95% CI**	**SE**
Jul-24	400	362	439	19.7544
Aug-24	390	351	429	19.8139
Sep-24	374	335	413	19.8731
Oct-24	375	336	414	19.9322
Nov-24	313	274	353	19.9911
Dec-24	307	268	346	20.0498
Jan-25	260	221	300	20.1084
Feb-25	245	205	284	20.1667
Mar-25	265	226	305	20.2250
Apr-25	283	243	322	20.2830
May-25	348	308	388	20.3409
Jun-25	363	323	403	20.3986
Total	**3,924**	**3,729**	**4,119**	**99.354**

Urbanization: Small Metro (MAPE), i.e., average forecast accuracy = 94%

Month	**Predicted**	**LL 95% CI**	**UL 95% CI**	**SE**
Jul-24	408	354	463	27.7391
Aug-24	436	380	491	28.2083
Sep-24	437	381	493	28.6698
Oct-24	443	386	500	29.1240
Nov-24	369	311	427	29.5713
Dec-24	375	316	434	30.0118
Jan-25	315	256	375	30.4460
Feb-25	298	237	358	30.8741
Mar-25	331	270	392	31.2963
Apr-25	363	301	425	31.7129
May-25	425	363	488	32.1241
Jun-25	412	349	476	32.5301
Total	**4,614**	**4,215**	**5,012**	**203.165**

Abbreviations: SE = Standard Error; CI = confidence interval; LL = lower limit; UL = upper limit.

Table 6. Forecasting Motor Vehicle Deaths by Age Group and Month, United States, July 2024-June 2025 Using Exponential Smoothing Model and July 2020-June 2024 data (48 months)

Ages: < 4 years (MAPE), i.e., average forecast accuracy = 82%

Month	**Predicted**	**LL 95% CI**	**UL 95% CI**	**SE**
Jul-24	28	16	40	6.1649
Aug-24	34	21	46	6.1898
Sep-24	24	12	36	6.2145
Oct-24	26	14	39	6.2391
Nov-24	25	12	37	6.2636
Dec-24	24	12	36	6.2881
Jan-25	18	-	30	6.3124
Feb-25	18	-	31	6.3366
Mar-25	28	15	40	6.3608
Apr-25	32	19	44	6.3848
May-25	33	20	45	6.4088
Jun-25	27	15	40	6.4327
Total	**316**	**252**	**380**	**32.586**

Ages: 5-14 years (MAPE), i.e., average forecast accuracy = 87%. (-) Data is not shown because of small numbers.

Month	**Predicted**	**LL 95% CI**	**UL 95% CI**	**SE**
Jul-24	84	68	100	8.3837
Aug-24	64	48	81	8.4403
Sep-24	65	48	82	8.4965
Oct-24	70	53	87	8.5523
Nov-24	65	48	82	8.6077
Dec-24	58	41	75	8.6628
Jan-25	48	31	65	8.7176
Feb-25	49	31	66	8.7720
Mar-25	59	41	76	8.8260
Apr-25	56	38	73	8.8798
May-25	79	62	97	8.9332
Jun-25	66	48	83	8.9863
Total	**762**	**666**	**858**	**49.034**

Table 6. Forecasting Motor Vehicle Deaths by Age Group and Month, United States, July 2024-June 2025 Using Exponential Smoothing Model and July 2020-June 2024 data (48 months)

Ages: 15-24 years (MAPE), i.e., average forecast accuracy = 97%

Month	**Predicted**	**LL 95% CI**	**UL 95% CI**	**SE**
Jul-24	618	559	677	30.0126
Aug-24	641	582	701	30.2935
Sep-24	604	544	664	30.5719
Oct-24	646	586	707	30.8478
Nov-24	555	494	616	31.1212
Dec-24	539	477	600	31.3923
Jan-25	494	432	556	31.6610
Feb-25	450	387	512	31.9274
Mar-25	520	456	583	32.1917
Apr-25	556	493	620	32.4538
May-25	612	548	676	32.7138
Jun-25	632	567	696	32.9717
Total	**6,867**	**6,496**	**7,237**	**188.918**

Ages: 25-34 years (MAPE), i.e., average forecast accuracy = 95%

Month	**Predicted**	**LL 95% CI**	**UL 95% CI**	**SE**
Jul-24	709	637	781	36.7369
Aug-24	708	633	783	38.3745
Sep-24	684	605	762	39.9450
Oct-24	684	603	765	41.4561
Nov-24	619	535	703	42.9140
Dec-24	590	503	677	44.3240
Jan-25	524	434	613	45.6905
Feb-25	490	398	583	47.0173
Mar-25	540	445	635	48.3076
Apr-25	576	479	673	49.5644
May-25	648	548	747	50.7901
Jun-25	643	541	745	51.9869
Total	**7,415**	**6,702**	**8,128**	**363.617**

Abbreviations: SE = Standard Error; CI = confidence interval; LL = lower limit; UL = upper limit.

Table 6. Forecasting Motor Vehicle Deaths by Age Group and Month, United States, July 2024-June 2025 Using Exponential Smoothing Model and July 2020-June 2024 data (48 months)

Ages: 35-44 years (MAPE), i.e., average forecast accuracy = 97%

Month	**Predicted**	**LL 95% CI**	**UL 95% CI**	**SE**
Jul-24	601	548	655	27.2128
Aug-24	602	545	658	28.9112
Sep-24	603	543	663	30.5153
Oct-24	613	550	676	32.0392
Nov-24	526	460	591	33.4938
Dec-24	533	464	601	34.8878
Jan-25	478	407	549	36.2282
Feb-25	411	337	484	37.5207
Mar-25	497	421	573	38.7702
Apr-25	522	443	600	39.9807
May-25	554	473	635	41.1555
Jun-25	550	467	633	42.2978
Total	**6,489**	**5,894**	**7,084**	**303.631**

Ages: 45-54 years (MAPE), i.e., average forecast accuracy = 96%

Month	**Predicted**	**LL 95% CI**	**UL 95% CI**	**SE**
Jul-24	516	474	559	21.6286
Aug-24	509	464	553	22.7189
Sep-24	509	463	556	23.7593
Oct-24	495	447	544	24.7560
Nov-24	445	394	495	25.7141
Dec-24	417	365	469	26.6377
Jan-25	382	328	436	27.5304
Feb-25	347	291	403	28.3950
Mar-25	408	350	465	29.2341
Apr-25	426	367	485	30.0498
May-25	464	404	525	30.8438
Jun-25	449	387	511	31.6180
Total	**5,368**	**4,930**	**5,806**	**223.459**

Abbreviations: SE = Standard Error; CI = confidence interval; LL = lower limit; UL = upper limit.

Table 6. Forecasting Motor Vehicle Deaths by Age Group and Month, United States, July 2024-June 2025 Using Exponential Smoothing Model and July 2020-June 2024 data (48 months)

Ages: 55-64 years (MAPE), i.e., average forecast accuracy = 95%

Month	**Predicted**	**LL 95% CI**	**UL 95% CI**	**SE**
Jul-24	526	477	575	25.0760
Aug-24	538	487	588	25.8033
Sep-24	539	487	591	26.5107
Oct-24	567	513	620	27.1997
Nov-24	478	423	532	27.8717
Dec-24	479	423	535	28.5278
Jan-25	427	369	484	29.1692
Feb-25	377	319	435	29.7968
Mar-25	414	355	474	30.4115
Apr-25	455	394	516	31.0139
May-25	487	425	549	31.6049
Jun-25	489	426	552	32.1850
Total	**5,776**	**5,354**	**6,199**	**215.422**

Ages: 65-74 years (MAPE), i.e., average forecast accuracy = 96%

Month	**Predicted**	**LL 95% CI**	**UL 95% CI**	**SE**
Jul-24	384	344	424	20.3194
Aug-24	396	353	438	21.7243
Sep-24	395	350	441	23.0438
Oct-24	432	384	479	24.2916
Nov-24	393	343	443	25.4785
Dec-24	402	350	454	26.6124
Jan-25	334	280	388	27.7000
Feb-25	304	248	361	28.7465
Mar-25	347	288	405	29.7561
Apr-25	337	277	397	30.7327
May-25	359	297	421	31.6791
Jun-25	387	323	450	32.5981
Total	**4,469**	**4,008**	**4,931**	**235.494**

Abbreviations: SE = Standard Error; CI = confidence interval; LL = lower limit; UL = upper limit.

Table 6. Forecasting Motor Vehicle Deaths by Age Group and Month, United States, July 2024-June 2025 Using Exponential Smoothing Model and July 2020-June 2024 data (48 months)

Ages: 75-84 years (MAPE), i.e., average forecast accuracy = 93%

Month	Predicted	LL 95% CI	UL 95% CI	SE
Jul-24	238	199	276	19.4173
Aug-24	268	228	308	20.3314
Sep-24	266	224	307	21.2062
Oct-24	287	244	330	22.0463
Nov-24	301	256	346	22.8555
Dec-24	292	246	339	23.6370
Jan-25	243	195	291	24.3936
Feb-25	225	175	274	25.1273
Mar-25	237	186	287	25.8402
Apr-25	243	191	295	26.5340
May-25	270	217	324	27.2101
Jun-25	266	212	321	27.8698
Total	**3,136**	**2,752**	**3,520**	**195.858**

Ages: 85+ years (MAPE), i.e., average forecast accuracy = 93%

Month	Predicted	LL 95% CI	UL 95% CI	SE
Jul-24	106	83	129	11.7757
Aug-24	120	96	144	12.0783
Sep-24	121	96	145	12.3735
Oct-24	124	99	149	12.6618
Nov-24	117	92	143	12.9437
Dec-24	130	104	156	13.2196
Jan-25	112	85	138	13.4898
Feb-25	85	58	112	13.7548
Mar-25	107	80	135	14.0147
Apr-25	105	77	133	14.2699
May-25	117	89	146	14.5206
Jun-25	105	76	134	14.7671
Total	**1,349**	**1,158**	**1,540**	**97.43**

Abbreviations: SE = Standard Error; CI = confidence interval; LL = lower limit; UL = upper limit.

Tables (Study 2)

Table 7. Percentage of Motor Vehicle Deaths by Selected Characteristics, United States, Jul 2020-Jun 2024

	Period					
	Jul 2020-Jun 2022		**Jul 2022-Jun 2024**		**Overall**	
Characteristics	**Count**	**Percent**	**Count**	**Percent**	**Count**	**Percent**
Overall	**85,527**	**100.0**	**82,843**	**100.0**	**168,370**	**100.0**
Sex						
Female	22,902	26.8	22,052	26.6	**44,954**	**26.7**
Male	62,625	73.2	60,791	73.4	**123,416**	**73.3**
Urbanization						
Large Central Metro	21,673	25.3	20,509	24.8	**42,182**	**25.1**
Large Fringe Metro	17,695	20.7	17,489	21.1	**35,184**	**20.9**
Medium Metro	19,280	22.5	18,938	22.9	**38,218**	**22.7**
Small Metro	8,700	10.2	8,241	9.9	**16,941**	**10.1**
Micropolitan (non-metro)	9,605	11.2	9,398	11.3	**19,003**	**11.3**
Non-core (non-metro)	8,574	10.0	8,268	10.0	**16,842**	**10.0**
Age group*						
5-24 years	14,682	17.2	14,226	17.2	**28,908**	**17.2**
25-44 years	30,922	36.2	28,639	34.6	**59,561**	**35.4**
45-64 years	24,526	28.7	23,062	27.8	**47,588**	**28.3**
65+ years	15,397	18.0	16,916	20.4	**32,313**	**19.2**
Race*						
AI/AN	488	0.6	443	0.5	**931**	**0.6**
Asian	395	0.5	468	0.6	**863**	**0.5**
Black or African American	16,124	18.9	14,323	17.3	**30,447**	**18.1**
More than one race	244	0.3	264	0.3	**508**	**0.3**
White	67,078	78.4	65,872	79.5	**132,950**	**79.0**

Abbreviations: AI/AN = American Indian/Alaska Native.

* Age group < 5 years and Native Hawaiian or Other Pacific race small count data not reported because of statistical reliability and data confidentiality; column percentages may not add up to 100% due to rounding.

Table 8. Negative Binomial Regression Analysis of Crash Counts, United States, Jul 2020-Jun 2024

Parameter		**Estimate**	**SE**	**Wald 95% CL**		**Chi-Square**	**p-value**
Intercept		7.0432	0.0245	6.9953	7.0912	82759.9	<.0001
Sex	Female	-1.101	0.0324	-1.1646	-1.0374	1152.36	<.0001
	Male	0	0	0	0	.	Ref.
Age	05-24 years	-0.7935	0.0314	-0.855	-0.7319	639.01	<.0001
	45-64 years	-0.2987	0.0302	-0.3579	-0.2395	97.72	<.0001
	65+ years	-0.9494	0.0319	-1.0119	-0.8869	886.31	<.0001
	25-44 years	0	0	0	0	.	Ref.
Urban	Large Fringe Metro	-0.2648	0.0301	-0.3239	-0.2057	77.15	<.0001
	Medium Metro	-0.1761	0.03	-0.2349	-0.1173	34.46	<.0001
	Micropolitan (non-metro)	-0.958	0.0319	-1.0205	-0.8954	902.13	<.0001
	Non-core (non-metro)	-1.1478	0.0326	-1.2117	-1.0839	1239.83	<.0001
	Small Metro	-0.9922	0.032	-1.0549	-0.9294	960.64	<.0001
	Large Central Metro	0	0	0	0	.	Ref.
Month	May	-0.0896	0.0187	-0.1262	-0.0531	23.09	<.0001
	July	-0.0547	0.0186	-0.0911	-0.0183	8.67	0.0032
	June	-0.0993	0.0187	-0.1359	-0.0627	28.3	<.0001
	April	-0.1949	0.0189	-0.2319	-0.1579	106.57	<.0001
	March	-0.2492	0.019	-0.2864	-0.2119	171.9	<.0001
	August	-0.0273	0.0185	-0.0636	0.0089	2.18	0.1397
	January	-0.2591	0.019	-0.2964	-0.2218	185.45	<.0001
	December	-0.1161	0.0188	-0.1529	-0.0794	38.34	<.0001
	February	-0.387	0.0194	-0.4249	-0.349	399.74	<.0001
	November	-0.1067	0.0187	-0.1434	-0.07	32.55	<.0001
	September	-0.0454	0.0186	-0.0817	-0.009	5.98	0.0145
	October	0	0	0	0	.	Ref.
Dispersion		0.0043	0.0006	0.0033	0.0055		

Abbreviations: SE = Standard Error; CL = Confidence Limit; Ref = Referent group. Statistical significance $p < 0.05$.

Interpretations: Estimate for Female = -1.101 means that, holding all other variables constant, being female is associated with a decrease in the expected number of MV deaths compared to males.

Exponentiated Estimate = $e^{(-1.1)} \approx 0.33$ So, the expected count of MV deaths among females is about 33% that of males, or a 67% reduction. The p-value is < .0001, i.e., statistically highly significant, indicating that the effect of gender (female vs. male) on MV death counts is not due to random chance.

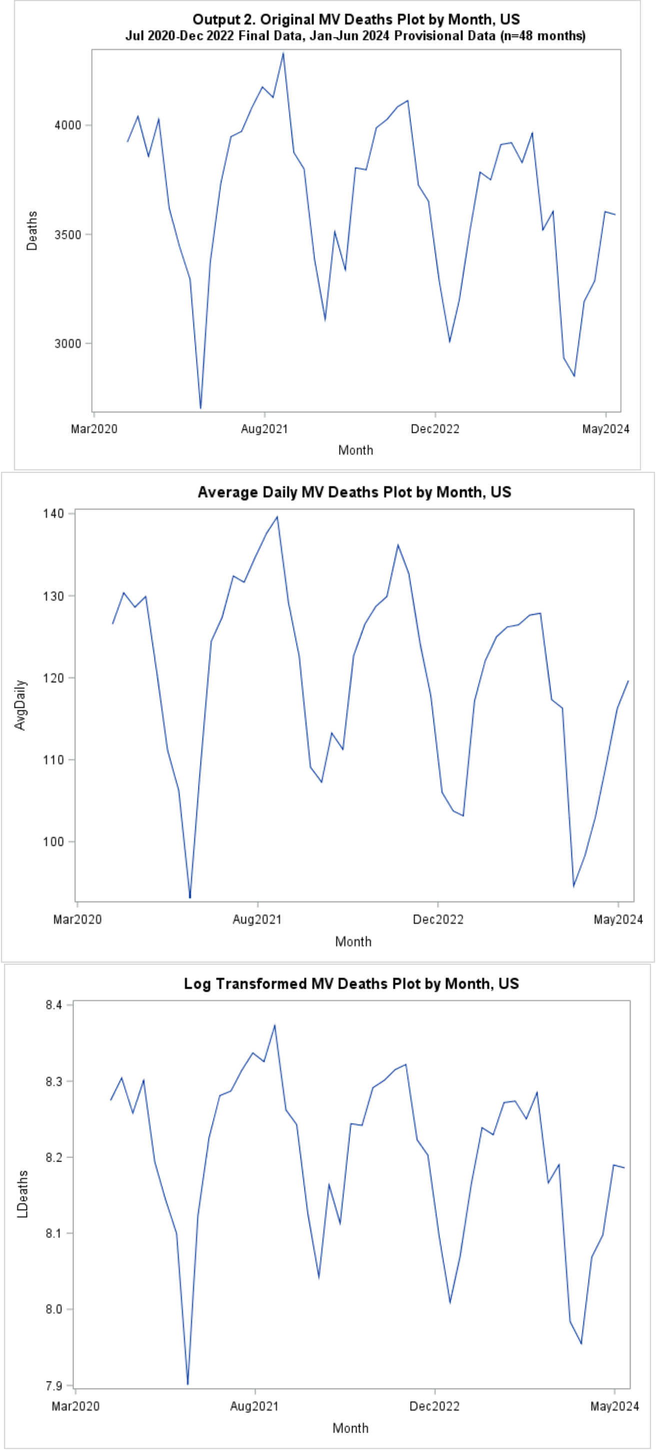
Appendix 1

Output (Study 1)

Output 1. Motor Vehicle Deaths, United States
Jul 2020-Dec 2022 Final Data, Jan-Jun 2024 Provisional Data (total=48 months) N = 175,600 deaths
Accessed 15 Jan 2025: CDC/NCHS https://wonder.cdc.gov/mcd-icd10-provisional.html

```
Obs  Month     Deaths

1    JUL2020   3923

2    AUG2020   4041

3    SEP2020   3858

--- Partial listing ---

46   APR2024   3287

47   MAY2024   3604

48   JUN2024   3590
```

Output 2. Motor Vehicle Deaths Plot by Month, US



Note: The study continued using the original scale for the analyses, as neither the Average Daily MV Death counts nor the Log-Transformed counts enhanced predictive accuracy or helped to smooth the trend.

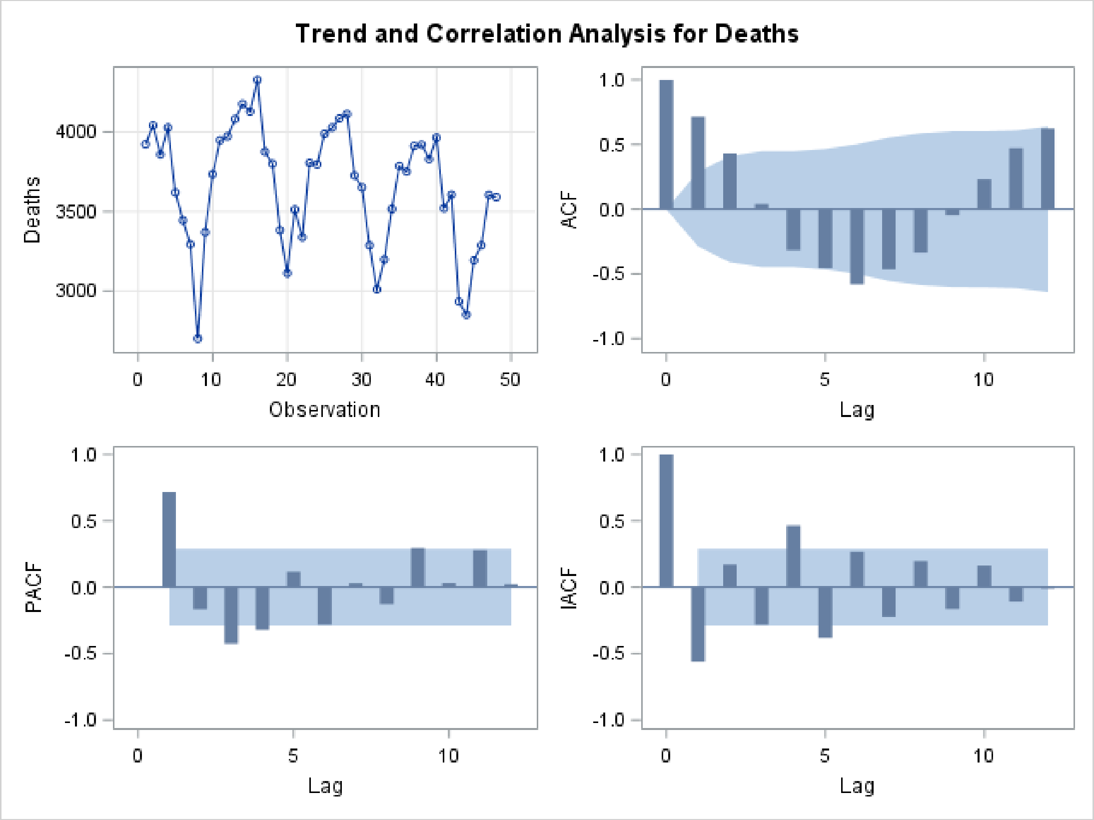
Output 3. ARIMA Identify Procedure: MV Deaths, United States, Jul 2020-Jun 2024

Name of Variable = Deaths

Mean of Working Series	**3658.333**
Standard Deviation	**375.324**
Number of Observations	**48**

Autocorrelation Check for White Noise

To Lag	Chi-Square	DF	Pr > ChiSq
6	72.4	6	<.0001
12	135.98	12	<.0001



Note. In the above deaths plot, we observe the presence of seasonality over the time series (i.e., nonstationary). A formal test and visual inspection of the autocorrelation function (ACF) plots also indicate that the crash data were nonstationary since the ACF decays very slowly. A significant chi-square statistic (p < .0001) indicates that the model does not fit well (Output 3; Brocklebank et al., 2018, p 56).

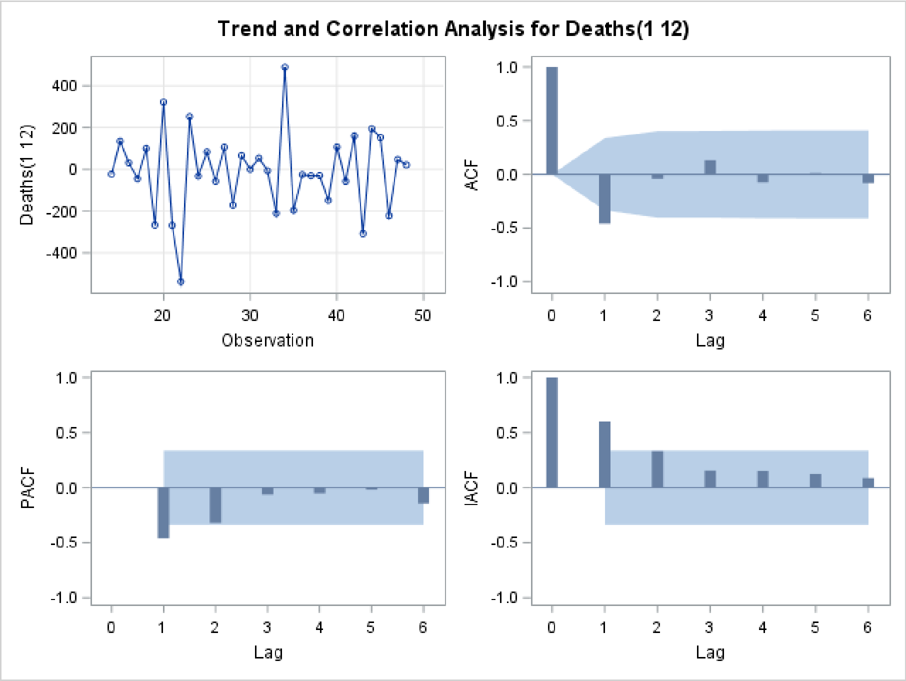
Autocorrelation Check for White Noise continued

Name of Variable = Deaths

Period(s) of Differencing	1,12
Mean of Working Series	-9.08571
Standard Deviation	191.6746
Number of Observations	35
Observation(s) eliminated by differencing	13

Autocorrelation Check for White Noise

To Lag	Chi-Square	DF	Pr > ChiSq
6	**9.36**	**6**	**0.1544**



Note: After n = 13, observations were eliminated by differencing to transform them into stationary, making it easier to model and predict future values. The death plot does not show any trend or seasonality (i.e., it is stationary). It has some spikes, but the series has a consistent long-run mean. A chi-square test (p = .154) > .05 indicates that the model does fit better.

Output 3. ARIMA Dickey-Fuller stationarity tests: MV Deaths, United States, Jul 2020-Jun 2024

Name of Variable = Deaths

Mean of Working Series	**3658.33**
Standard Deviation	**375.324**
Number of Observations	**48**

Autocorrelation Check for White Noise

To Lag	Chi-Square	DF	Pr > ChiSq
6	72.4	6	<.0001
12	135.98	12	<.0001

Augmented Dickey-Fuller Unit Root Tests

Type	Lags	Rho	Pr < Rho	Tau	Pr < Tau	F	Pr > F
Zero Mean	1	-0.2646	0.6186	-0.5	0.4951		
Single Mean	1	-18.2552	0.01	-3	0.0424	4.52	0.0684
Trend	1	-19.5577	0.0449	-3.01	0.1408	4.58	0.2949

Note: The "Augmented Dickey-Fuller Unit Root Tests" for Single Mean, Tau -3.00 (p = .0424) < .05, the null hypothesis test, which states that the data has a unit root (i.e., the series is nonstationary), was rejected (Output 3; Brocklebank et al., 2018, pp. 85-86). Furthermore, per Brocklebank et al. (2018), the adjustment for the lagged differences was motivated by large sample theory, and n = 48 months was not particularly large.

Output 4. ARIMA Forecasts with Moving Average, MV Deaths, United States, Jul 2020-Jun 2024

Maximum Likelihood Estimation

Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag
MA1,1	0.59408	0.13974	4.25	<.0001	1

Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag
MA1,1	0.59408	0.13974	4.25	<.0001	1

Variance Estimate	27206.9
Std Error Estimate	164.9451
AIC	458.1393
SBC	459.6947
Number of Residuals	35

Autocorrelation Check of Residuals

To Lag	Chi-Square	DF	Pr > ChiSq
6	1.31	5	0.9339
12	10.11	11	0.5202

Model for variable Deaths

Period(s) of Differencing	1,12

No mean term in this model.

Moving Average Factors

Factor 1:	1 - 0.59408 B**(1)

Moving Average

Output 5. Mean and Total Motor Vehicle Deaths, United States, Jul 2020-Dec 2022, Excluding Holdout Data (July 2023–June 2024) Used for Accuracy Testing.

Analysis Variable: Deaths

N (month)	Mean	Sum
36	3,705.39	133,394

Note: Descriptive statistics for the deaths

Output 6. ARIMA Forecasts with Moving Average, MV Deaths, United States, Jul 2020-Dec 2022, Excluding Holdout Data (July 2023–June 2024) Used for Accuracy Testing.

Maximum Likelihood Estimation

Parameter	Estimate	Standard Error	t Value	Approx Pr > |t|	Lag
MA1,1	0.53967	0.18275	2.95	0.0031	1

Variance Estimate	34714.92
Std Error Estimate	186.3194
AIC	307.0563
SBC	308.1918
Number of Residuals	23

Autocorrelation Check of Residuals

To Lag	Chi-Square	DF	Pr > ChiSq
6	1.21	5	0.9439
12	6.69	11	0.8234
18	11.52	17	0.8281

Model for variable Deaths

Period(s) of Differencing	1,12

No mean term in this model.

Moving Average Factors

Factor 1:	1 - 0.53967 B**(1)

Output 7. Mean Absolute Percentage Error (MAPE) and Accuracy Percentages, MV Deaths, US, Excluding Holdout Data (July 2023–June 2024) Used for Accuracy Testing.

Variable	**Percent**
Accuracy ESM Seasonal	98.463
Accuracy ESM Additive	98.478
Accuracy ESM Multiplicative	97.805
Accuracy ARIMA MA	97.716

Note. The Exponential Smoothing Models (ESM) seasonal and additive models achieved a notably high average accuracy percentage of *98.5%.* It demonstrates its potential to enhance the timeliness of the provisional data (eliminating data lag by 1 year).

Output (Study 2)

Output 8. Motor Vehicle Deaths, United States, Jul 2020-Jun 2024
Accessed 26 Jan 2025: CDC/NCHS https://wonder.cdc.gov/mcd-icd10-provisional.html

Fitted Normal Distribution for Deaths

Goodness-of-Fit Tests for Normal Distribution

Test		**Statistic**		**p Value**
Kolmogorov-Smirnov	D	0.1508034	Pr > D	<0.010
Cramer-von Mises	W-Sq	4.453887	Pr > W-Sq	<0.005
Anderson-Darling	A-Sq	27.3436122	Pr > A-Sq	<0.005

Output 8. Collinearity Diagnostics (Urbanization dataset)
Model: MODEL1
Dependent Variable: Deaths

Number of Observations = 576

Analysis of Variance

Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	3	16436109	5478703	279.01	<.0001
Error	572	11231986	19636		
Corrected Total	575	27668095			

Root MSE	140.12972	**R-Square**	0.594
Dependent Mean	292.30903	**Adj R-Sq**	0.5919
Coeff Var	47.9389		

Parameter Estimates

Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**Variance Inflation**
Intercept	**1**	91.4132	25.5841	3.57	0.0004	0
Sex	**1**	272.4375	11.6775	23.33	<.0001	1
Urban	**1**	-58.4881	3.4188	-17.11	<.0001	1
Age	**1**	-1.2208	5.2223	-0.23	0.8152	1

Collinearity Diagnostics

Number	**Eigenvalue**	**Condition Index**	**Proportion of Variation**			
			Intercept	**Sex**	**Urban**	**Age**
1	3.6511	1.0000	0.0038	0.0069	0.0126	0.0111
2	0.1812	4.4886	0.0008	0.0071	0.6534	0.3390
3	0.1305	5.2888	0.0125	0.3947	0.1664	0.4349
4	0.0371	9.9191	0.9829	0.5914	0.1676	0.2151

Output 9. Fit Statistics (smaller value is better)

Fit Statistics	**Poisson Model**	**Negative Binomial Model**
Pearson Chi-Square Value/Degrees of freedom	2.5	**1.3**
AIC	5,597.3	**5,412.8**
BIC	5,854.3	**5,674.2**

Fit statistics for the negative binomial were better compared to those of the Poisson model.

Output 10. Fitting a Poisson Model for Crash Count Data
The GENMOD Procedure using the Log link function

Dependent Variable	Deaths
Number of Observations	576

Class	Levels
Sex	2
Urban	6
Age	4
Month	12

Criteria For Assessing Goodness of Fit

Criterion	Value	Value/DF
Deviance	1,295.09	2.51
Scaled Deviance	1,295.09	2.51
Pearson Chi-Square	1,281.55	2.48
Scaled Pearson X2	1,281.55	2.48
Log Likelihood	827,603.71	
Full Log Likelihood	-2,739.67	
AIC (smaller is better)	5,597.34	
AICC (smaller is better)	5,611.06	
BIC (smaller is better)	5,854.35	

Algorithm converged.

Output 11. Negative Binomial Model to Account for Overdispersion

The GENMOD Procedure using the Log link function

Dependent Variable	Deaths
Number of Observations	576

Class	Levels
Sex	2
Urban	6
Age	4
Month	12

Criteria For Assessing Goodness of Fit

Criterion	Value	Value/DF
Deviance	680.73	1.32
Scaled Deviance	680.73	1.32
Pearson Chi-Square	666.80	1.29
Scaled Pearson X2	666.80	1.29
Log Likelihood	827,696.98	
Full Log Likelihood	-2,646.40	
AIC (smaller is better)	5,412.80	
AICC (smaller is better)	5,427.01	
BIC (smaller is better)	5,674.17	

Fit statistics for the negative binomial were better (smaller) compared to those of the Poisson model.
Algorithm converged.

Appendix 2

SAS Program (Study 1)

```
*******************************************************************************
By:      Tadesse Haileyesus
Data:    Accessed 15 Jan 2025: CDC/NCHS https://wonder.cdc.gov/mcd-icd10-provisional.html
*******************************************************************************;
DATA Total;
INPUT @1 Month monyy7. @9 Deaths numx5.0;
CARDS;
Jul2020 3923
Aug2020 4041
Sep2020 3858
Oct2020 4027
Nov2020 3619
Dec2020 3444
Jan2021 3293
Feb2021 2700
Mar2021 3369
--- Partial listing ---
Jan2024 2933
Feb2024 2850
Mar2024 3192
Apr2024 3287
May2024 3604
Jun2024 3590
;

* Different months have different numbers of days. Check if it has any effect;
DATA work.Total;
  SET Total;
  CharMon =substr(put(Month,monyy7.),1,3);
    IF CharMon IN ('SEP','APR','JUN','NOV') THEN N_Days = 30; ELSE N_Days = 31;
  LeapYr=substr(put(Month,monyy7.),1,5);
    IF LeapYr IN ('FEB18','FEB19','FEB21','FEB22','FEB23') THEN N_Days = 28;
    IF LeapYr IN ('FEB20','FEB24') THEN N_Days = 29;
  AvgDaily = Deaths / N_Days ;      * Average daily deaths;
  LDeaths = log(Deaths);            * Log transformed deaths data;
RUN;
PROC MEANS DATA = WORK.Total N MEAN SUM;
  VAR Deaths N_Days;                * Descriptive Stat;
RUN;
TITLE1 "Output 2. Original MV Deaths Plot by Month, US";
PROC SGPLOT DATA = WORK.Total;
  SERIES X=Month y=Deaths; * Y=AvgDaily; * Y=lDeaths;
  XAXIS VALUESFORMAT=monyy7.;
QUIT;
```

```
*Auto-regressive Integrated Moving Averages (ARIMA) checks;
TITLE1 "Output 3. ARIMA Identify Procedure: MV Deaths, United States, Jul 2020-Jun 2024";
PROC ARIMA DATA = WORK.Total;
  IDENTIFY VAR=Deaths nLag=12;         * # of lag limited to 12 (default 24) to ensure reliable results;
  IDENTIFY VAR=Deaths(1) nLag=12;    * 1st Obs will be eliminated by differencing (n=48-1 months);
  IDENTIFY VAR=Deaths(1,12) nLag=6; * 1+12 Obs will be eliminated, removing yearly seasonality;
RUN;
TITLE1 "Output 3. ARIMA Augmented Dickey-Fuller (ADF) Stationarity Tests";
PROC ARIMA DATA = WORK.Total;
  IDENTIFY VAR=Deaths STATIONARITY=(adf=(1)) OUTCOV=adf;
RUN;
TITLE1 "Output 4. ARIMA Forecasts with Moving Average, MV Deaths, United States, Jul 2020-Jun 2024";
PROC ARIMA DATA = WORK.Total plots=(forecast(forecast));
  IDENTIFY VAR=Deaths(1,12) noprint;
  ESTIMATE q=(1) NOCONSTANT METHOD=ml;                * Fit the MA(1) model;
  FORECAST LEAD=12 INTERVAL=month ID=Month OUT=outf_ARIMA_MA  ;
RUN; QUIT;
TITLE2 "Before the Holdout Data Deleted for Accuracy Test: Mean Absolute Percentage Errors (MAPEs)" ;
PROC ARIMA DATA = WORK.Total plots(only)=(forecast(forecast));
  IDENTIFY VAR=Deaths(1,12) nLag=6 NOPRINT;
  ESTIMATE q=(1) (12) NOCONSTANT METHOD=ml; * MA(q) (12) specify multiplicative structure;
  FORECAST LEAD=12 INTERVAL=month ID=Month OUT=outf_ARIMA_Mult_q_MA;
RUN; QUIT;
* Accuracy test: Using a holdout test data
* Determine the total number of observations to output only the requested one;
DATA _NULL_ ;
    SET WORK.TOTAL NOBS=num_obs;
    CALL symputx('num_obs', num_obs);
    STOP;
RUN; * Exclude the Last 12 monthly data;
DATA MV_holdout;
    SET WORK.TOTAL ;
    IF _N_ <= (&num_obs - 12) THEN OUTPUT;
RUN;
TITLE1 "Output 5. Mean and Total, Motor Vehicle Deaths, United States, Jul 2020-Dec 2022 ";
TITLE2 "After the Holdout data (Jul 2023-Jun 2024) was deleted for the accuracy test.";
PROC MEANS DATA = WORK.MV_holdout N MEAN SUM ;
  VAR Deaths;
RUN;
* Final Forecast comparisons;
TITLE1 "Output 6. ARIMA Forecasts with Moving Average, MV Deaths, United States, Jul 2020-Dec 2022" ;
TITLE2 "After the Holdout data (Jul 2023-Jun 2024) was deleted for the accuracy test.";
PROC ARIMA DATA = MV_holdout plots(only)=(forecast(forecast));
  IDENTIFY VAR=Deaths(1,12)  NOPRINT;
  ESTIMATE q=(1)   NOCONSTANT METHOD=ml;
  FORECAST LEAD=12 INTERVAL=month ID=Month OUT=outf_ARIMA_MA;
RUN; QUIT;
```

```
TITLE1 "Output 6.1. ARIMA Forecasts with Seasonal Multiplicative Moving Average, MV Deaths, US" ;
TITLE2 "Jul 2020-Dec 2022, after the Holdout data (Jul 2023-Jun 2024) deleted for the accuracy test.";
PROC ARIMA DATA = MV_holdout plots(only)=(forecast(forecast));
  IDENTIFY VAR=Deaths(1,12)  NOPRINT;
  ESTIMATE NOCONSTANT METHOD=ml;
  ESTIMATE q=(1)  (12) NOCONSTANT METHOD=ml;
  FORECAST LEAD=12 INTERVAL=month ID=Month OUT=outf_ARIMA_Mult_MA;
RUN; QUIT;;
TITLE1 "Output 6.2. Seasonal Method: ESM (Exponential Smoothing Model) Forecast, MV Deaths, US" ;
TITLE2 "After the Holdout data (Jul 2023-Jun 2024) was deleted for the accuracy test.";
PROC ESM DATA = MV_holdout outest=betas outfor=for_ESM_Seasonal lead=12 outsum=totals plot
=(modelforecasts);
  forecast Deaths / method =seasonal;
  ID Month interval=month;
RUN;
PROC PRINT DATA = betas NOOBS;
PROC PRINT DATA = totals; * For overall total output SE;
PROC PRINT DATA = WORK.accu_ESM_Seasonal;
  FORMAT Month monyy7.; RUN;
TITLE1 "Output 6.3. Additive Winters Method: ESM (Exponential Smoothing Model), MV Deaths, US" ;
TITLE2 "After the Holdout data (Jul 2023-Jun 2024) was deleted for the accuracy test.";
PROC ESM DATA = MV_holdout outest=betas outfor=for_ESM_Additive lead=12 outsum=totals plot =
(modelforecasts);
  forecast Deaths/method=addwinters;
  ID Month interval=month;
RUN;
PROC PRINT DATA = betas NOOBS;
PROC PRINT DATA = totals; RUN;
TITLE1 "Output 6.4. Multiplicative Winters Method: ESM Forecast, MV Deaths, US, Jul 2020-Dec 2022";
TITLE2 "After the Holdout data (Jul 2023-Jun 2024) was deleted for the accuracy test";
PROC ESM DATA = MV_holdout outest=betas outfor=for_ESM_Multiplicative
  lead=12 outsum=totals plot = ( modelforecasts);
  forecast Deaths / method =winters;
  ID Month interval=month;
RUN;
PROC PRINT DATA = betas noobs; RUN;
PROC PRINT DATA = totals; RUN;
```

```
* Determine the total number of observations to output only the requested one;
DATA _NULL_ ;
  SET for_ESM_Multiplicative NOBS=num_obs;
  CALL symputx('num_obs', num_obs);
  STOP; RUN;
* Keep the Last 24 obs - 12 with no actual death counts;
data accu_ESM_Multiplicative;
  SET for_ESM_Multiplicative ;
  IF _N_ > (&num_obs - 24) AND ACTUAL NE .  THEN OUTPUT;
RUN;
DATA accu_ESM_Multiplicative;
 SET accu_ESM_Multiplicative;
  e2_ESM_Multiplicative = error**2;
       Abs_Pct_Err_ESM_Multiplicative = 100*abs(error)/actual;
       Accuracy_ESM_Multiplicative = 100 - Abs_Pct_Err_ESM_Multiplicative;
RUN;
PROC PRINT DATA = accu_ESM_Multiplicative;
 FORMAT Month monyy7.; RUN;
DATA all;
 MERGE accu_ESM_Seasonal accu_ESM_Additive accu_ESM_Multiplicative
    WORK.accu_ARIMA_MA ;* accu_ARIMA_Mult_MA ;
RUN;
TITLE1 "Output 7. Mean Absolute Percentage Error (MAPE) and Accuracy Percentages, MV Deaths, US" ;
TITLE2 "After the Holdout data (Jul 2023-Jun 2024) was deleted for the accuracy test.";
PROC MEANS N MEAN DATA = all;
 VAR Abs_Pct_Err_ESM_Seasonal Abs_Pct_Err_ESM_Additive
   Abs_Pct_Err_ESM_Multiplicative Abs_Pct_Err_ARIMA_MA
   Accuracy_ESM_Seasonal Accuracy_ESM_Additive
   Accuracy_ESM_Multiplicative Accuracy_ARIMA_MA  ;
RUN; * Abs_Pct_Err_ARIMA_Mult_MA Accuracy_ARIMA_Mult_MA ;
TITLE1 "Output 8. Final Seasonal Method Forecast: ESM (Exponential Smoothing Model), MV Deaths, US" ;
TITLE2 "Jul 2020-Jun 2024 Final and Provisional Data, Jul 2024-Jun 2025 Forecast Data";
PROC ESM DATA = WORK.Total outest=betas outfor=Final_ESM_Seasonal lead=12 outsum=totals plot
=(modelforecasts); *(forecasts modelforecasts);
 forecast Deaths / method =seasonal; *linear; * no ML method ;
 ID Month interval=month;
RUN;
PROC PRINT DATA = betas NOOBS; RUN;
PROC PRINT DATA = totals; RUN; * Output overall total SE;
```

SAS Program (Study 2)

```
********************************************************************************
By:     Tadesse Haileyesus
Data:   Accessed 15 Jan 2025: CDC/NCHS https://wonder.cdc.gov/mcd-icd10-provisional.html
********************************************************************************;
TITLE1 "Regression Analysis: Motor Vehicle Deaths, United States, Jul 2020-Jun 2024";
TITLE3 "Accessed 26 Jan 2025: CDC/NCHS https://wonder.cdc.gov/mcd-icd10-provisional.html ";

PROC MEANS DATA = DataOut.By_month N SUM ; VAR Deaths; RUN; * Descriptive Stat for Total ;

PROC SORT DATA = DataOut.By_month; BY Sex ;
PROC MEANS DATA = DataOut.By_month N SUM ; VAR Deaths; BY Sex; RUN;

* Create a density curve to show the distribution of a numeric variable, Deaths;
PROC PRINT DATA = DataOut.By_month (OBS=15); RUN;
TITLE1 "Output 8. Regression Analysis: Motor Vehicle Deaths, United States, Jul 2020-Jun 2024";
PROC SGPLOT DATA = DataOut.By_month;
  HISTOGRAM Deaths;
  DENSITY Deaths;
  DENSITY Deaths / TYPE=kernel;
RUN;

ODS SELECT moments basicmeasures goodnessoffit;
PROC UNIVARIATE DATA = DataOut.By_month;
 VAR Deaths;
 HISTOGRAM / NORMAL;
RUN;

PROC GENMOD DATA = DataOut.By_month;
   CLASS Sex Urban Age_Grp Month_name; * Char variables specified in the CLASS statement;
   MODEL Deaths =Sex Urban Age_Grp Month_name
        / DIST=poi LINK=log TYPE3;
   TITLE1 "Output 10. Fitting a Poisson Model for Crash Count Data";
RUN;
PROC GENMOD DATA = DataOut.By_month PLOTS(UNPACK) =ALL;
   CLASS Sex Urban Age_Grp Month_name;
   MODEL Deaths =Sex Urban Age_Grp Month_name
        / OBSTATS DIST=negbin LINK=log TYPE3 ;
   TITLE1 "Output 11. Negative Binomial Model to Account for Overdispersion";
RUN;
QUIT;
```