STATISTICAL MODELLING OF INJURY MORTALITY IN GAUTENG AND MPUMALANGA PROVINCES OF SOUTH AFRICA

by

RAMOOKANA JOHANNES LEBOGO

[redacted]

DISSERTATION

Submitted in fulfillment of the requirements for the degree of

MASTER OF SCIENCE

in

STATISTICS

in the

FACULTY OF SCIENCE AND AGRICULTURE
(School of Mathematical and Computer Sciences)

at the

UNIVERSITY OF LIMPOPO

SUPERVISOR: Prof. D Maposa

CO-SUPERVISOR: Dr. A Boateng (KNUST, Ghana)

MAY 2023

Declaration

I, **Ramookana Johannes Lebogo**, declare that the dissertation hereby submitted to the University of Limpopo, for the degree of Master of Science in Statistics has not been submitted by me for a degree at this or any university; that it is my work in design and in execution, and that all material contained herein has been duly acknowledged by means of list of references.

Signature:.....

Date: 19 MAY 2023

Abstract

Injury is a truly global health issue with massive societal and economic consequences. In this study, Negative-Binomial integer-valued generalised autoregressive conditional heteroscedasticity (NB-INGARCH) and autoregressive integrated moving average (ARIMA) techniques have been compared and used to build models for both Mpumalanga and Gauteng monthly injury mortality data. The best model was chosen using the root mean square error (RMSE). The best model is the one with the lowest RMSE value. The ARIMA$(1,1,1) \times (1,1,1)_{12}$ model had the lowest RMSE, making it the most suitable model for both MP and GP monthly injury mortality data. The results identified ARIMA$(1,1,1) \times (1,1,1)_{12}$ as an appropriate model for predicting Mpumalanga and Gauteng monthly injury mortality with the lowest root mean square error. ARIMA$(1,1,1)\times (1,1,1)_{12}$ model is applied to forecast the injury mortality for the next two years. Furthermore, the forecasted results of ARIMA$(1,1,1) \times (1,1,1)_{12}$ model show a decrease of injury mortality in 2020 as compared to 2019. A multifaceted approach to reduce injury mortality is needed. Regulating alcohol sales and raising alcohol prices prevent all forms of violence, while improving drinking environments prevent youth violence. A Graduated Driver Licensing system could benefit the youth driver population to reduce transport accidents.

Keywords:

ARIMA model, Forecasting , NB-INGARCH model and Injury mortality.

Dedication

I dedicate this dissertation to my beautiful wife, Melidah Mapeu Lebogo and my little girl, Mohau Lebogo. I also dedicate this work to my colleagues in the Department of Statistics and Operations Research.

Acknowledgments

It gives me great pleasure to thank a few people whom this dissertation would have not been completed without them. I give thanks to the Lord God of Hosts, whose name is Jesus Christ for giving me wisdom, strength and guidance to successfully complete this dissertation. Firstly, I would like to thank both Prof D Maposa and Dr A Boateng as my supervisor and co-supervisor, respectively, for the excellent supervision, expert guidance and valuable contribution to insure that this dissertation is successfully completed. Secondly, a special thanks to Mr KN Maswanganyi and Mr TF Maja for valuable contributions.

Lastly, I would also like to thank Statistics South Africa for the data and the Department of Statistics and Operations Research for the opportunity to further my postgraduate studies.

Contents

List of Figures

List of Tables

List of Abbreviations and Acronyms

ABS	Australia Bureau of Statistics
ACF	Autocorrelation Function
ADF	Augmented Dickey-Fuller
AIC	Alkaike Information Criterion
AIDS	Acquired Immune Deficiency Syndrome
AR	Autoregression
ARCH	Autoregressive Conditional Heteroskedasticity
ARIMA	Autoregressive Integrated Moving Average
BIC	Bayesian Information Criterion
CDC	Disease Control and Prevention
Child PIP	Child Healthcare Problem Identification Programme
CODURF	Cause of Death Unit Record Files
Df	Degrees of Freedom
DSPs	Disease Surveillance Point System
GARCH	Generalised Autoregressive Conditional Heteroskedasticity

GBD	Global Burden of Disease
GDL	Graduated Driver Licensing
GLM	Generalised Linear Model
GP	Gauteng
GSA	Geographical Service Area
HIV	Human Immunodeficiency Virus
IM	Injury Mortality
INGARCH	Integer-valued Generalised Autoregressive - Conditional Heteroscedasticity
JB	Jarque Bera
KPSS	Kwiatkowski-Phillips-Schmidt-Shin
LMSS	Local Mortality Surveillance System
LSE	Least Square Estimates
MA	Moving Average
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MLE	Maximum Likelihood Estimates
MME	Method of Moments Estimates
MP	Mpumalanga
MSE	Mean Square Error
NB	Negative Binomial

NB-INGARCH	Negative Binomial Integer-valued Generalised - Autoregressive Conditional Heteroscedasticity
NDP	National Development Plan
NIMSS	National Injury Mortality Surveillance System
PACF	Partial Autocorrelation Function
PP	Phillips-Perron
PPIP	Perinatal Problem Identification Programme
p-value	Probability Value
QMLE	Quasi-conditional Maximum Likelihood Estimation
Q-Q	Quantile-Quantile
RMSE	Root Mean Square Error
SAR	Seasonal Autoregressive
SDGs	Sustainable Development Goals
SMA	Seasonal Moving Average
StatsSA	Statistics South Africa
SW	Shapiro Wilk
TB	Tuberculosis
USA	United States of America
WHO	World Health Organisation
WISQARS	Web-Based Injury Statistics Query and Reporting System

Chapter 1

Introduction and background

1.1 Introduction

An injury, is defined as a bodily cut at the organic level resulting from acute exposure to energy. This energy may be as a result of mechanical, thermal, electrical and chemical, which could interact with the body in amounts or rates that exceed the threshold of physiological tolerance (Robertson, 2015). Injuries are the major causes of mortality, public health threats and disability at any age worldwide, which may be avertable (James et al., 2020). Mortality is defined as the state of being subject to death, which may be caused by the circumstances of communicable diseases, non-communicable diseases and injuries (WHO, 2019). Mortality caused by injury or injury mortality (IM) are classified as mortality due to intentional injuries such as suicide and homicide, and also unintentional injuries such as road traffic, falls, poisoning, burn and drowning.

IM contributes to almost 4.4 million deaths per year globally, which is about 12055 deaths per day (WHO, 2021). According to World Health Organisation

(WHO) report of 2021, deaths caused by unintentional injuries and intentional injuries accounts to 3.16 million and 1.25 million every year, respectively. The report further indicated that over 30% of these deaths resulted from road traffic, 17% from suicide, 10% from homicide and 7% from drowning.

Africa, just like any continent around the world, continue to witness a high burden of IM. For instant Tyson et al. (2015), indicated that regardless of lack of vital or quality data on the causes of death in sub-Saharan Africa, few studies conducted found that there is a high proportions of deaths due to injuries. Despite Africa having just 2% of the world's motor vehicles, it accounts for 16% of worldwide road traffic deaths and has the highest road fatality rate of all WHO regions (WHO, 2021). The road traffic IM has been increasing for the three last decade and African region continues to have the highest road traffic deaths (Schlottmann et al., 2017). For instance, Nigeria and South Africa have the highest mortality rate on road traffic with 33.7 and 31.9 deaths per 100000 population per year, respectively.

South Africa has high mortality levels resulting from a unique quadruple disease burden, which is burdened with a diverse spectrum of diseases, including infectious diseases, chronic and degenerative diseases, malnutrition and childbirth-related conditions and a disproportionately large burden of injuries (Goosen et al., 2003; Brysiewicz, 2001).

Injuries impose as a fourth major burden on the South African population, which is driven in particular by the high incidence of homicide, while homicide and road traffic collisions are the leading causes of IM in South Africa (Matzopoulos et al., 2019). According to Statistics South Africa (StatsSA) report of 2021, more than a tenth of all deaths that occurred in South Africa in 2018 were due to IM, (StatsSA, 2021). It was observed that the majority of injury causes

of death resulting from unintentional injury were 82% in South Africa. South Africa is one of the few places in the world where rates of intentional injury exceed the rates of unintentional injury (WHO, 2019). The 2012 national burden of disease study found that homicide was the eighth leading cause of overall premature death in South Africa, while being the second leading cause of premature deaths for males, after HIV/AIDS (Matzopoulos et al., 2015). The injury burden was very high, particularly for homicide, which was approximately six times the global average (Matzopoulos et al., 2015).

Within individual countries, the frequency and the patterns of IM vary between rural and urban areas (Kmet and Macarthur, 2006). The rate of urbanisation in South Africa has increased dramatically over the years. The provincial estimates show that Gauteng has the largest of population, with 15.8 million and Mpumalanga accounted to the fifth with 4.7 million, (StatsSA, 2021). According to Gantchev et al. (2015), Gauteng (urban) and Mpumalanga (rural) had equally higher overall IM rate found among children from Gauteng being 31.7 per 100000 deaths and Mpumalanga with 29.2 per 100000 deaths. Furthermore, in particular passenger related motor vehicle deaths were more evident among children in rural areas than urban areas, also burn was more common in the urban than the rural areas. According to StatsSA (2021) the 13.7% of IM was in Eastern Cape which was the leading province, while Gauteng and Mpumalanga were fourth and fifth with 12% and 10.9%, respectively.

The United Nations General Assembly adopted a resolution on 25 September 2015, where a 2030 agenda for Sustainable Development called for action by all countries to eradicate poverty, reduce mortality and to realise the human rights of all. This consisted of 17 Sustainable Development Goals (SDGs) and 169 targets (Cf, 2015). The most important SDGs pertaining to reduction of injury and violence include: “eliminate all forms of violence against all women and girls”,

"significantly reducing all forms of violence and related rates everywhere", "by 2020, halve the number of global deaths and injuries from road traffic". The 2014-2019 strategic framework in South Africa has the 2030 National Development Plan (NDP), which provides for an increase in South Africans life expectancy by 2030 of at least 70 years, and for a 50% decreasing in the levels of violence, road traffic injury and other injuries compared to 2010, among others (dNational Department of Health, 2015).

The aim of this study is to investigate mortality caused by injuries in South Africa using various statistical techniques in the Gauteng and Mpumalanga provinces. Whereas Gauteng is defined as predominantly urban, Mpumalanga is mainly rural in nature. Although there are no agreed-upon universal criteria delineating rural and urban areas, they may be distinguished by difference along several dimensions, including infrastructure, social service, non-agricultural employment, income and population density (Bank, 2011).

1.2 Problem statement

Mortality is a major concern in South Africa and worldwide according to (WHO, 2019). Based on literature and South African National Development Plan (NDP), which was adopted to minimise injuries, accidents and abuse by 50% in 2030, a broad national or provincial study is needed to help address these issues in South Africa. The current study will attempt to address these challenges at the provincial level by identifying the factors that contribute to injury mortality.

1.3 Motivation

A reduction in the proportion of IM all over the world is important. The loss of working group through injury is a big motivating factor to carry out this study

as these negatively affect the economy of a country. An understanding of various factors that contribute to injury mortality could help the country and the world to become a safer place to live in and by so doing, reducing mortality resulting from injuries. This study will therefore employ statistical methods to investigate and model IM in the two provinces of South Africa. The use of statistical techniques such as econometric and time series techniques would allow more efficient use of the available injury mortality information. According to the literature study, there has not been much statistical research in the field of mortality caused by injuries in South Africa using these proposed count data and time series models, which might result in limited information on potential risk factors contributing to IM. The motivation for selecting Gauteng and Mpumalanga provinces in this study, is that there are nine provinces in South Africa and of those nine, Mpumalanga, Limpopo, Eastern Cape are known as rural provinces while the rest are urban provinces (Atkinson, 2014).

The proposed study will use the available IM dataset obtained from Statistics South Africa on mortality and causes of death in Gauteng and Mpumalanga provinces, as well as some statistical techniques, to identify various factors associated with IM in these two provinces, motivated by the projected increase in IM. The identification of these contributing factors using various statistical techniques, will among others, assist the government and various departments in reducing mortality caused by injuries.

1.4 Aim

The study aims to investigate mortality caused by injuries using various statistical techniques in the Gauteng and Mpumalanga provinces of South Africa.

1.5 Objectives

The objectives of the study are to:

(i) Identify factors associated with mortality related injury.

(ii) Compare injury mortality of Gauteng and Mpumalanga provinces.

(iii) Model the pattern of injury mortality in these provinces.

(iv) Perform a comparative analysis of various statistical methods used.

(v) Forecast the injury mortality in these two provinces using the best time series model.

1.6 Methodology

In this study, mortality and causes of death data from StatsSA for the years 2008-2018 will be utilised. The analysis of this data will be conducted using various statistical models, namely, count data models such as Poisson and Negative Binomial models as well as time series models such as autoregressive integrated moving average (ARIMA), Poisson integer-valued generalised autoregressive conditional heteroscedasticity (INGARCH) and Negative Binomial integer-valued generalised autoregressive conditional heteroscedasticity (NB-INGARCH). SAS, R and SPSS statistical packages will be used for both data management and analysis purposes.

1.7 Significance of the study

The findings of this study are useful for intervention measures of the government agencies such as StatsSA and WHO. This study will be useful in setting

up appropriate plans regarding the incidence and trend of mortality from non-natural injuries so that prevention measures can be put in place to reduce mortality from injuries. The study will also provide the two provinces of South Africa with information, such as warnings, awareness, among others. The findings may also be useful in tracking and assessing the operations of the government and different agencies involved. Furthermore, this study will also act as a reference point for other researchers seeking to model mortality from injuries.

1.8 Structure of the dissertation

Chapter 1 has provided the introduction and background of the study, problem statement and motivation of the research study. The aim, objectives and significance of the study are also provided in Chapter 1. Chapter 2 presents the literature review, while Chapter 3 provides the methodology of the research study. Chapter 4 presents the data analysis in the form of tables and figures. Interpretations and discussions of the results are also provided in Chapter 4. Chapter 5 provides the concluding remarks and recommendations of the study.

Chapter 2

Literature review

2.1 Introduction

This chapter addresses the global epidemiology of injuries, followed by South Africa's estimates for injury mortality (IM) then injury differences in urban and rural areas and the overview of IM.

2.2 Global estimates of injury mortality

In 2019, global estimates for injuries reported almost 4.4 million deaths per year, with 12055 deaths per day, while unintentional injuries (drowning, road traffic, falls, burn, poisoning, natural disasters) claimed the lives of 3.16 million people and intentional injuries (such as homicide, suicide, war) claimed the lives of 1.25 million people, with road traffic accounting for 30% of these deaths, suicide for 17%, and homicide 10% (WHO, 2021). Around three-quarters of deaths worldwide were caused by road traffic, while four-fifths of homicide were

committed by males. Furthermore, among those aged 5-29 years, three of the top five causes of mortality were injury-related, including road traffic, homicide and suicide.

2.2.1 Injury mortality statistics globally

Abio et al. (2020) conducted a study on the trends in mortality from external causes in the Republic of Seychelles between 1989 and 2018. The data were extracted from the national vital statistics register and Negative Binomial models were used for causes of deaths. Road traffic injuries, drowning, suicide, poisoning, homicide, and falls are examples of external causes of deaths that may be characterized depending on the nature of the sources of injuries (Holder, 2001). External causes accounted for 8.5% of all deaths from 1989 to 2018. Drowning was the leading cause of IM, accounting for 22% of deaths, followed by road traffic injuries, which accounted for 18% of deaths, and 9% suicide, 8% poisoning and 7% homicide. Males, on the other hand, accounted for 78% of all injury mortality, while females accounted for 22%. Males had the highest mortality rates from external causes, with 19% drowning, 14% road traffic, and 8% suicide, whereas females had the highest mortality rates from external causes, with 3% road traffic, 3% drowning, and 2% homicide. The majority of these deaths occurred in people aged 20 to 39 and 40 to 59 which accounting 37% and 27%, respectively.

Roth et al. (2018) reported on Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries (one of the country was Ethiopia) and territories from 1980–2017 and negative binomial models were used for causes of deaths. In 2017, injuries were responsible for 12% of all deaths. Road traffic injuries were the leading cause of death in 2017, accounting for 1.24 million deaths of all injury deaths in the year. In 2017, 794000 people died from suicide, 696000 died from falls, 405000 died from homicide,

and 295000 died from drowning. Anteneh and Endris (2020) studied injury-related adult mortality in Addis Ababa, Ethiopia, from 2007 to 2012 and 2015-2016, using data from the Addis Ababa Mortality Surveillance program's verbal autopsy. They discovered that injury-related mortality accounted for 7% of all deaths, whereas non-injury deaths accounted for 93%. Road traffic claimed the most lives for both male and females, accounting for 40% of all injury deaths and 27% of suicide deaths. Males accounted for 80% of IM, whereas suicide accounted for 30% of males' deaths and 15% of females. The age group 15-34 years accounted for 44% of homicide and suicide, followed by 35-54 years with 32%.

In a study in United States of America (USA), (Heron, 2019) reported on leading causes of death in 2017. Data of this report was based on information from all deaths certificates files in the 50 states and District of Columbia. The application of regression analysis was used on the mortality data of 50 states and District of Columbia. In 2017, the top ten causes of deaths accounted for 74% of all deaths in the USA, with road traffic and suicide placed third and tenth, with 6.0% and 1.7%, respectively. Males and females, accounting for 7.6%, and 4% of all deaths, respectively. The relative burden of mortality from these causes was significantly larger at aged 1 to 9, accounting for 31.8% of all deaths, 40.6% of deaths among aged 10 to 24, and 34.6% among aged 25-44. Murphy et al. (2018) found that unintentional injury was the main cause of IM in the USA, with 40.4 per 100000 population.

Henley et al. (2019) reported the trends in IM in Australia, the data that was presented in these report was a series of Australia Bureau of Statistics (ABS) Cause of Death Unit Record Files (CODURF) for years 1999 to 2017. Age-standardised mortality rates per 100000 population were calculated for unintentional and intentional injuries. Injury was the cause of 13144 deaths in

2016-2017, accounting for 8.2% of all deaths. Unintentional injury was the leading cause of IM, accounting for 75% of all deaths, and intentional for 25%. Falls involved in 38% of IM which was the leading cause of death, followed by suicide with 23%. The majority of deaths occurred at ages 65 and over, accounting for 53%, while 43% of males and 68% of females.

In a study assessing deaths from 1999-2015 and visits to emergency department in 2001-2015 for children and adolescents groups through web-based injury statistics query and reporting system (WISQARS) in the USA, (Ballesteros et al., 2018). The application of regression analysis was used on the WISQARS. In 2015, unintentional injury accounted for 61% of IM, homicide accounted for 20% and suicide accounted for 19%. Males accounted for 69% of deaths due to injury-related, while 31% accounted by females. There were 12977 IM among people aged 0 to 19, for a mortality rate of 16.2 per 100000 population. More than 58% of deaths were accounted on the age 15 to 19. These findings support Curtin et al. (2018), who found an increase in injury among children and adolescents aged 10-19 years between 1999 and 2016, with unintentional injury leading as the cause, followed by suicide and homicide. Additionally, youth, 15-19 years old accounted for 55% of all deaths from injury. Furthermore aged 15 to 19 years old suffered the most injuries as a results of motor vehicle traffic.

2.3 South African estimates of injury mortality

In 2018, IM accounted for 12% of all causes of mortality in South Africa. Unintentional injuries accounted for about 80% of IM deaths, whereas intentional injuries accounted for only 20%. Homicide was the second leading cause of injury, accounting for 14% of all injuries and 1.7% of all deaths. Road traffic were the third most prevalent cause IM, accounting for 11.4% of IM and 1.4% of all deaths, while suicide accounted for 1% of IM and 0.2% of all deaths. This shows

how injuries contribute to the quadruple burden of illness, which encompasses HIV/AIDS, Tuberculosis (TB), communicable diseases (i.e. maternal causes, perinatal conditions, nutritional deficiencies), and non-communicable diseases (i.e. cancer, diabetes, heart disease and asthma) (StatsSA, 2021).

2.3.1 South African injury mortality statistics

Meel (2017) conducted a study to check the pattern of non-natural deaths in the Transkei Sub-region of South Africa. A record review was undertaken from 1996 to 2015 of 24693 medico-legal autopsies performed at Mthatha Forensic Pathology Laboratory. The regression approaches were used to determine the incidence. In that region, the total average of non-natural mortality was 205 per 100000 per population in 2015. The leading cause of IM was unintentional, which accounted for 51% of all deaths, while intentional injury accounted for 49%. The major cause of death in IM was homicide, which accounted for 45% of all deaths, followed by road traffic, which accounted for 24% of all deaths. In males, homicide was the leading cause of death, accounting for 49% of all deaths, followed by road traffic at 22%, while in females, road traffic injury was the leading cause of death, accounting for 30% of all deaths, followed by homicide at 27%. The majority of deaths occurred between the ages 11 and 44, accounting for 64% of all deaths, with males accounting for 54% and females for 10%. Males and females aged 21 to 30 accounted for 25% and 4% of IM, respectively, which was the highest age group.

Pillay-van Wyk et al. (2016) conducted a study on mortality trends and differentials in South Africa from 1997 to 2012. They analysed causes of death from a data for 1997 -2012. These data was extracted from StatsSA. These data were adjusted for completeness using indirect demographic techniques for adults and comparison with survey and census estimates for child mortality. The multinomial logistic regression was applied to the data by age, sex, province and popu-

lation group, to smooth out sampling fluctuations in the cause fractions. Injury-related cause were responsible for 9.6% of all deaths. Homicide remained the leading cause of IM in 1997 and 2012, with 7.3% in 1997 and 3.5% in 2012, moving from third to eighth in 1997 and 2012, respectively, while road traffic stayed in ninth place with 3.6% in 1997 and 3.3% in 2012. In terms of provinces, Western Cape had accounted for 9.5% of IM and was the second top cause of deaths in 2012, followed by 5.8% in Gauteng. Road traffic injuries were the second leading cause of deaths in Mpumalanga, accounted for 5.7% of all deaths, followed by Limpopo with 5.4%. Using data from the Local mortality Surveillance system (LMSS), the child Healthcare Problem Identification Programme (Child PIP), and the Perinatal Problem Identification Programme (PPIP), Reid et al. (2016) reviewed the deaths of children aged 5 and under in the Metro West geographical service area (GSA) of the Western Cape between 1 January and 31 December 2011 with particular reference to cause and location. IM causes accounted for 9% of deaths.

In a study in Western Cape, Prinsloo et al. (2016) reported on validating homicide rate in the Western Cape province from the 2009 IM survey. The mortality data were sourced from Western Cape's provincial IM Surveillance System to complete the national sample. Age-standardised mortality rates per 100000 population were calculated for homicide, suicide, unintentional injury, and transport. In 2009, the Western Cape's age-standardised mortality rate was 40.1 per 100000 population, placing fourth among nine provinces and close to the national average of 38.4 per 100000 population. In the Western Cape, there was a little of variation between metro and non-metro regions. While the Eastern Cape has the highest non-metro homicide rate (57.9 per 100000 population), KwaZulu-Natal has the highest metro homicide rate (72.3 per 100000 population).

2.4 Urban-rural injury mortality differences

A study conducted to investigate trends in urban-rural mortality disparity in China from 2010 to 2016 (Li et al., 2020), they utilised data from disease surveillance point system (DSPs) collected by the Chinese center for Disease Control and Prevention (CDC). Chi-squared test were used to compare differences in rates between urban and rural residents. IM accounted for 7.8% of all the deaths, where urban injury deaths accounted for 6.2% of all the deaths in urban areas, whereas rural injury deaths accounted for 8.5% of all deaths in rural areas. A study in China conducted by Leilei et al. (2019) provided an overview of burden of injury in China in 2017 and the study was aimed to measure the change in this burden between 1990 and 2017, and to explore the underlying factors influencing these change. The Global Burden of Disease (GBD), injury, and risk factors for non-fatal and fatal injury outcomes at the national and 31 subnational levels were used in 2017. Injuries accounted for 7% of total deaths. The five leading causes of IM in urban and rural areas were road traffic, fall, suicide, drowning and poisoning, which accounted to 80% of all injury-related deaths. The mortality rate were significantly higher in rural areas (Li et al., 2020).

Moy et al. (2017) reported on leading causes of death in non-metropolitan (rural) and metropolitan (urban) areas from 1999 to 2014. They utilised data from National Vital Statistics System in USA to calculate age-adjusted rates and potential excess deaths for rural and urban for the five leading causes of deaths. Poisson regression was used on both the number and the excess deaths. During 1999-2014 the age-adjusted deaths rate on unintentional injury under IM were higher in rural with 58% than urban with 38%. Rural IM of 50 per 100000 were higher than those in urban areas of 30 per 100000 population. The age-adjusted deaths rate on unintentional injury were approximately 50% higher in rural areas than urban areas for most of this period. 65% of unintentional in-

jury deaths in rural areas were potentially excess deaths, compared with 29.2% in urban areas.

In a study in northern Finland, (Raatiniemi et al., 2016) reported on IM differences in urban and rural areas over a five-year period from 2007 to 2021. Raatiniemi et al. (2016) utilised data coded according to ICD-10 from Finnish cause-of-death registry. The Kruskal-Wallis test were used to compare differences in rates between urban and rural. Rural IM of 65 per 100000 were higher than those in urban areas of 45 per 100000. General level falls classified as low-energy trauma by the authors, suicide and road traffic incidents were the leading causes of deaths. No significant differences were found in the rates of low-energy trauma, suicide or homicide between urban and rural areas.

Swart et al. (2012) conducted a population based study to examine whether the incidence and pattern of fatal injuries among children differ in rural and urban areas of South Africa utilising National Injury Mortality Surveillance System (NIMSS) for the period 2007. All deaths were among children below 15 years of age in Gauteng (urban) and Mpumalanga (rural) who died in 2007. The cross-sectional method was used to analyse all deaths among children below 15 years of age in Gauteng and Mpumalanga who died in 2007. In Gauteng, the vast majority of child injuries were unintentional (89%) followed by homicide (9%), and suicide (2%). A similar trend was observed in Mpumalanga, where 87.5% of child injury deaths were inadvertent, 10.5% were homicide, and 2% were suicide. Pedestrian injuries were the leading cause of child injury in Gauteng, followed by burns, drowning, passenger-related injuries, and fall, which accounted for 69.8% of all child injury deaths in the province. In Mpumalanga, pedestrian injuries, followed by passenger-related deaths, drowning, burns and poisoning were the main causes of child injury deaths.

Garrib et al. (2011) conducted a study on IM rates and associated factors in rural South Africa from 2000-2007. In the study they used population-based mortality data collected by a demographic surveillance system on all the resident and non-resident members of 11000 households. Deaths and person-years of observation (pyo) were aggregated for individuals between 01 January 2000 and 31 December 2007. The regression analyses were used to analyse mortality data. Injury-related causes accounted for 8.9% of all deaths, with a 142.4 per 100000 pyo IM rate. Homicide was the most prevalent cause of injury deaths in both females and males, accounting for 50% of all injury deaths. Road traffic deaths accounted for 26% of IM, and suicide was the third most common cause, accounting for 8% of IM. Homicide mortality rates were leading in males all ages, peaking at 289.5 deaths per 100000 pyo in the 30-39 year age group. Homicide mortality rates in women peaked in the 70-79 year age group, with 96.8 mortality per 100000 person-years. In all age categories, males had considerably higher road traffic mortality rates than female. All females' suicides happened between the age of 10-40, while 54% of male suicides occurred between the ages of 20 and 29, with a mortality rate of 54.3 deaths per 100000 population. In terms of areas, females in urban regions had approximately 50% of the probability of dying from injury as females in rural areas. Males in urban region had 60% higher risk of injury death than males in rural areas.

2.5 Overview of global and national estimates of injury mortality

These studies reveal that unintentional injury is a leading contributor to IM worldwide. Road traffic injuries seems to be a prominent source on IM in some of these nations, with rural areas reporting more deaths than urban areas. Contrast too international studies, the literature about South Africa reveal that intentional injury is a leading contributor to IM. Homicide seems to be a promi-

nent source on IM in South Africa. According to a study of the literature, there are few nationally representative urban-rural studied on IM in South Africa. This study will fill up the gaps in the literature.

According to the literature study, there has not been much statistical research in the field of mortality caused by injuries in South Africa using these proposed count data and time series models, which might result in limited information on potential risk factors contributing to injury mortality. This study will therefore employ statistical methods to investigate and model IM in the two provinces of South Africa. The use of statistical techniques such as econometric and time series techniques would allow more efficient use of the available IM information.

Chapter 3

Methodology

Introduction

This chapter deals with a detailed description of the data source and study area and the methods used for this research. The chapter addressed the possible probability distributions of count data models such as Poisson and negative binomial distributions as well as time series models such as autoregressive integrated moving average (ARIMA), Poisson integer-valued generalised autoregressive conditional heteroscedasticity (INGARCH) and Negative Binomial integer-valued generalised autoregressive conditional heteroscedasticity (NB-INGARCH).

3.1 Data source and study area

The study uses secondary data obtained from Statistics South Africa (StatsSA) database on mortality and causes of death from the year 2008-2018. StatsSA

records 48 characteristics or factors for every deceased person, such as province, manner of death by victim age and sex, day of death, month of death, year of death, education of deceased and cause of death (homicide, suicide, transport accidents (road traffic), falls, burn, drowning, poisoning and natural disasters), among others. Accordingly, this current study will utilise StatsSA data of these two provinces which are Gauteng and Mpumalanga. The selected provinces are two of the nine provinces of South Africa. The motivation behind the choice of these two provinces is that Gauteng is the economic hub of South Africa and Africa, although it is the smallest of South Africa's nine provinces by land-size (StatsSA, 2019), Gauteng also is one of the provinces which comprises the largest share (25.8%) of the South African population, while Mpumalanga is one of the provinces with about (7.8%) of the South African population (StatsSA, 2019). SAS, R and SPSS statistical packages will be used for both data management and analysis purposes.

3.2 Analytical procedures

3.2.1 Overview of generalised linear models

Basic count data regression models can be represented and understood using the generalised linear model (GLM) framework that emerged in the statistical literature in the early1970s (Nelder and Wedderburn, 1972). The GLM, is an extension of linear modelling process that follows probability distribution other than normal and has residuals that are not normally distributed, it includes linear regression models, analysis of variance models, logistic regression models, Poisson regression models, zero-inflated Poisson regression models, Negative-Binomial models, log-linear models, as well as many other models. The dependent variable (injury mortality) values are predicted from a linear combination of predictor variables, which are connected to the variable via a link function (Armitage et al., 2008).

3.2.2 Overview of time series models

The time series modelling of injury mortality data in this study will be done using the Box-Jenkins autoregressive integrated moving average (ARIMA), Poisson integer-valued generalised autoregressive conditional heteroscedasticity (INGARCH) and Negative Binomial integer-valued generalised autoregressive conditional heteroscedasticity (NB-INGARCH) models. The advantage of these proposed models over other models are that Poisson INGARCH and NB-INGARCH have the ability to handle discreteness in count data whereas the ARIMA model has the strength which results from its distributional assumption underlying the estimation process. Furthermore, the comparison of these time series models will be performed using accuracy measures such as root mean square error (RMSE). Finally, the forecasting of injury mortality will be done based on the selected model between the ARIMA and Poisson INGARCH or NB-INGARCH.

3.3 Generalised linear models

As precedent, GLM count models under consideration in this study are Poisson and Negative Binomial regression.

A generalised model consists of three components listed hereunder:
1. A random (exponential family) component, which specify the conditional distribution of the response variable, Y_i , given the explanatory variables, x_{ij}.
2. A linear function of the regression variables, called the linear predictor

$$\eta_i = \alpha + \beta_i X_{i1} + ... + \beta_k X_{k1} = x_i^{'}\beta, \tag{3.1}$$

on which the expected value μ_i of Y_i depends.
3. An invertible link function,

$$g(\mu_i) = \eta_i, \tag{3.2}$$

which transforms the expectation of the response to the linear predictor. The inverse of the link is sometimes called the mean function

$$g^{-1}(\eta_i) = \mu_i. \tag{3.3}$$

For traditional linear models in which the random component consists of the assumption that the response variable follows the normal distribution, canonical link function is the identity link. The identity link specifies that the expected mean of the response variable is identical to the linear predictor, rather than to non-linear function of the linear predictor. That is, for the normal linear model, the link function

$$g(\mu_i) = \mu_i. \tag{3.4}$$

The GLM is an extension of the linear model to include response variables that follow any probability distribution in the exponential family of distributions. Many commonly used distributions in the exponential family are the normal, binomial, Poisson, exponential, gamma and inverse Gaussian distributions. In addition, several other distributions are in the exponential family and they include the Beta, multinomial, Dirichlet, and Pareto. There are other several distributions which are not in the exponential family but are used for statistical modelling and they include the student's t and uniform distributions.

The exponential family

GLMs may be used to model variables following distributions in the exponential family with density functions

$$f(y;\theta,\varphi) = exp\left\{\frac{y\theta - b(\theta)}{a(\varphi)} + c(y;\varphi)\right\} \tag{3.5}$$

or,

$$\boldsymbol{log}(y;\theta,\varphi) = \frac{y\theta - b(\theta)}{a(\varphi)} + \boldsymbol{c}(y;\varphi), \tag{3.6}$$

where φ is a dispersion parameter $a(\varphi)$, $b(\theta)$ and $\boldsymbol{c}(y;\varphi)$ are known functions,for distributions in the exponential families, the conditional variable of Y is a function of the mean, μ together with a dispersion parameter, φ, that is,

$$\boldsymbol{E}(Y_i) = \mu_i = b^{'}(\theta) \tag{3.7}$$

and

$$\boldsymbol{var}(Y_i) = \sigma_i^2 = b^{''}(\theta)a(\varphi), \tag{3.8}$$

where $b^{'}(\theta)$ and $b^{''}(\theta)$ are the first and second derivatives of $b(\theta)$. The dispersion parameter is usually fixed to one for some distributions.

The link function

In theory, link function $\eta_i = g(\mu_i)$ can be any monotonic, differentiable function. In practice, only a small set of link functions are actually utilised. In particular, links are chosen such that the *inverse link* $\mu_i = g^{-1}(\eta_i)$ is easily computed, and so that g^{-1} maps from $X_i\beta = \eta_i \in \Theta$ into the set of admissible values for μ_i. A log link is usually used for Poisson model, since while $\mu_i = g(\mu_i) \in \Theta$, because Y_i is a count, we have $\mu_i \in 0, 1, ...$ For binomial data, the link function maps from $0 < \mu_i < 1$ to $\mu_i \in \Theta$.

3.4 Model components

The canonical treatment of GLMs is from McCullagh and Nelder (1989), and this review closely follows their notation and approach we begin by considering

the familiar linear regression model,

$$Y_i = X_i\beta + \epsilon_i, \tag{3.9}$$

where $i = 1, 2, ..., n$: Y_i is a dependent variable, $X_i\beta$ is a vector of k independent variables or predictors, β is a k-by-1 vector of unknown parameters and ϵ_i are zero-mean stochastic disturbances. Typically, the ϵ_i are assumed to be independent across observations with constant variance σ_i^2 , and distributed normally. That is, the normal linear regression model is characterised by the following features

1. **Random component**: the Y_i are assumed to have independent normal distribution with $E(Y_i) = \mu_i$, with constant variance σ^2, or Y_i $i\tilde{i}dN(\mu, \sigma^2)$ if it is not normally distributed then $y_i \sim P(\mu_i)$.

2. **Systematic component**: specifies the explanatory or independent variables for the model: $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p$. The covariates x_i combine linearly with the coefficients to form the linear predictor

$$\eta_i = X_i\beta. \tag{3.10}$$

3. **Link between the random and systematic components**: the linear predictor $\eta_i = X_i\beta$ is a function of the mean parameter μ_i via a *link* function, $g(\mu_i)$. It should be noted that for normal linear model g is an identity.

3.4.1 Linear predictor

The log-linear model

Suppose that we have a sample of n observations, $y_1, y_2, y_3, ..., y_n$ which can be treated as realisations of independent Poisson random variables, with $y_i \sim P(\mu_i)$, and suppose that we want to let the μ_i (and therefore the variance) de-

pend on a vector of explanatory variables x_i. We could entertain a simple linear model of the form

$$\mu_i = x_i^{'}\beta, \tag{3.11}$$

but this model has the disadvantage that the linear predictor on the right hand side can assume any real value, whereas the Poisson mean on the left hand side, which represent an expected count, has to be non-negative. A straightforward solution to this problem is to use the model of logarithm of the mean instead of using a linear model. Thus, we take logs calculating

$$\eta_i = \mathit{log}(\mu_i), \tag{3.12}$$

and assume that the transformed mean follows a linear model

$$\mu_i = x_i^{'}\beta, \tag{3.13}$$

thus, we consider a generalised linear model with $link\ log$. Combining these two steps in one we can write the log-linear model as

$$\mathit{log}(\mu_i) = x_i^{'}\beta, \tag{3.14}$$

in this model the regression coefficient β_j represents the expected change in the log of the mean per unit change in the predictor x_j. In other words, increasing x_j by one unit is associated with an increase β_j in the log of the mean.
Exponentiating equation, we obtain a multiplicative model for the mean:

$$\mu_i = \mathit{exp}(x_i^{'}\beta), \tag{3.15}$$

in this model, an exponentiated regression coefficient $\mathit{exp}(\beta_j)$ represents a multiplicative effect of the j^{th} predictor on the mean. Increasing x_j by one unit multiplies the mean by a factor $\mathit{exp}(\beta_j)$.

A further advantage of using the log link stems from the empirical observation that with count data the effect of predictors are often multiplicative rather than additive. That is, one typically observes small effects for small counts, and large effects for large counts. If the effect is in fact proportional to the count, working in the log scale leads to much simpler model.

3.4.2 Link function

Fisher scoring log-linear model

Fisher scoring algorithm is a form of Newton-Raphson method used in statistics to solve maximum likelihood equations numerically. Nelder and Wedderburn (1972) applied Fisher scoring algorithm to estimate $\hat{\beta}$ in generalised linear models. The Fisher scoring algorithm for Poisson regression models with canonical link would be considered, where it would be modelled as

$$\eta_i = g(\mu_i) = log(\mu_i), \tag{3.16}$$

the derivative of the link is easily seen to be

$$g^{'} = \frac{1}{\mu_i}, \tag{3.17}$$

specifically, given an initial estimate β, the algorithm update it to be β^{new} by

$$\beta^{new} = \beta + \left\{ \boldsymbol{E}\Big(- \frac{\partial L}{\partial \beta \partial \beta^T} \Big) \right\}^{-1} \frac{\partial L}{\partial \beta}, \tag{3.18}$$

where both derivatives are evaluated at β, and the expectation is evaluated as if β were the true parameter values. β is then replaced by β^{new} and the updating is repeated until convergence. It can be shown that for a GLM, the updating

equation can be rewritten as

$$\beta^{new} = \beta + \left(\mathbf{X}^T\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{W}_z, \tag{3.19}$$

where, z is the n-vector with i^{th} component

$$z_i = \left(Y_i - \mu_i\right)y^{'}\mu_i \tag{3.20}$$

and, $\mathbf{W}$ is the $n \times n$ diagonal matrix with

$$\mathbf{W}_i = \left\{g^{'}(\mu_i)^2 b^{''}\right\}^{-1}, \tag{3.21}$$

$$\mathbf{W}_i = \left\{(\mu_i)\frac{1}{\mu_i^2}\right\}^{-1} \tag{3.22}$$

and this simplifies to

$$\mathbf{W}_i = \mu_i. \tag{3.23}$$

3.4.3 Probability distributions

The Poisson regression model

The Poisson distribution is a discrete probability distribution that represents the probability of a certain number of events occurring in a given amount of time provided these events occur at a specified average rate and each count occurs independently of the time since the previous event. The Poisson distribution may also be used to calculate the number of occurrences in various intervals (Kutner et al., 2005).

Poisson regression is a technique for describing count data as a function of a set of explanatory variables (Lee, 1986). It has been widely employed in human and veterinary epidemiological research to explore the incidence and mortality of chronic illnesses throughout the previous two decades (Gardner et al.,

1995). Poisson regression has also been used to estimate injury mortality in many places of the world when analysing mortality data. Poisson regression has been used to compare exposed and unexposed populations and to determine the causes of injury mortality, among other things.

Simeon-Denis Poisson (1781–1840) was the first to introduce the Poisson distribution, which he described in his probability theory in 1838. The work focused on a set of random variables N that count, 31 among other things, the number of discrete occurrences (also known as "arrivals") that occur during a particular time interval, (Haight, 1967). If the expected number of occurrences in this interval is μ_i , then the probability that there are exactly y_i occurrences (y_i being a non-negative integer, $y_i = 0, 1, 2, ...$) is equal to

$$f(y_i, \mu_i) = \frac{\mu_i^{y_i} e^{-\mu_i}}{y_i!}, \tag{3.24}$$

where
y_i is the number of occurrences of an event- the probability of which is given by the function $f(y_i, \mu_i)$, μ_i is a positive real number, equal to the expected number of occurrences that occur during the given interval.

The parameter μ_i is only the mean number of occurrences, y_i but also its variance

$$\sigma_{y_i}^2 = \boldsymbol{E}(y_i^2) - [\boldsymbol{E}(y_i)]^2 = \mu_i, \tag{3.25}$$

thus, the number of observed occurrences fluctuates about its mean, μ_i with a standard deviation

$$\sigma_{y_i} = \sqrt{\mu_i}, \tag{3.26}$$

as a function of y_i , this is the discrete probability mass function. The Poisson distribution can be derived as a limiting case of the binomial distribution. The Poisson distribution can be applied to system with a large number of possible events, each of which is rare. The Poisson distribution is sometimes called a Poissonian, analogous to the term Gaussian for Gauss or normal distribution.

Assumptions of Poisson distribution are:

- Observations are independent.
- Probability of occurrence in a short interval is proportional to the length of the interval.
- Probability of another occurrence in such a short interval is zero.

Verification of Poisson distribution as a member of exponential family:

The Poisson distribution belongs to the exponential family as defined by Nelder and Wedderburn, (Nelder and Wedderburn, 1972). Taking logarithm of the Poisson distribution function, we obtain

$$log\, f_i(y_i) = y_i log(\mu_i) - \mu_i - log(y_i!), \tag{3.27}$$

where
y_i is the number of occurrences of the count or event,the probability of which is $f(y_i)$. μ_i is the expected number of occurrences that occur during the given interval. Looking at the coefficient of y_i we observe immediately that the link function is $log(\mu_i)$ and the canonical parameter (θ_i) is given by

$$\theta_i = log(\mu_i), \tag{3.28}$$

therefore, the canonical link is the logarithm. Solving for μ_i we obtain the inverse link

$$\mu_i = e^{\theta_i} \tag{3.29}$$

and we see that we can write the second term in equation 3.27 as

$$b(\theta_i) = e^{\theta_i}, \tag{3.30}$$

the last term is a function of y_i only, so we identify from equation 3.27 that

$$c(y_i, \varphi) = log(y_i), \tag{3.31}$$

finally, it should be noted that we take the dispersion parameter ($\varphi = 1$), just as it is in the binomial case and we verify that Poisson distribution belongs to the exponential family.

Verification of equal mean and variance:

Differentiating the cumulation function $b(\theta_i)$ we have

$$\mu_i = b^{'}(\theta) = e^{\theta_i} = \mu_i \tag{3.32}$$

and differentiating again we have

$$\sigma_i^2 = b^{''}(\theta) = e^{\theta_i} = \mu_i, \tag{3.33}$$

hence, the mean is equal to the variance. In spite of its recent wide application, Poisson regression model remains partly poorly known especially if compared with other regression techniques, like linear, logistic and Cox regression models (Kutner et al., 2005). The Poisson regression model assumes that the sample of n observations, y_i are observations on independent Poisson variables Y_i with

mean μ_i. If this model is correct, the equal variance assumption of classic linear regression is violated, since the Y_i have means equal to their variance. So we fit the generalised linear model,

$$\boldsymbol{log}(\mu_i) = x_i^{'}\beta, \tag{3.34}$$

we say that the Poisson regression model is generalised linear model with Poisson error and a log link, so that

$$\mu_i = \boldsymbol{exp}(x_i^{'}\beta), \tag{3.35}$$

this implies that one unit increases in an x_i are associated with multiplication of μ_i by $\boldsymbol{exp}(\beta_i)$.

Model identification

The primary equation of the model is

$$P(Y_i = y_i) = \frac{e^{-\mu}\mu^{y_i}}{y_i!}, y_i = 1, 2, ..., \tag{3.36}$$

the most common formulation of this model is the log-linear specification as in equation

$$\boldsymbol{log}(\mu_i) = x_i^{'}\beta, \tag{3.37}$$

the expected number of events per period is given by

$$\boldsymbol{E}(y_i|x_i) = \mu_i = \boldsymbol{e}^{x_i^{'}\beta}, \tag{3.38}$$

thus

$$\frac{d\boldsymbol{E}(y_i|x_i)}{dx_i} = \beta e^{x_i^{'}\beta} = \beta_i\mu_i. \tag{3.39}$$

The major assumption of Poisson model is

$$\boldsymbol{E}(y_i|x_i) = \mu_i = e^{x_i^{'}\beta} = Var(y_i|x_i). \tag{3.40}$$

This assumption would be tested. According to Hilbe (2011), If, $Var(y_i|x_i) > \boldsymbol{E}(y_i|x_i)$ then there is overdispersion.If, $Var(y_i|x_i) < \boldsymbol{E}(y_i|x_i)$ then there is underdispersion.

The Negative Binomial model

The Negative Binomial regression is more flexible than the Poisson and is frequently used to study count data with overdispersion (Hilbe, 2011). In fact, the Negative Binomial regression model is in many ways equivalent to the Poisson regression model because the Negative Binomial model could be viewed as a Poisson-gamma mixture model (Hilbe, 2011). However, the difference is that the Negative Binomial regression model had a free dispersion parameter. In other words, the Poisson regression model can be considered as a Negative Binomial regression with an ancillary or heterogeneity. In the Negative Binomial regression model, a random parameter reflecting unexplained between subject differences is included Gardner et al. (1995), that is, the negative binomial regression adds an overdispersion parameter to estimate the possible deviation of the variance from the expected value under Poisson regression.

The major assumption of the Poisson model in equation 3.40 is that the Poisson model does not allow for over or underdispersion. A richer model is obtained by using the Negative Binomial distribution instead of Poisson distribution. Instead of equation

$$P(Y_i = y_i) = \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}, \tag{3.41}$$

For mathematical convenience, the gamma distribution is usually assumed for $\mu_i = e^{x_i^{'}\beta}$ (Berk and MacDonald, 2008). After normalising the gamma distribu-

tion to have an expected value, one arrives at

$$g(\mu_i) = \frac{\theta^\theta}{\Gamma(\theta)} e^{-\theta\mu_i} \mu_i^{\theta-1}, \tag{3.42}$$

where Γ denotes the gamma distribution, and θ is a parameter to be specified a prior or estimated. Intergrating over μ_i, the density for y_i, conditional only on the predictors, is

$$f(y_i|x_i) = \frac{\Gamma(\theta + y_i)}{\Gamma(y_i + 1)\Gamma(\theta)} \left(\frac{\mu_i}{\mu_i + \theta}\right)^{y_i} \left(1 - \frac{\mu_i}{\mu_i + \theta}\right)^{\theta}, \tag{3.43}$$

this Negative Binomial distribution can be shown to have conditional mean μ_i and conditional variance $\mu_i\Big(1 + \eta^2\mu_i\Big)$ with $\eta^2 = \frac{1}{\theta}$. Note that the parameter η^2 is not allowed to vary over the observations. As before, the conditional mean function is modelled as

$$\boldsymbol{E}(y_i|x_i) = \mu_i = e^{x_i^{'}\beta}. \tag{3.44}$$

The conditional variance function is the given by

$$Var(y_i|x_i) = e^{x_i^{'}\beta}\Big(1 + \eta^2 e^{x_i^{'}\beta}\Big). \tag{3.45}$$

Using maximum likelihood, we can then estimate the regression parameter β, and also the extra parameter η. The parameter η measures the degree of over (or under) dispersion. The limit case $\eta = 0$ corresponds to the Poisson model.

3.5 Maximum likelihood estimation

Maximum likelihood estimation (MLE) involves estimating the regression parameters specifically using the maximum likelihood estimation (Myung, 2003). The likelihood function for n independent Poisson observations is a product of

probabilities given by

$$\boldsymbol{L}(\theta|X,Y) = \prod_{i=a}^{b} \frac{e^{y_i\theta' x}\boldsymbol{e}^{-e^{\theta' x}}}{y!}, \tag{3.46}$$

if $\boldsymbol{prob}(y_i) = \frac{e^{-\mu_i}\mu_i^{y_i}}{y_i!}$ is the probability function of Poisson distribution. Taking logarithm of equation 3.45 and ignoring the constant involving $\boldsymbol{log}(y_i!)$, we find that the log-likelihood function as

$$\boldsymbol{log}\ \boldsymbol{L}(\beta) = \sum_{i=0}^{n}[-\mu_i + y_i x_i^{'}\beta] \tag{3.47}$$

$$= \sum_{i=1}^{n}[-e^{x_i^{'}\beta} + y_i x_i^{'}\beta], \tag{3.48}$$

where $\mu_i = e^{x_i^{'}\beta}$ (Kane, 1948). The parameters of this equation can be estimated using maximum likelihood method

$$\frac{\partial L}{\partial \beta} = \sum_{i=1}^{n}(y_i - e^{x_i^{'}\beta})x_i = 0 \tag{3.49}$$

and

$$\frac{\partial^2 L}{\partial \beta \partial \beta'} = -\sum_{i=1}^{n}(e^{x_i^{'}\beta}x_i^{'}x_i), \tag{3.50}$$

which is the Hessian of the function and with typical element

$$\frac{\partial^2 L}{\partial \beta \partial \beta'} = -\sum_{i=1}^{n}(e^{x_i^{'}\beta}x_{ij}x_{il}); j,l = 1,2,...,p, \tag{3.51}$$

as

$$\frac{\partial^2 L}{\partial \beta_j \partial \beta_l} = -\sum_{i=1}^{n}(e^{x_i^{'}\beta}x_{ij}x_{il}), \tag{3.52}$$

does not involve the y data

$$k_{jl} = \boldsymbol{E}\left(\frac{\partial^2 L}{\partial \beta_j \partial \beta_l}\right) = -\sum_{i=1}^{n}(e^{x_i^{'}\beta}x_{ij}x_{il}); j,l = 1,2,...,p \tag{3.53}$$

and the information matrix is

$$\boldsymbol{K} = \sum_{i=1}^{n} (e^{x_i^{'}\beta} x_i^{'} x_i), \tag{3.54}$$

there is no closed form solution to,

$$\frac{\partial L}{\partial \beta} = \sum_{i=1}^{n} (y_i - e^{x_i^{'}\beta}) x_i = 0, \tag{3.55}$$

so the MLE for β must be obtained numerically, However, as the Hessian is negative definite for all x and β, the MLE($\hat{\beta}$) is unique, if it exists.
From

$$\frac{\partial^2 L}{\partial \beta_j \partial \beta_l} = -\sum_{i=1}^{n} (e^{x_i^{'}\beta} x_{ij} x_{il}) \; and \; k_{jl} = \boldsymbol{E}\left(\frac{\partial^2 L}{\partial \beta_j \partial \beta_l} \right) = -\sum_{i=1}^{n} (e^{x_i^{'}\beta} x_{ij} x_{il}), \tag{3.56}$$

$$k_{jlr} = \boldsymbol{E}\left(\frac{\partial^3 L}{\partial \beta_j \partial \beta_l \partial \beta_r} \right) = -\sum_{i=1}^{n} (e^{x_i^{'}\beta} x_{ij} x_{il} x_{ir}) \tag{3.57}$$

and

$$k_{jl}^{(r)} = (\frac{\partial kjl}{\partial \beta_r}) = -\sum_{i=1}^{n} (e^{x_i^{'}\beta} x_{ij} x_{il} x_{ir}), j, l, r = 1, 2, 3, ..., p. \tag{3.58}$$

To make matters more transparent, consider the case of a single covariate and an intercept. Then x_i is a scalar observation and

$$\boldsymbol{L} = \sum_{i=1}^{n} \big[-\mu_i + y_i(\beta_1 + \beta_2 x_i) - \boldsymbol{log}(y_i) \big], \tag{3.59}$$

where $\mu_i = \boldsymbol{exp}(\beta_1 + \beta_2)$, for $i = 1, 2, 3, ..., n$. The first order conditions, $\frac{\partial L}{\partial \beta} = 0$ yield a system of K equations (one for each β) of the form

$$\frac{\partial L}{\partial \beta} = \sum_{i=1}^{n} (y_i - e^{x_i^{'}\beta}) x_i = 0, \tag{3.60}$$

where $\hat{y}_i = e^{x_i^{'}\hat{\beta}}$ is the fitted value of y_i. The predicted/fitted value has as usual been taken as the estimated value $\left(y_i|x_i\right)$. This first order condition tells us that the vector of residual is orthogonal to the vectors of explicative variables.

3.6 Test of hypotheses

Likelihood ratio tests for log-linear models can easily be constructed in terms of deviances. In general, the differences in deviances between two nested models has approximately in large samples a chi-square distribution with degrees of freedom equal to the difference in the number of parameters between the models, under the assumption that the smaller model is correct. One can also construct Wald tests, based on the fact that maximum likelihood estimator $\hat{\beta}$ has approximately in large samples of a multivariate normal distribution with mean equal to the true parameter value β and variance-covariance matrix,

$$var(\hat{\beta}) = X^{'}WX, \tag{3.61}$$

where X is the model matrix and W is the diagonal matrix of estimation weights.

3.6.1 Likelihood ratio test

A likelihood test on the slopes serves as a sample test on the overall fit of the model and is analogous to the F-test in the traditional regression model (Woolf, 1957). The model with only the intercept is nothing but the mean of the counts, or

$$\mu_1 = \bar{y}, \tag{3.62}$$

where

$$\bar{y} = \sum_{i=1}^{n} \frac{y_i}{n}, \tag{3.63}$$

the corresponding log-likelihood is

$$L_R = n\bar{y} + log(\bar{y}) \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} log(y_i!), \tag{3.64}$$

where R stands for the restricted mode, as opposed to the unrestricted model with $K-1$ slope parameters. the last term in $\sum_{i=1}^{n} log(y_i!)$ can be dropped, as long as it is also dropped in the calculation of the maximum likelihood

$$L_u = \sum_{i=1}^{n} \Big[- e^{x_i^{'}\beta} + y_i x_i^{'} \beta - log(y_i!) \Big], \tag{3.65}$$

for the unrestricted model L_u using

$$L = e^{x_i^{'}\hat{\beta}_t}, \tag{3.66}$$

the likelihood ratio test is then

$$LR = 2\big(L_u - L_R\big), \tag{3.67}$$

which follows a χ^2 distribution with $K-1$ degrees of freedom.

3.6.2 Goodness of fit test

In order to assess the adequacy of the Poisson regression model you should first look at the basic descriptive statistics for the event count data. If the count mean and variance are significantly different (equivalent in a Poisson distribution) then the model is likely to be overdispersed or underdispersed. The model analysis option gives a scale parameter (sp) as a measure of overdispersion; this is equal to the Pearson chi-square statistic divided by the number of observations minus the number of parameters (covariates and intercept). Underdispersion is very uncommon to various forms of count data especially with

accident data.

The variances of the coefficients can be adjusted by multiplying by $sp.$ the goodness of fit test statistics and residuals can be adjusted by dividing by $sp.$ Using a quasi-likelihood approach sp could be integrated with regression, but this would assume a known fixed value for sp, which is seldom the case. A better approach to overdispersed Poisson models is to use a parametric alternative model, the negative binomial.

The deviance (likelihood ratio) test statistic, G^2, is the most useful summary of the adequacy of the fitted model (Woolf, 1957). It represents the change in deviance between the fitted model and model with a constant term and no covariates; therefore G^2 is not calculated if no constant is specified. If this test is significant then the covariates contribute significantly to the model.

The deviance goodness of fit test reflects the fit of the data to Poisson distribution in the regression. If this test is significant then a red asterisk is shown by the $p - value$, and you should consider other covariates and other error distributions such as Negative Binomial.
The deviance function is

$$deviance = 2\sum_{i=1}^{n} y_i ln\left[\frac{y_i}{\hat{\mu}_i}\right] - \left(y_i - \hat{\mu}_i\right), \tag{3.68}$$

where y is the number of events, n is the number of observations and $\hat{\mu}_i$ is the fitted Poisson mean. The first term is identical to the binomial deviance, representing twice a sum of observed times log of observed over fitted. The second term, a sum of differences between observed and fitted values, is usually zero, because MLE's in Poisson models have the property of reproduction. The second term, a sum of differences between observed and fitted values, is usually zero,

because MLE's in Poisson models have the property of reproducing marginal totals, as noted above.

The log-likelihood function is

$$L = \sum_{i=1}^{n} y_i ln\big(\hat{\mu}_i\big) - \hat{\mu}_i - ln(y_i!), \tag{3.69}$$

the maximum likelihood regression proceeds by iteratively re-weighted least squares, using singular value decomposition to solve the linear system at each iteration, until the change in deviance is within the specified accuracy. The Pearson Chi-square residual is

$$r_p = \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}, \tag{3.70}$$

for large samples the distribution of the deviance is approximately a chi-squares with $n - p$ degrees of freedom, where n is the number of observations and p the number of parameters. Thus, deviance can be used directly to test the goodness of fit of the model. An alternative measure of goodness of fit is Pearson's chi-squared statistic, which is defined as the Pearson goodness of fit test statistic is

$$\chi^2 = \sum_{i=1}^{n} \frac{y_i - \mu_i}{\sqrt{\hat{\mu}_i}}, \tag{3.71}$$

the deviance residual is (Cook and Weisberg, 1982)

$$r_d = sign(y_i - \hat{\mu}_i)\sqrt{deviance(y_i, \hat{\mu}_i)}, \tag{3.72}$$

the Freeman-Turkey, variance stabilized, residual is (Freeman and Tukey, 1950)

$$r_{ft} = \sqrt{y_i} + sqrty_i + 1 - sqrty_i\hat{\mu}_i, \tag{3.73}$$

the standardized residual is

$$r_s = \frac{(y_i - \hat{\mu}_i)}{\sqrt{1 - h_i}}, \tag{3.74}$$

where h is the leverage (diagonal of the hat matrix).

3.7 Time series models

The time series modelling of IM data in this study will be done using the Box-Jenkins autoregressive integrated moving average (ARIMA) and Poisson integer-valued generalised autoregressive conditional heteroscedasticity (INGARCH) or Negative Binomial integer-valued generalised autoregressive conditional heteroscedasticity (NB-INGARCH) models. The advantage of these models over other models are that Poisson INGARCH and NB-INGARCH has the ability to handle discreteness in count data whereas the ARIMA model has the strength which results from its distributional assumption underlying the estimation process.

3.8 ARIMA model

ARIMA is one of the most traditional methods of non-stationary time series analysis. In contrast to the regression models, the ARIMA model allows time series to be explained by its past, or lagged values and stochastic error terms. An ARIMA model can be understood by outlining each of its components as provided in the succeeding subsections.

3.8.1 Autoregressive model

The model x_t(*injury mortality*) is autoregression (AR) of order p if there exists constants $\phi_1, \phi_2, ..., \phi_p$ such that

$$x_t = \phi_1 x_{t-1} + ... + \phi_p x_{t-p} + z_t, \tag{3.75}$$

where z_t is a white noise with mean zero and constant σ_z^2.

3.8.2 Moving average

The model x_t is moving average (MA) of order q if there exists constants $\theta_1, \theta_2, ..., \theta_q$ such that

$$x_t = \theta_1 z_{t-1} + ... + \theta_q z_{t-q} + z_t, \tag{3.76}$$

where z_t is a white noise with mean zero and constant σ_z^2.

Autoregressive moving average

The autoregressive moving average (ARMA(p, q)) model containing AR and MA is said to be an ARMA model of order p and q, respectively, such that

$$x_t = \phi_1 x_{t-1} + ... + \phi_p x_{t-p} + \theta_1 z_{t-1} + ... + \theta_q z_{t-q}. \tag{3.77}$$

The equation 3.76 can also be written using the backshift operator B, such that

$$\phi(B) x_t = \theta(B) z_t, \tag{3.78}$$

where $\phi(B)$, $\theta(B)$ are polynomial of order p, q.

3.8.3 Autoregressive integrated moving average

The ARMA model can be used only with stationary time series data. In the case of non-stationarity, the autoregressive integrated moving average model (ARIMA), which is a time-series forecasting approach that is used in predicting the future value of a variable from its own past value, will be used for non-stationary time series data (Pankratz, 2009). It uses autoregression (AR) and moving average (MA), and incorporates a differencing d order to remove trend and /or seasonality. The model is expressed with the following equation

$$\phi_p(B)(1-B)^d x_t = \theta_q(B) z_t, \tag{3.79}$$

selecting appropriate values for parameter p and q requires testing and optimisation. To choose the values, one must inspect visual observations for the data to determine trend and/or seasonality. Visualisation in the form of auto-correlation function (ACF) and partial auto-correlation function (PACF) chart have proven to be useful in determining the values for parameters p and q. However, in case large numbers need to be forecasted, manual visual inspection will not be possible. One solution for this problem is to incorporate the use of software packages that facilitate automated selection of ARIMA's parameters and chooses the best suited one.

3.9 Poisson INGARCH model

3.9.1 ARCH model

Lets us start with the autoregressive conditional heteroskedasticity (ARCH) models, which were developed by R.F Engle III (Engle, 1982). The ARCH models are motivated by a specific drawback of the AR models. If looking at the

conditional mean and variance of an AR(p) process, then

$$E\big[X_t|X_{t-1}, X_{t-2}, \cdots\big] = \alpha_1 X_{t-1} + \cdots + \alpha_p X_{t-p} + \mu_\epsilon, \tag{3.80}$$

varies in time according to the last p observations, while

$$V\big[X_t|X_{t-1}, X_{t-2}, \cdots\big] = \sigma_\epsilon^2, \tag{3.81}$$

is constant in time. But it is common to observe clusters of large or low volatility, a phenomenon that cannot be reproduced by the AR models.
Let $\big(\epsilon_t\big)_{\mathbb{Z}}$ be square-integrable white noise $E\big[\epsilon_t\big] = 0$ and $V\big[\epsilon_t\big] = 1$. Then $\big(X_t\big)_{\mathbb{Z}}$ defined by

$$X_t = \sigma_t \cdot \epsilon_t, \tag{3.82}$$

where $\sigma_t^2 = \beta_0 + \alpha_1 X_{t-1}^2 + \cdots + \alpha_p X_{t-p}^2$ with $\beta_0, \alpha_p > 0$ and $\alpha_1, \cdots, \alpha_{p-1} \geq 0$, and where ϵ_t is required to be independent of $\big(X_t\big)_{s<t}$(causality), is said to be an ARCH(p) process.
As a result, we obtain the time-varying condition variances

$$V\big[X_t|X_{t-1}, X_{t-2}, \cdots\big] = \sigma_t^2 = \beta_0 + \alpha_1 X_{t-1}^2 + \cdots + \alpha_p X_{t-p}^2, \tag{3.83}$$

so now, the AR(p)-like recursion is not applied to the observation but to their conditional variance. In contrast, the conditional variance remains constant in time provided that the requirement $\sum_{j=1}^{p} \alpha_j < 1$ is satisfied. In fact constant condition again guarantees the existence of a unique causal (weakly) stationary solution of the ARCH recursion. So although an ARCH(p) process is obviously not serially independent by construction, it is serially uncorrelated. However, looking at the process of squared observations, autocorrelation becomes visible. If the weakly stationary and causal ARCH(p) process $\big(X_t\big)_{\mathbb{Z}}$ has existing fourth-order moments, there we can represent the squared process $(X_{t-p}^2)_{\mathbb{Z}}$ by an AR(p)-

like recursion,

$$X_t^2 = \alpha_1 X_{t-1}^2 + \cdots + \alpha_p X_{t-p}^2 + v_t, \tag{3.84}$$

with the $\big(v_t\big)_{\mathbb{Z}}$ being weak white noise having the mean $E\big[v_t\big] = \beta_0$. Therefore the autocorrelated function of the squared process satisfies the Yule-Walker given by

$$\rho(k) = \sum_{i=1}^{p} \alpha_i \rho(|k-i|), \tag{3.85}$$

for $k = 1, 2, ...$

3.9.2 GARCH model

A few years after the of ARCH model, Engle's student T. Bollerslev proposed the generalised autoregressive conditional heteroskedasticity (GARCH) model, where the conditional variance not only depended on the past observations but also on past conditional variances (Bollerslev, 1986). So for GARCH process of order (p, q), abbreviated as GARCH(p, q) process, the recursion

$$\sigma_t^2 = \beta_0 + \alpha_1 X_{t-1}^2 + \cdots + \alpha_p X_{t-p}^2 + \beta_1 \sigma_{t-1}^2 + \cdots + \beta_q \sigma_{t-q}^2, \tag{3.86}$$

is required to be satisfied. The condition for the existence of a weakly stationary and causal solution now become $\alpha_1 + \cdots + \alpha_p + \beta_1 + \cdots + \beta_q < 1$.

3.9.3 INGARCH model

The ordinary ARMA models cannot be used for count process $\big(X_t\big)_{\mathbb{Z}}$, as they are not able to preserve of the discrete range because of "multiplication problem" (Fokianos, 2012). A way to circumventing this problem is to define a count process model by linear regression of the conditional means

$$M_t := E\big[X_t | X_{t-1}, X_{t-2} ...\big] \tag{3.87}$$

if the count at time t is generated using a count distribution having mean M_t, it is guaranteed that the outcomes are integer values. Note that this approach also shares analogies with an ARCH model, where the autoregression is defined at the level of conditional variance. ARCH models are generalised beyond pure autoregression by also including past conditional variances into the model recusion to give GARCH.

In the abovementioned count context, one may hence include the previous condition mean as a "feedback" term. This leads to the definition of the integer-valued generalised autoregressive conditional heteroscedasticity (INGARCH), abbreviated as INGARCH(p, q) model with $p \geq 1$ and $q \geq 0$, which is based on the assumption of the conditional mean M_t satisfying

$$M_t = \beta_0 + \sum_{i=1}^{p} \alpha_i X_{t-i} + \sum_{j=1}^{q} \beta_j M_{t-j}, \tag{3.88}$$

with $\beta_0 > 0$ and $\alpha_1, \cdots, \alpha_p, \beta_1, \cdots, \beta_q \geq 0$ (if $q = 0$, we call it an INARCH(p) model). Depending on the choice of the conditional distribution family, different INGARCH models are obtained.

3.9.4 Poisson INGARCH model

Count time series $x_1, \cdots, x_T$ are discret-valued time series having a quantitative range consisting of nonnegative intergers from $\mathbb{N}_0 = 0, 1, \cdots$ (Fokianos et al., 2009). The set of all integers is denoted $\mathbb{Z} = \cdots, -1, 0, 1$.
Let $\big(X_t\big)_{\mathbb{Z}}$ be a process with range $\mathbb{N}_0$. The process $\big(X_t\big)_{\mathbb{Z}}$ follows the Poisson INGARCH(p, q) model with $p \geq 1$ and $q \geq 0$

- X_t, conditioned on $X_{t-1}, X_{t-2}, \cdots$, is Poisson distributed according to Poi(M_t), where
- the conditional mean $M_t := E\big[X_t | X_{t-1}, X_{t-2}, \cdots\big]$ satisfies

$M_t = \beta_0 + \sum_{i=1}^{p} \alpha_i X_{t-i} + \sum_{j=1}^{q} \beta_j M_{t-j}$,

with $\beta_0 > 0$ and $\alpha_1, \cdots, \alpha_p, \beta_1, \cdots, \beta_q \geq 0$.

The default choice is a conditional Poisson distribution, i.e., X_t, conditioned on $X_{t-i}, \cdots$, is Poisson distributed according to $Poi(M_t)$, (Ferland et al., 2006). The resulting Poisson INGARCH model is sometimes also referred to as a linear Poisson autoregressive model (Fokianos et al., 2009). This model has been discussed by several authors including Ferland et al. (2006), Fokianos et al. (2009) and Weiß (2009). The stochastic properties of Poisson INGARCH model have been derived by Ferland et al. (2006) and Weiß (2009). For

$$\alpha_\bullet + \beta_\bullet := \sum_{i=1}^{p} \alpha_i + \sum_{j=1}^{q} \beta_j < 1, \tag{3.89}$$

the Poisson INGARCH process exists and is strictly stationary, with finite first- and and second-order moments (Weiß, 2009). For $p = q = 1$, all moments exist (Ferland et al., 2006) and mixing properties have been established by Neumann (2011). Due to linear conditional mean, the unconditional mean equals

$$\mu = \frac{\beta_0}{1 - \alpha_\bullet - \beta_\bullet}, \tag{3.90}$$

and the variance

$$V\big[X_t\big] = \mu + V\big[M_t\big], \tag{3.91}$$

and autocovariances can be computed by solving a set of Yule-Walker equations (Weiß, 2009):

$$\gamma(k) = \sum_{i=1}^{p} \alpha_i \gamma(|k-i|) + \sum_{j=1}^{min(k-1,q)} \beta_j \gamma(k-j) + \sum_{j=k}^{q} \beta_j \gamma_M(j-k), \tag{3.92}$$

$\gamma_M(l) = \sum_{i=1}^{min(l,p)} \alpha_i \gamma_M(|l-i|) + \sum_{i=l+1}^{p} \alpha_i \gamma(i-l) + \sum_{j=1}^{q} \beta_j \gamma_M(|l-j|)$,

where $\gamma(k) := Cov\big[X_t, X_{t-k}\big]$ and $\gamma_M(l) := Cov\big[M_t, M_{t-l}\big]$ (Weiß, 2009).
Despite the fact that the conditional Poisson distribution is equidispersed (variance equal the mean), the unconditional distribution exhibits overdispersion, i.e., the dispersion ratio $\frac{\sigma^2}{\mu} > 1$. In the purely autoregressive case of an INARCH(p) model (i.e., if $q = 0$), the Yule-Walker equations imply that ARMA-like autocorrelation function (ACF) satisfies

$$\rho(k) = \sum_{i=1}^{p} \alpha_i \rho(|k-i|), \tag{3.93}$$

so except the restriction to non-negative coefficients α_i, equation 3.91 is identical to the Yule-Walker equations of an ordinary AR(p) model. Consequently, the model order of an INARCH model can be identified by using the ACF and PACF.

3.9.5 Negative Binomial INGARCH model

As an alternative to the conditional Poisson distribution, Zhu (2011) and Christou and Fokianos (2014) considered the Negative Binomial distribution $\boldsymbol{NB}(r, p_t)$ with the parameters $r > 0$ and $p_t = \frac{r}{\lambda_t + r}$ where λ_t is, for instance, of the form equation 3.89. We still have $E(X_t | X_u, u < t) = \lambda_t$, but the conditional variance $\frac{\lambda_t + \lambda_t^2}{r}$ is larger than the conditional variance of the Poisson case, which reflects the conditional overdispersion that is suspected to be present on real series (Christou and Fokianos, 2014). The condition on equation 3.88 entails the existence of an ergodic and strictly stationary X_t. In the case $(p, q) = (1, 1)$, it can be shown (Christou and Fokianos, 2014), that the stationary solution is such that $E(X_t^2) < \infty$ if and only if

$$\big(\alpha_0 + \beta_0\big)^2 + \frac{\alpha_0^2}{r} < 1, \tag{3.94}$$

writing α_0 and β_0 instead of α_{01} and β_{01}. Always in the case $(p, q) = (1, 1)$, it can be shown that $E(X_t^4) < \infty$ if and only if

$$\big(\alpha_0 + \beta_0\big)^4 + \frac{6\alpha_0^2\big(\alpha_0 + \beta_0\big)^2}{r} + \frac{\alpha_0^3\big(11\alpha_0 + 8\beta_0\big)^2}{r^2} + \frac{6\alpha_0^4}{r^3} < 1. \tag{3.95}$$

The conditions ensuring the existence of $E(X_t^2)$ are much more complicated for the general orders p and q , (Zhu, 2011).

3.9.6 Quasi maximum likelihood estimation

The estimation of the model's parameter was determined by quasi-conditional maximum likelihood estimation (QMLE) as explained by Ahmad and Francq (2016). This is denoted by

$$\theta = (\beta_0, \beta_1, \cdots, \beta_p, \alpha_1, \cdots, \alpha_q, \eta_1, \eta_2, \cdots, \eta_r)^T, \tag{3.96}$$

is the vector of regression parameter and the parameter space for the INGARCH model regardless of the distributional assumption which is taken to be

$$\Theta = \left\{\theta \in \mathbb{R}^{p+q+r+1} : \beta_0 > 0, \beta_1, \cdots, \beta_p, \alpha_1, \cdots, \alpha_q, \eta_1, \cdots, \eta_r \geq 0, \sum_{k=1}^{p} \beta_k + \sum_{l=1}^{q} \alpha_l < 1\right\}, \tag{3.97}$$

the intercept β_0 is essential to be positive while all other parameter must be nonnegative to ensure positivity of the conditional mean M_t. For the log-linear model, the parameter space is taken to be

$$\Theta = \left\{\theta \in \mathbb{R}^{p+q+r+1} : |\beta_1|, \cdots, |\beta_p|, |\alpha_1|, \cdots, |\alpha_q| < 1, \left|\sum_{k=1}^{p} \beta_k + \sum_{l=1}^{q} \alpha_l\right| < 1\right\}. \tag{3.98}$$

According to Mütze et al. (2019), the estimation to be negative binomial parameter does not rely on the additional dispersion parameter, ϕ. This allows utilising a quasi-maximum likelihood approach based on the Poisson Likelihood to

estimate the regression parameter, θ. The QMLE approach is preferred for simplicity and its practicality on deriving consistent estimators when the model for M_t has been correctly specified. The conditional quasi log-likelihood function up to a constant is as follow:

$$l(\theta) = \sum_{t=1}^{n} logp_t(y_t;\theta) = \sum_{t=1}^{n}(y_t ln(M_t(\theta)) - M_t(\theta)), \tag{3.99}$$

where $p_t(y;\theta)$ is the probability density function of a Poisson distribution. In case of the Poisson assumption it holds if σ_2, where σ_2 is the variance. The QMLE of θ is the solution of non-linear constrained optimisation problem.

$$\hat{\theta} = argmax_{\theta\in\Theta} l(\theta). \tag{3.100}$$

3.10 Test for normality

Normality tests are used to deduce whether a set of data is adequately modelled by the normal distribution and how probable it is for a random variable underlying the data to be normally distributed (Subramoney et al., 2021). The hypotheses tested are,
H_0 : Data is normally distributed.
H_1 : Data is not normally distributed.
The normal distribution's most important properties are that it is symmetric (zero skewness) and has a kurtosis of three.

3.10.1 Jarque-Bera test

The well-known test for normality of Jarque and Bera (JB) is a goodness-of-fit test that compares the difference of the skewness and kurtosis of sample data to those of the normal distribution (Thadewald and Büning, 2007). The JB test

statistic is defined as,

$$JB = \frac{N}{6}\left[s^2 + \frac{(K-3)^2}{4}\right], \tag{3.101}$$

with S, K and N denoting the sample skewness, sample kurtosis and sample size, respectively. For large sample sizes, the test statistic is compared to a chi-squared distribution with two degrees of freedom, i.e., $\chi^2(2)$. Normality is rejected if the test statistic is greater than chi-squared value. The chi-squared approximation needs fairly large sample sizes in order for it to be accurate.

3.10.2 Shapiro-Wilk test

The Shapiro-Wilk (SW) test has been found to be the most powerful normality test and is highly recommended by researchers as the best choice for testing the normality of a set of data (Ghasemi and Zahediasl, 2012). The test statistic is given by,

$$W = \frac{(\sum_{i=1}^{n} a_i x_i)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \tag{3.102}$$

where x_i are the ordered random sample value and a_i are constants generated from the covariances, variance and means of the sample of size n. For small values of W, it indicates that the sample is not normally distributed and thus the null hypothesis is rejected. This test also has limitations in a sense that the test has bias by sample sample size (the larger the sample, the more accurate the results).

3.10.3 Quantile-Quantile plot

The Quantile-Quantile (Q-Q) plot is a graphical device used to assess the validity of the theoretical distributional assumption for a particular data set, for example a normal distribution or an exponential distribution (Velez and Morales, 2015). Generally, the main idea is to calculate the theoretically expected values for each data point based on the distribution in question. If the data points fall

approximately on a straight line, the data will consequently follow the assumed distribution. This approach is merely a visual check and thus subjective, but it enables one to get an idea whether the assumed distribution is plausible.

3.11 Testing for unit root and stationarity

The statistical basis for estimation and forecasting depends on a time series being covariance stationary. However, a great amount of economic and financial time series exhibit characteristics of time series that is non-stationary. Thus, prior to further analysis, identifying the form of the trend and thereafter removing it from the time series is an essential task (Zivot and Wang, 2003). Trend removal methods depend on whether a time series is trend stationary or difference stationary. Unit root tests can be used to distinguish whether a trending time series should be differenced or regressed against deterministic functions of time.

3.11.1 Unit root test

Unit root test is used to test whether a time series variable is a non-stationary (Fuller, 2009). In time series analysis, a unit root arises when either autoregressive or moving average polynomials of an ARIMA model has a unit root is in a unit circle. A unit root that has either of these polynomial has important implications for modelling. Consider AR(1)

$$x_t - x_{t-1} = \big(\rho - 1\big)x_{t-1} + z_t, \tag{3.103}$$

$$\nabla x_t = \phi x_{t-1} + z_t, \tag{3.104}$$

where, $z_t \sim WN(0, \sigma_z^2)$

$$\tau_{nc} = \frac{\phi}{SE(\phi)}, \tag{3.105}$$

τ_{nc} is test statistic with no constant, n is number of observations. where $SE(\phi)$ is the standard error of the coefficient

$$SE(\phi) = \sqrt{\frac{s^2}{\sum_{t=1}^{n}\left(x_{t-1} - \bar{x}\right)}}, \tag{3.106}$$

$$s^2 = \sum_{t=1}^{n} \frac{(\nabla x_t - \phi x_{t-1})^2}{n-3}, \tag{3.107}$$

s^2 is a sample variance,τ_{nc} is a test statistic with a no constant n is number of observations and $\bar{x}$ is the sample mean of $x_1, x_2, \cdots, x_{n-1}$. we reject the null hypotheses if $\tau_{nc} > t_{\frac{\alpha}{2},n-1}$ where $t_{\frac{\alpha}{2},n-1}$ is tabulated value with the degrees of freedom $n-1$.

3.11.2 Dickey-Fuller test

The Dickey-Fuller test is used to check whether a unit root is present in an autoregressive model and whether or not the data should be differenced (Phillips and Perron, 1988). To test for unit root with drift, the AR(1) is considered and is given by

$$\nabla x_t = \alpha_0 + \phi x_{t-1} + z_t, \tag{3.108}$$

$$\tau_{cd} = \frac{\phi}{SE(\phi)}, \tag{3.109}$$

where $SE(\phi)$ is the standard error of the coefficient

$$SE(\phi) = \sqrt{\frac{s^2}{\sum_{i=1}^{n}\left(x_{t-1} - \bar{x}\right)}}, \tag{3.110}$$

$$s^2 = \sum_{i=1}^{n} \frac{(\nabla x_t - \alpha_0 - \phi x_{t-1})^2}{n-3}, \tag{3.111}$$

s^2 is a sample variance,τ_{cd} is a test statistic with a constant, n is number of observations and $\bar{x}$ is the sample mean of $x_1, x_2, \cdots, x_{n-1}$. we reject the null hypotheses if $\tau_{cd} > t_{\frac{\alpha}{2},n-1}$ where $t_{\frac{\alpha}{2},n-1}$ is tabulated value with the degrees of freedom $n-1$.

3.11.3 Augmented Dickey-Fuller test

Augmented Dickey-Fuller (ADF) is used to test for a unit root with a drift and deterministic time trend

$$\nabla x_t = \alpha_0 + \alpha_1 + \phi x_{t-1} + z_t, \tag{3.112}$$

$$\tau_{ct} = \frac{\phi}{SE(\phi)}, \tag{3.113}$$

where $SE(\phi)$ is the standard error of the coefficient

$$SE(\phi) = \sqrt{\frac{s^2}{\sum_{i=1}^{n} \left(x_{t-1} - \bar{x}\right)}}, \tag{3.114}$$

$$s^2 = \sum_{t=1}^{n} \frac{(\nabla x_t - \alpha_0 + \alpha_1 - \phi x_{t-1})^2}{n-3}, \tag{3.115}$$

s^2 is a sample variance,τ_{ct} is a test statistic with a drift, n is number of observations and $\bar{x}$ is the sample mean of $x_1, x_2, \cdots, x_{n-1}$. we reject the null hypotheses if $\tau_{ct} > t_{\frac{\alpha}{2},n-1}$ where $t_{\frac{\alpha}{2},n-1}$ is tabulated value with the degrees of freedom $n-1$.

3.11.4 Phillips-Perron test

Phillips and Perron (PP) proposed non-parametric test statistic that rectify autocorrelation and heteroscedasticity in the errors (Phillips and Perron, 1988). It is a modification of the ADF test in a sense that autocorrelation does not affect

the asymptotic distribution of the test statistic. Using an AR(1) model to illustrate the non-parametric modified test statistic for the three cases are given as follows,

Case I: *Constant only (Models with drift)*

$$Z_p = T(\hat{\rho} - 1) - \frac{(s^2 - s_e^2)}{2T^{-2}\sum_{t=1}^{T}(y_{t-1} - \bar{y}_{-1})^2}, \tag{3.116}$$

where $\bar{y}_{-1} = \frac{\sum_{i=1}^{T-1} y_t}{T-1}$

Case II: *Trend only (No drift)*

$$Z_p = T(\hat{\rho} - 1) - \frac{(s^2 - s_e^2)}{2T^{-2}\sum_{t=1}^{T} y_{t-1}^2}, \tag{3.117}$$

$$Z_t = \frac{s_e}{s} t_{\hat{p}} - \frac{1}{2}\frac{(s^2 - s_e^2)}{s(T^{-2}\sum_{t=1}^{T} y_{t-1}^2)^{\frac{1}{2}}}, \tag{3.118}$$

where, $D_X = det(X^{'}X)$ and the regression are $X = (1, t, y_{t-1})$. The consistent estimates of variance parameters are defined as follows,

$$s_e^2 = T^{-1}\sum_{t=1}^{T} e_t^2 \qquad s^2 = \lim_{T\to+\infty}\sum_{i=1}^{T} E\left(\frac{1}{T}\sum_{t=1}^{T} e_t^2\right). \tag{3.119}$$

The PP test tends to be more powerful than the ADF test and is more robust to general forms of heteroscedasticity found in the error terms. However, test has been observed to have serious size distortion when autocorrelations of e_t are negative and is more prone to model misspecification, in other words, order of ARMA model.

3.11.5 Kwiatkowski-Phillips-Schmidt-Shin test

The Kwiatkowski, Phillips, Schmidts and Shin (KPSS) test, contrary to the two unit root tests mentioned above, tests for stationarity as the null hypothesis. The absence of unit root in KPSS test is not evidence of stationarity, but rather

proof of stationarity around a trend. This is a distinguishing feature of the test, since it depicts that it is possible for a time series to be non-stationary and possess no unit root and yet be trend-stationary (Shin and Schmidt, 1992).
Consider the model,

$$y_t = \Delta t + \zeta_t + e_t, \tag{3.120}$$

where e_t is a stationary process and ζ_t is a random walk given by,

$$\zeta_t = \zeta_{t-1} + \mu_t \qquad \mu_t \sim iid(0, \sigma_\mu^2), \tag{3.121}$$

the test statistic is as follows,

$$KPSS = \frac{1}{T^2} \frac{\sum_{t=1}^{T} S_t^2}{\hat{\sigma}_\infty^2}, \tag{3.122}$$

where, $S_t = \sum_{i=1}^{t} \hat{e}_i$ is a partial sum and $\hat{\sigma}_\infty^2$ is an heteroscedasticity and autocorrelation consistent (HAC) estimator of the variance of $\hat{e}_t$. This is Lagrange multiplier (LM) test for constant parameters. As opposed to testing the null hypothesis of stationarity to trend stationarity, the test is formulated similarly, except that the error terms are obtained as residuals from the regression of y_t on an intercept only.

3.12 Model identification

Model identification helps to decide which model is most appropriate for a given set of data. The following steps outline an approach to this problem.

- Plot the time series. Identify if there is any unusual observations. Decide if a transformation is necessary to stabilize the variance, if necessary, transform the data to achieve stationarity in the variance.
- Consider if the (possibly transformed) data appear stationary from ACF and PACF time plots. If the time plot shows the data scattered horizon-

tally around a constant mean, or equivalently, the ACF and PACF drop to or near zero quickly, which indicates that the data are stationary. If the time plot is not horizontal, or the ACF and PACF do not drop to zero, non-stationarity is implied.

- When the data appear non-stationary, it can be made stationary by differencing. For non-seasonal data, take the first differences of the data. For seasonal data, take seasonal differences of data. Check if these still appear stationary. If there are still non-stationary take first difference of the differenced data.

- When stationary has been achieved, the autocorrelation is examined to see if any pattern remains.

3.12.1 Autocorrelation function

Autocorrelation function (ACF) of a stationary process with mean μ and variance σ^2 and covariance γ_k, then ACF is defined as

$$\rho_{(k)} = \frac{\gamma_k}{\gamma_0} = \frac{\gamma_k}{\sigma^2}, \tag{3.123}$$

where γ_0 is the variance of the series and γ_k is the covariance of lag k.

3.12.2 Partial autocorrelation function

Partial autocorrelation function (PACF) is the difference between x_t and x_{t+k} with their linear dependency on the intervening variable $x_{t+k}, \cdots, x_{t+k-1}$. The PACF is defined as

$$\Phi_{kk} = corr(x_t, x_{t+k}|x_{t+1}, \cdots, x_{t+k-1}), \tag{3.124}$$

for $k = 1, 2, 3$.

3.13 Parameter estimation

Parameter estimation deals with the problem of estimating the parameters of an ARIMA model once has been specified. The commonly used method of estimating parameters are: method of moments estimates (MME), least square estimates (LSE) and maximum likelihood estimates (MLE).

3.13.1 Method of moments estimates

The method of moments consists of equating sample moments to corresponding theoretical moments and solving the results equations to obtain unknown parameters. The estimates of γ_k and ρ_k are as follows

$$\gamma_k = \frac{1}{2}\sum_{t=1}^{n-k}\big(x_t - \bar{x}\big)\big(x_{t+k} - \bar{x}\big), \tag{3.125}$$

$$\rho_k = \frac{\sum_{t=1}^{n-k}\big(x_t - \bar{x}\big)\big(x_{t+k} - \bar{x}\big)}{\sum_{t=1}^{n}\big(x_t - \bar{x}\big)^2}. \tag{3.126}$$

3.13.2 Least square estimates

The method of least square estimation is an estimation procedure developed for standard regression models. The least square estimation for

$$y_t = \phi_1 x_{t-1} - x_t, \tag{3.127}$$

for $t = 1, 2, 3, \cdots, n$ is denoted by

$$\hat{\phi} = \frac{n\sum_{t=1}^{n} x_t y_t - \sum_{t=1}^{n} x_t \sum_{t=1}^{n} y_t}{n\sum_{t=1}^{n} x_t^2 - \big(\sum_{t=1}^{n} x_t\big)^2}, \tag{3.128}$$

where the estimate ϕ is best linear unbiased estimator of ϕ.

3.13.3 Maximum likelihood estimates

The maximum likelihood is a function of unknown parameters in the model with the observed data held fixed. For ARIMA models, likelihood function L will be a function of the Φ's, Θ's and σ_z^2 given the observations $y_1, \cdots, y_n$. The advantage of maximum likelihood is that all of the information in the data is most probable rather than just the first and second moments.

3.14 Model diagnostics

Model diagnosis is used to test the goodness of fit of a model. If the fit is poor, suggests appropriate modifications. There are two complementary approaches that can be used for analysis: analysis of residuals from the fitted model and analysis of over parameterised models.

3.14.1 Residual analysis

Residual analysis can be used to check if the model is correctly specified and if the parameter estimates are reasonably close to the true values. The assumption that it should have the same properties of white noise. That is, it satisfies the properties

- Independence
- Identically
- Normally distributed random variables with zero mean and common variance σ_z^2

3.14.2 The Ljung-Box test

The Ljung-Box is a test that takes into account the magnitude of residuals autocorrelations as a group to check for model adequacy (Ljung and Box, 1978).

The test is given by

$$Q = N \sum_{k=1}^{k} \gamma_{z,k}^2, \tag{3.129}$$

alternative test is given by

$$Q = N(N+2) \sum_{k=1}^{k} \left(\frac{\gamma_{z,k}^2}{N-K} \right), \tag{3.130}$$

where N is the number of terms in the differenced series and k is the number of lags and (z, k) denote the autocorrelation at lag k of the residual $\hat{z}_t$
If $Q > \chi_{k,1-\alpha}^2$ we reject null hypothesis and conclude that the random term z_t from the estimates model are correlated and that the estimated model may be adequate.

3.14.3 The Lagrange multiplier (LM) test

A time series that displays conditional heteroscedasticity and / or autocorrelation in the squared series is said to process autoregressive conditional heteroscedastic (ARCH) effects. Engle's ARCH test is a Langrange multiplier test to investigate the importance of these ARCH effects. consider a residual series defined as,

$$e_i = y_t - \hat{\mu}_t, \tag{3.131}$$

where $y_t = \mu_t + \epsilon_t$ is a time series with μ_t being the conditional mean of the process and ϵ_t is the innovation with zero mean and unit variance. If autocorrelation in the original time series is accounted for, then the residuals will be correlated with mean zero. However, there may still be a possibility that the residuals are serially dependent. The null hypothesis for the test is that there are no ARCH effects and it is tested as follows:
$H_0 : \alpha_1 = \alpha_2 = ... = \alpha_m = 0$ (No ARCH effects among the lags considered)
$H_1 : \epsilon_t^2 = \alpha_0 = \alpha_1 \epsilon_{t-1}^2 + ... + \alpha_m \epsilon_{t-m}^2 + \mu_t$ (at least one of the α_i coefficients is

significant), where μ_t is a white noise error process.

In order to conduct the test, the lag m needs to be specified. This can be chosen by comparing log-likelihood values for different choices of the lag length. The log-likelihood ratio test or the Akaike information criterion/Bayesian information criterion can be used to compare values.The test statistic for the ARCH test is the general F statistic for the regression on the squared residuals. Under the null hypothesis, the F statistic follows a chi-square distribution with m degrees of freedom. A significantly large critical value indicates rejection of the null hypothesis.

3.15 Model assessment

The main purpose of this section is finding model that is adequate representations of the observed data. However, there are models that all fits the observed data to a similar degree, making it difficult to choose which model is the best. There are many statistical methods developed to search for the best model, namely; the stepwise regression, likelihood ratio tests, Alkaike information criterion (AIC) and Bayesian information criterion (BIC). This study wishes to only concentrate on the AIC and BIC because the first two methods has some limitations when comparing more than one models.

3.15.1 Akaike information criterion

Akaike (1973) define Akaike information criterion (AIC) as a useful statistic for statistical model selection and evaluation. The procedure was developed by Akaike which is given by:

$$\mathrm{AIC} = -2\mathrm{log}(\mathrm{L}) + 2\mathrm{K}, \tag{3.132}$$

where K is the number of parameters in the model and L is the likelihood function. One important advantage of AIC is that, it is simple and easy to use. Furthermore, another important aspect of AIC is that, the best model chosen does not imply the true model but in fact it means the model is best among competing models. The selection rule state that, the best model will be the one with the lower value of AIC.

3.15.2 Bayesian information criterion

The development of Bayesian information criterion (BIC) use the idea of AIC, by early 1978 Glideon Schwarz added a penalty term to the AIC equation in which resulted in the procedure called the Bayesian information criterion (BIC), that is:

$$\text{BIC} = -2\text{log}(L) + \text{K}\text{log}(n), \tag{3.133}$$

where L is the maximised value of the likelihood function, n is the number of observations and K is the number of parameters in the model. The selection rule state that, the best model will be the one with the lower value of BIC.

3.16 Forecasting

Forecasting is the prediction of the future values based on available data. The objective of forecasting is to produce an optimum forecast that has no error which leads to the concept of minimum square error forecast (Abraham and Ledolter, 2009). This forecast will produce an optimum future value with the minimum error of mean square error.

3.16.1 Mean square error

Mean square error (MSE) is a measure of average square deviation of forecasted values. Like mean absolute percentage error (MAPE), the opposite signed er-

rors do not offset one another, MSE gives an overall idea of the error occurred during forecasting (Prasad and Rao, 1990). It penalised extreme errors occurred while forecasting. The MSE does not provide any idea about the direction of overall errors and sensitive to the change of scale and data transformations. The mean square is given by

$$MSE = \frac{1}{n}\sum_{t=1}^{n}|e_t^2|, \tag{3.134}$$

where $(e_t = y_t - \hat{y}_t)$ is forecast error, y_t is the vector of the observed value of the variable being predicted and $\hat{y}_t$ is the predicted value.

3.16.2 Mean absolute error

Mean absolute error (MAE) measures the average absolute deviation of forecasted values from the original ones also known as mean absolute deviation (MAD). It shows the magnitute of the overall error, occurred due to forecasting (Willmott and Matsuura, 2005). The MAE is defined by

$$MSE = \frac{1}{n}\sum_{t=1}^{n}|e_t|, \tag{3.135}$$

where $(e_t = y_t - \hat{y}_t)$ is forecast error, y_t is the vector of the observed value of the variable being predicted and $\hat{y}_t$ is the predicted value.

3.16.3 Mean absolute percentage error

The Mean absolute percentage error (MAPE) represents the percentage of the average absolute error occurred. It is independent of the scale measurement, but affected by data transformations (De Myttenaere et al., 2016). It does not show the direction of the error. It also does not penalise the extreme deviations.

The mean absolute percentage error is given by

$$MAPE = \left(\frac{1}{n} \sum_{t=1}^{n} \left| \frac{e_t}{y_t} \right| \right) \times 100\%, \tag{3.136}$$

where $(e_t = y_t - \hat{y}_t)$ and y_t are forecast error and data at time t respectively.

3.16.4 Root mean square error

Root mean square error (RMSE) is just the square root of MSE. All properties of MSE holds also RMSE (Prasad and Rao, 1990). The root mean square error is given by

$$RMSE = \sqrt{\frac{1}{n} \sum_{t=1}^{n} |e_t^2|}, \tag{3.137}$$

where $(e_t = y_t - \hat{y}_t)$ is forecast error, y_t is the vector of the observed value of the variable being predicted and $\hat{y}_t$ is the predicted value.

Chapter 4

Results and discussion

4.1 Introduction

This chapter focuses on the preliminary analysis of data on the number of people on injury mortality in both Gauteng (GP) and Mpumalanga (MP) provinces; and the detailed analysis of data using various statistical tools such as Poisson integer-valued generalised autoregressive conditional heteroscedasticity (INGARCH), Negative Binomial (NB) integer-valued generalised autoregressive conditional heteroscedasticity (INGARCH) and Box-Jenkins autoregressive integrated moving average (ARIMA).

4.2 Data description

This dissertation was based on secondary data obtained from Statistics South Africa (StatsSA). The study considered causes of death only on injury for the period 2008-2018. The number of people died by injury was used as the response variable in all the models. The time series data cover the whole of GP and MP provinces on monthly injury mortality. SAS, R and SPSS programming software

were used for data analysis.

4.3 Descriptive statistics

In this section, we present the distribution of injury mortality and summary statistics. The total injury mortality for the year 2008 to 2018 (11 years) is 78627. Therefore, the comparison of injury mortality of GP and MP is presented on Table 4.1.

Table 4.1: Injury mortality on provinces

Province	Number of deaths	Percentage
MP	15264	19.4
GP	63363	80.6

Table 4.1 reveals that GP has the highest proportion of 80.6% deaths due to injury as compared to 19.4% of MP. In addition, Table 4.2 shows the total number of deaths and percentages in years for which injury mortality occurred in MP between 2008 and 2018.

Table 4.2: Injury mortality from year 2008-2018 for MP

Year	Number of deaths	Percentage
2008	1523	10
2009	990	6.5
2010	1031	6.8
2011	1235	8.1
2012	1313	8.6
2013	1488	9.7
2014	1723	**11.3**
2015	1533	10
2016	1523	10
2017	1454	9.5
2018	1451	9.5

Table 4.2 reveals that year 2008 recorded 1523 (10%) deaths due to injury. However, it can be observed that there is massive decrease from 1523 in 2008 to

990 (6.5%) in 2009. In each year of 2010, 2011, 2012 , 2013 and 2014, there was steady increase in the number of deaths due to injury, ranging from 1031 (6.8%), 1235 (8.1%), 1313 (8.6%), 1488 (9.7%) and 1723 (11.3%), respectively. In contrast, in 2015, 2016, 2017 and 2018, there was gradual decrease in the number of deaths due to injury, ranging from 1533 (10%), 1523 (10%), 1454 (9.5%) and 1451 (9.5%), respectively. Moreover, 2014 was a year with highest number of deaths due to injury, 1723 (11.3%) while 2009 had the lowest, 990 (6.5%). Generally, the distribution of number of deaths due to injury since 2008 till 2018 shows an upward and downward trend over time. Subsequently, Table 4.3 shows the total number of deaths and percentages in years for the injury mortality which occurred in GP between 2008 and 2018.

Table 4.3: Injury mortality from year 2008-2018 for GP

Year	Number of deaths	Percentage
2008	6312	10
2009	6021	9.5
2010	5757	9.1
2011	5393	8.5
2012	5610	8.9
2013	6352	10
2014	6726	**10.6**
2015	6057	9.6
2016	5371	8.5
2017	4929	7.8
2018	4835	7.6

The results in Table 4.3 reveal that year 2008 recorded 6312 (10%) deaths due to injury. However, it can be observed that there is a steady decrease in the year 2009, 2010 and 2011 in number of deaths due to injury, ranging from 6021 (9.5%), 5757 (9.1%) and 5393 (8.5%), respectively. In a year 2012, 2013 and 2014, there was another gradual increase in the number of deaths due to injury, ranging from 5610 (8.9%), 6352 (10%) to 6726 (10.6%), respectively; but in 2015, 2016, 2017 and 2018, there is another steady decrease in the number of deaths due to injury, ranging from 6057 (9.6%), 5371 (8.5%), 4929 (7.8%) and 4835

(7.6%), respectively.

Table 4.4: Injury mortality from year 2008-2018 for MP and GP

Year	*(GP)	Percentage	*(MP)	Percentage	*Total	Percentage
2008	6312	8.03	1523	1.94	7835	9.97
2009	6021	7.66	990	1.26	7011	8.92
2010	5757	7.32	1031	1.31	6788	8.63
2011	5393	6.86	1235	1.57	6628	8.43
2012	5610	7.13	1313	1.67	6923	8.8
2013	6352	8.08	1488	1.89	7840	9.97
2014	6726	8.55	1723	2.19	8449	**10.74**
2015	6057	7.70	1533	1.95	7590	9.65
2016	5371	6.83	1523	1.94	6894	8.77
2017	4929	6.27	1454	1.85	6383	8.12
2018	4835	6.15	1451	1.85	6286	8

* Number of deaths.

Table 4.4 shows 7835 (9.97%) injury-related deaths in 2008. However, the number of injury-related deaths was 7011 (8.92%), 6788 (8.63%), and 6628 in 2009, 2010, and 2011, respectively (8.43%). The number of deaths from injuries had steadily increased in 2012, 2013 and 2014 with 6923 (8.8%), 7840 (9.97%) and 8449 (10.74%), respectively; but in 2015, 2016, 2017 and 2018 there was another steady decline in the number of deaths from injuries, from 7590 (9.65%), 6894 (8.77%), 6383 (8.12%) and 6286 (8%). In addition, 2014 was a year with a high number of deaths from injuries: 8449 (10.74%) and 2018 was the lowest: 6286 (8%). Consequently, GP has the highest proportion of 8.55% deaths from injuries compared to 2.19% in MP. In general, the distribution of the number of injury-related deaths from 2008 to 2018 shows an upward and downward trend over time. This result is not surprising since Matzopoulos et al. (2015) showed that injury mortality decreased from 2004 to 2011. StatsSA (2021) showed that 2014 had the highest injury mortality and between 2010 and 2015 there was a steady increase in the proportion of deaths from injury mortality among GP.

Table 4.5: Injury mortality from January to December for MP

Month	Number of deaths	Percentage
January	1205	7.9
February	1215	8
March	1279	8.4
April	1171	7.7
May	1103	7.2
June	1158	7.6
July	1319	8.6
August	1372	9
September	1280	8.4
October	1371	9
November	1207	7.9
December	1584	**10.4**

Table 4.5 shows that 1205 (7.9%) persons died as a result of an injury in January. Conversely, the number of deaths from injuries increased in February and March, ranging from 1215 (8%) to 1279 (8.4%), respectively. In April and May, the number of deaths due to injury decreased steadily, from 1171 (7.7%) to 1103 (7.2%), respectively. However, the number of injury-related deaths increased steadily in June, July, and August, with 1158 (7.6%), 1316 (8.6%) and 1372 (9%), respectively. Additionally, there was a decrease and increase in the number of deaths due to injury in September, October, November, and December, ranging from 1280 (8.4%), 1371 (9%), 1207 (7.9%) and 1584 (10.4%), respectively. Meanwhile, December had the highest number of deaths due to injury: 1584 (10.4%), while May had the lowest: 1103 (7.2%). In general, the distribution of injury-related deaths from January to December shows an upward and downward tendency throughout time. In addition, Table 4.6 shows the number of deaths and percentages for which injury mortality occurred in GP between January and December.

Table 4.6: Injury mortality from January to December for GP

Month	Number of deaths	Percentage
January	4565	7.2
February	4634	7.3
March	5240	8.3
April	4834	7.6
May	4976	7.9
June	5265	8.3
July	5602	8.8
August	5666	8.9
September	5500	8.7
October	5515	8.7
November	5568	8.8
December	5998	**9.5**

In January, there were 4565 (7.2%) injury-related deaths, according to Table 4.6. Consequently, the number of deaths due to injury increased in February and March, with 4634 (7.3%) and 5240 (8.3%), respectively. The number of injuries that resulted in death decreased by 4834 (7.6%) in April. Meanwhile, there was another constant increase in the number of deaths due to injury in May, June, July, and August, ranging from 4976 (7.9%), 5265 (8.3%), 5602 (8.8%), and 5666 (8.9%), respectively. Furthermore, the number of injury-related deaths decreased by 5,500 (8.7%) in September. There was an increase in the number of injury-related deaths in October, November, and December, ranging from 5515 (8.7%) to from 5515 (8.7%), 5568 (8.8%), and 5998 (9.5%), respectively. The month with the most injury deaths was December, with 5998 (9.5%), and January had the fewest, with 4565 (7.2%). In general, the distribution of the number of deaths due to injury from January to December shows an increase and decreasing tendency throughout time.

Table 4.7: Injury mortality from month January to December for MP and GP

Month	*(GP)	Percentage	*(MP)	Percentage	*Total	Percentage
January	4565	5.81	1205	1.53	5770	7.34
February	4634	5.89	1215	1.55	5849	7.44
March	5240	6.66	1279	1.63	6519	8.29
April	4834	6.15	1171	1.49	6005	7.64
May	4976	6.33	1103	1.40	6079	7.73
June	5265	6.70	1158	1.47	6423	8.17
July	5602	7.12	1319	1.68	6921	8.8
August	5666	7.21	1372	1.74	7038	8.95
September	5500	7.00	1280	1.63	6780	8.63
October	5515	7.01	1371	1.74	6886	8.75
November	5568	7.08	1207	1.54	6775	8.62
December	5998	7.63	1584	2.01	7582	**9.64**

* Number of deaths.

The total number of deaths and percentages by month for which injury mortality occurred in both MP and GP combined between January and December are shown in Table 4.7. According to Table 4.7, there were 57770 (7.34%) deaths due to injury in January. Conversely, there is an increase in the number of deaths due to injury in the months of February and March, with 5849 (7.44%) and 6519 (8.29%), respectively. In April, there was a decrease of 6005 (7.64%) in the number of injuries deaths. Moreover, there was a steady increase in the number of deaths due to injury in the months of May, June, July, and August, ranging from 6079 (7.73%), 6423 (8.17%), 6921 (8.8%) and 7038 (8.95%), respectively. Additionally, the number of deaths due to injury decreased and increased in September, October, November, and December, with 6780 (8.63%), 6886 (8.75%), 6775 (8.62%), and 7582 (9.65%), respectively. As a result, GP has the largest proportion of 7582 (7.63%) deaths due to injury, while MP has the lowest rate of 2.01%. In general, the distribution of injury-related deaths exhibits an increase and decrease tendency over time. This is consistent with the fact that December has the highest injury mortality rate (StatsSA, 2021).

Table 4.8: Injury mortality gender for MP

Gender	Number of deaths	Percentage
Males	11506	75.4
Females	3758	24.6

The results in Table 4.8 reveal that males recorded 11506 (75.4%) deaths due to injury, while females recorded 3758 (24.6%) deaths due to injury. Furthermore, Table 4.9 shows the number of deaths and percentages of injury mortality in GP by gender.

Table 4.9: Injury mortality gender for GP

Gender	Number of deaths	Percentage
Males	49801	78.6
Females	13562	21.4

Table 4.9 reveals that males, recorded 49801 (78.6%) deaths due to injury, while females recorded 13562 (21.4%) deaths due to injury.

Table 4.10: Injury mortality on gender for MP and GP

Gender	*(GP)	Percentage	*(MP)	Percentage	*Total	Percentage
Males	49801	63.39	11506	14.63	61307	78.02
Females	13562	17.25	3758	4.78	17320	22.03

* Number of deaths.

Table 4.10 shows the total number of deaths and gender percentages for injury mortality in MP and GP. According to Table 4.10, males accounted for 61307 (78.02%) of all injuries, while females accounted for 17320 (22.03%). The proportion of mortality attributable to injury in GP is 63.39%, compared to 14.63% in MP. This is similar with the findings of Matzopoulos et al. (2015), who found that males injury mortality was consistently and considerably greater than females. Males accounted for a higher percentage than females, according to Abio et al. (2020) and (StatsSA, 2021).

Table 4.11: Injury mortality on age-group for MP

Age-group	Number of deaths	Percentage
0-14 years old	1951	12.8
15-34 years old	7003	**45.9**
35-54 years old	4217	27.6
55-74 years old	1642	10.8
75+ years old	451	3

The distribution of the number of deaths due to injury in MP is shown in Table 4.11. According to the findings, the age group 0-14 years old had 1951 (12.8%) deaths due to injury. The age group 15-34, on the other hand, had the highest number of injury-related deaths, with 7003 (45.9%). Conversely, there was a steady decrease in the number of deaths due to injury in the age groups 35-54, 55-74, and 75+, with 4217 (27.6%), 1642 (10.8%), and 451 (3%), respectively. In addition, Table 4.12 shows the number of deaths and percentages for which injury mortality occurred in GP for age-group.

Table 4.12: Injury mortality on age-group for GP

Age-group	Number of deaths	Percentage
0-14 years old	5480	8.6
15-34 years old	30186	**47.6**
35-54 years old	18928	29.9
55-74 years old	6688	10.6
75+ years old	2081	3.3

Table 4.12 reveals that the age group 0-14 years old had 5480 (8.6%) deaths due to injury in GP. Meanwhile, the age group 15-34 recorded the highest number of deaths due to injury, 30186 (47.6%). Furthermore, there was a steady decrease in the number of injury mortality in the age-groups 35-54, 55-74, and 75+, with 18928 (29.9%), 6688 (10.6%), and 2081 (3.3%), respectively.

Table 4.13: Injury mortality on age-group for MP and GP

Age-group	*(GP)	Percentage	*(MP)	Percentage	*Total	Percentage
0-14 years old	5480	6.97	1951	2.48	7431	9.45
15-34 years old	30186	38.39	7003	8.91	37189	**47.3**
35-54 years old	18928	24.07	4217	5.36	23145	29.43
55-74 years old	6688	8.51	1642	2.09	8330	10.6
75+ years old	2081	2.65	451	0.57	2532	3.22

* Number of deaths.

Table 4.13 presents the total number of deaths and gender percentages for injury mortality in MP and GP. According to Table 4.13, the age group 0-14 years old had 7431 (9.45%) deaths due to injury. Meanwhile, the age group 15-34 had the highest number of deaths due to injury, with 37189 (47.3%) deaths. Furthermore, there was a steady decrease in the number of deaths due to injury in the age-groups 35-54, 55-74, and 75+, ranging from 23145 (29.43%), 8330 (10.6%), and 2532 (3.22%), respectively. The proportion of mortality due to injury in GP is 38.39%, compared to 8.91% in MP. This is consistent with the WHO (2021) 2021 report, which states that people aged 5 to 29 had the highest injury rate. This is consistent with the WHO (2021) 2021 report that those aged 5-29 have the highest injury mortality rate, and Meel (2017) found that 64% of deaths occurred between the ages of 11-44.

Table 4.14: Causes of injury mortality for MP

Causes of death	Number of deaths	Percentage
Sequelae of external causes of mortality	22	0.1
Homicide	3279	21.5
Other external causes of accidental injury	1488	9.7
Complications of medical and surgical care	545	3.6
Drowning	681	4.5
Falls	52	0.3
Suffocation	412	2.7
Forces of nature	287	1.9
Fire/Burns	991	6.5
Transport accidents	4247	**27.8**
Poisoning	1497	9.8
Suicide	1758	11.5
Starvation	5	0

The distribution of the number of deaths due to injury in MP is shown in Table 4.14. The findings show that transport accidents resulted in 4247 (27.8%) deaths due to injury, the largest number of deaths. This was followed by 3279 (21.5%) homicides and suicides 1758 (11.5%) of the deaths that were caused by injury. Starvation was the cause of injury death with the fewest injuries, 5 (less than 1%).

Table 4.15: Causes of injury mortality for GP

Causes of death	Number of deaths	Percentage
Sequelae of external causes of mortality	149	0.2
Homicide	21850	**34.5**
Other external causes of accidental injury	3875	6.1
Complications of medical and surgical care	3228	5.1
Drowning	1607	2.5
Falls	241	0.4
Suffocation	1139	1.8
Forces of nature	1037	1.6
Fire/Burns	5045	8.0
Transport accidents	12741	20.1
Poisoning	4811	7.6
Suicide	7619	12.0
Starvation	21	0.0

The number of deaths and percentages of injury mortality in GP for the main causes of injury mortality are shown in Table 4.15. According to the findings in Table 4.15, homicide recorded 21850 (34.5%) deaths due to injury in MP, the highest number of deaths. Following that were 12741 (20.1%) transport accidents and 7619 (12%) suicides. The cause of injury mortality with the fewest deaths was starvation, which accounted for 21 (less than 1%) of all injury deaths.

Table 4.16: Causes of injury mortality for MP and GP

Causes	*GP	Percentage	*MP	Percentage
Sequelae of external causes of mortality	149	0.19	22	0.03
Homicide	21850	27.79	3279	4.17
Other external causes of accidental injury	3875	4.93	1488	1.89
Complications of medical and surgical care	3228	4.11	545	0.69
Drowning	1607	2.04	681	0.87
Falls	241	0.31	52	0.07
Suffocation	1139	1.45	412	0.52
Forces of nature	1037	1.32	287	0.37
Fire/Burns	5045	6.42	991	1.26
Transport accidents	12741	16.20	4247	5.40
Poisoning	4811	6.12	1497	1.90
Suicide	7619	9.69	1758	2.24
Starvation	21	0.03	5	0.01

* Number of deaths.

Table 4.16 reveals that homicide claimed the lives of 25129 (31.96%) people. The majority of deaths happened as a result of injuries. This was followed by 16988 (21.6%) deaths caused by transport accidents and 9377 (11.98%) deaths caused by suicide. Starvation, 26 (less than 1% of injury mortality) is the cause of injury mortality with the lowest number. GP has the highest proportion of homicide deaths at 27.79%, while MP has 4.17%. Similarly, GP has the highest proportion of deaths related to transport accidents (16.20%). This is consistent with Pillay-van Wyk et al. (2016), who found that homicide was the most common cause of injury mortality. Similarly, Meel (2017) and StatsSA (2021) found

that homicide was the most common cause of injury mortality.

Table 4.17: Summary statistics for MP

min	max	mean	median	Stdev	skewness	kurtosis
57	199	115.6	116	26.4	0.227530	-0.181914

The results in Table 4.17 reveal the range of deaths due to injury, starting from 57 as the minimum number of deaths that occurred in MP to 199 as the maximum number of deaths that occurred in MP due to injury. However, there is an average and deviation of approximately 116 and 26, respectively, in the number of deaths due to injury on a monthly basis. Furthermore, the monthly injury mortality mirrors a normal distribution with a skewness of 0.23 and a kurtosis value of -0.18, which suggests that the data will follow heavy-tailed distributions.

Table 4.18: Summary statistics for GP

min	max	mean	median	Stdev	skewness	kurtosis
311	671	480	481.5	69.819229	0.135806	-0.492410

The results in Table 4.18 show the range of mortality due to injury, from 311 as the lowest number of deaths in GP to 671 as the highest number of deaths in GP related to injury. On a monthly basis, there is an average and deviation of 480 and approximately 70 deaths due to injury, respectively. However, the monthly injury mortality follows a normal distribution with a skewness of 0.14 and a kurtosis of -0.49, indicating that the distribution is flat and has thin tails when the excess value of kurtosis is negative (less than 3).

4.4 Time series analysis in MP

In this section, we analyse the time series data for MP to identify the plausible model and underlying pattern and make predictions.

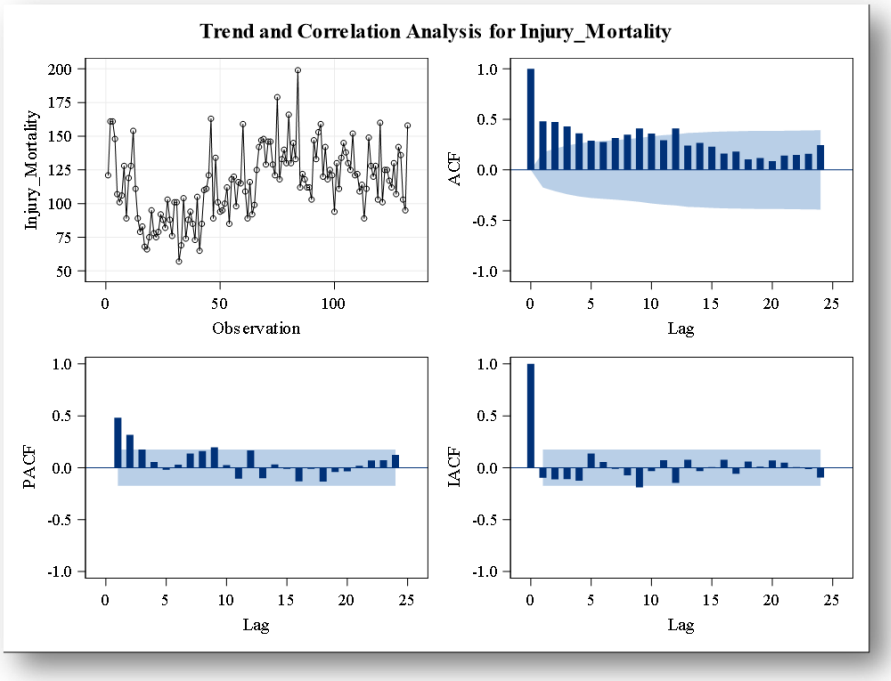

Figure 4.1: Trend and correlation plots.

The results in Figure 4.1 reveal a time series plot that indicates the monthly injury mortality cases, which have many trends and do not seem to be stationary in either mean or variance, showing upward and downward movement, which shows seasonality. The autocorrelation function (ACF) plot shows that the data is not stationary, and the plot also indicates that there is a seasonal pattern. The plot shows a non-seasonal significant peak at lag 1 until 10, and a seasonal significant peak at lag 12. Furthermore, the partial autocorrelation function (PACF) shows that the data is not stationary. The plot indicates that there is a seasonal pattern. The plot also shows a non-seasonal significant peak at lags 1,

2, 3, and 9 and a seasonal significant peak at lag 12.

Table 4.19: Tests for stationarity

Name	t-Stat	$p$value
ADF	-3.469	0.048
PP	-86.458	$<$0.010
KPSS	1.0007	$<$0.010

The Augmented Dickey-Fuller (ADF) and Phillips-Perron (PP) test statistics with p-values of 0.048 and less than 0.01, respectively, are shown in Table 4.19; the null hypothesis of non-stationarity of injury mortality is consequently rejected at the 5% level of significance. At the 5% level of significance, we reject the null hypothesis of stationarity using the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test statistic with a p-value less than 0.01. Finally, the ADF and PP results confirm that the injury mortality cases are stationary, however the KPSS tests confirm that the injury mortality cases are non-stationary, indicating that this data requires differencing.

Table 4.20: Test for normality

Name	t-Stat	p-value
JB	1.271	0.530
SW	0.992	0.616

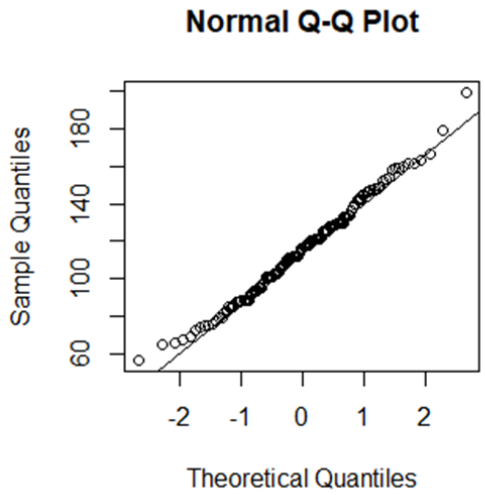

Figure 4.2: Normal Q-Q plot.

According to the findings in Table 4.20, there is evidence of platykurtic behaviour in injury mortality. Jarque Bera (JB) and Shapiro Wilk (SW) have p-values of 0.530 and 0.616, respectively. Using JB and SW p-values from Table 4.20, we reject the null hypothesis of a Gaussian distribution at the 5% significance level. This reveals that the injury mortality instances are normally distributed. Additionally, the Q-Q plot in Figure 4.2 demonstrates that injury mortality cases depart from the normal distribution at both tails. Table 4.20 shows that the statistics from the JB and SW tests do not support the claim of non-normality.

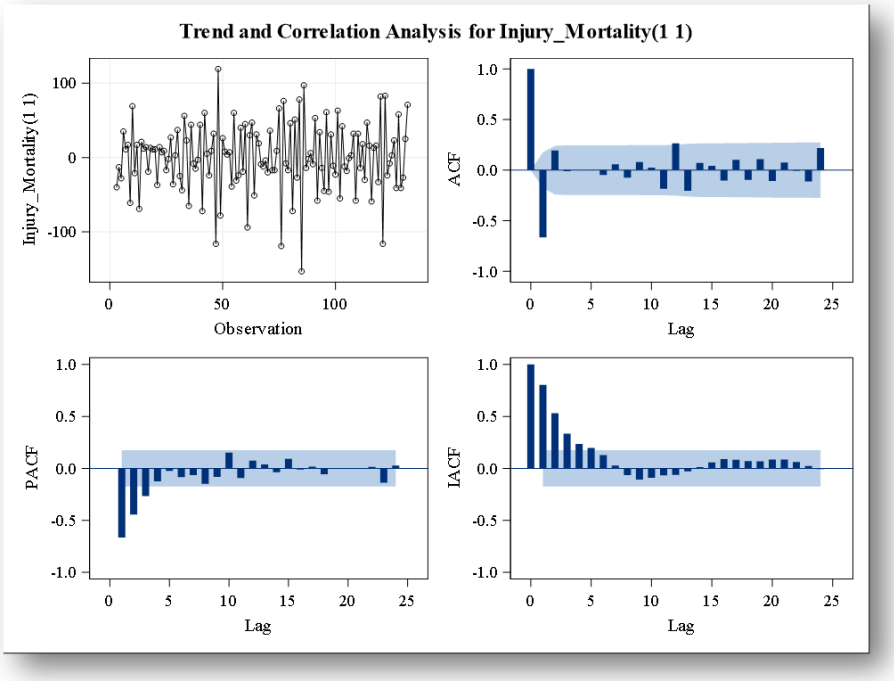

Figure 4.3: Trend and correlation for first differencing.

Figure 4.3 reveals that the time series plot is stationary after first differencing and that seasonality exists. The ACF plot displays a seasonally significant peak at lag 1 and a non-seasonally significant peak at lag 12. The PACF plot of first difference monthly injury mortality data reveals a non-seasonal significant peak at lags 1, 2 and 3.

Table 4.21: Tests for stationarity for first differencing

Name	t-Stat	*p*value
ADF	-6.8699	<0.010
PP	-171.57	<0.010
KPSS	0.043816	>0.100

Table 4.21 shows that the *p*-values for the ADF and PP test statistics are less than 0.01; thus, the null hypothesis of non-stationarity of injury mortality is re-

jected at the 5% level of significance. The results show that we fail to reject the null hypothesis of stationarity at the 5% level of significance for the KPSS test statistic with a p-value greater than 0.10. Finally, the ADF, PP, and KPSS tests show that the injury mortality cases are stationary after the first differencing.

Table 4.22: Test for normality first differencing

Name	t-Stat	p-value
JB	3.1379	0.2083
SW	0.99026	0.4919

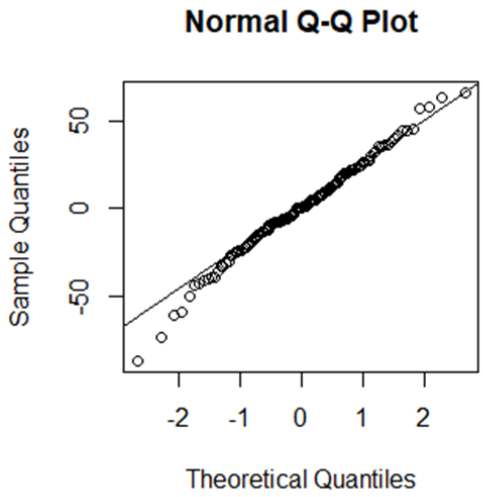

Figure 4.4: Normal Q-Q plot for first differencing.

Table 4.22 shows that there is evidence of platykurtic behaviour in injury mortality. We reject the null hypothesis of a Gaussian distribution at the 5% level of significance using the test statistics from Table 4.22, Jarque Bera (JB), and Shapiro Wilk (SW). The Q-Q plot shows that injury mortality instances are normally distributed. This indicates that injury mortality cases stray from the typical distribution on both ends. The JB and SW test statistics in Table 4.22 support the conclusion of normality.

4.5 Poisson and NB-INGARCH models in MP

We present the research analysis on both the Poisson and the Negative Binomial (NB) INGARCH models in this section. The analysis for MP is presented first. The ACF plot in Figure 4.3 includes a few spikes after the first lag, and two of them exceed the confidence interval, indicating a significant association between the value itself and the values corresponding to its first lag. There is a significant drop-off following the first lag. There are a couple of spikes after the first lag on the PACF plot, and three spikes that exceed the confidence interval, indicating a significant correlation between the value itself and the values corresponding to its first, second, and third lags. There is a notable decrease after the third lag.

4.5.1 Model identification

In this section, we estimate the performance of various competing models with the hope of choosing the best one. The one with the lowest BIC and AIC should be our choice. We will follow the most commonly used of these approaches, which is to start with the most complex model, and see which model can be dropped. The following models are constructed using the ACF and PACF mentioned in Figure 4.3: INGARCH models that best fit the data based on observing the ACF and PACF of the first differencing. The ACF and PACF were examined, and the following three models were summarised.

Table 4.23: Poisson and NB-INGARCH models

Model	AR	MA	Link	distribution	AIC	BIC
A	1	1	Ident	Poisson	1384.829	1393.478
A	1	1	Ident	Neg.Bin	1184.195	1195.727
A	1	1	log	Poisson	1381.128	1389.776
A	1	1	log	Neg.Bin	1183.355	1194.887
B	1	2	Ident	Poisson	1638.99	1647.638
B	1	2	Ident	Neg.Bin	1240.003	1251.535
B	1	2	log	Poisson	1398.953	1407.601
B	1	2	log	Neg.Bin	1187.681	1199.212
C	1	3	Ident	Poisson	1425.192	1433.84
C	1	3	Ident	Neg.Bin	1194.403	1205.934
C	1	3	log	Poisson	1417.38	1426.028
C	1	3	log	Neg.Bin	1192.66	1204.191

The results in Table 4.23 reveal that the logarithmic link function models present better results across the table for the criterion. The logarithmic function is the best link function since it minimises the loss of information and improves the results in the criterion compared to the identity link function. Nevertheless, the majority of the model variation shows a Negative Binomial distribution, which presents better scores in the Akaike information criterion (AIC) and worse results in the Bayesian information criterion (BIC). Model A performs slightly better across both criteria and is thus the preferred model.

Table 4.24: Parameter estimation on Poisson INGARCH

Parameters	Estimate	SE	CI(lower)	CI(upper)
(Intercept)	0.580	0.1519	0.282	0.877
beta(1)	0.2740	0.0686	0.140	0.408
alpha(1)	0.542	0.0598	0.425	0.659

$AIC = 1381.128, BIC = 1387.776$

Table 4.25: Parameter estimation on NB-INGARCH

Parameters	Estimate	SE	CI(lower)	CI(upper)
(Intercept)	5.9144	4.3177	-2.548	14.377
beta(1)	0.2740	0.0686	0.140	0.408
alpha(1)	0.6775	0.0864	0.508	0.847
sigmasq	0.0258			

$AIC = 1184.195, BIC = 1195.727$

Tables 4.24 and 4.25 reveal that the AIC is 1381.128 and the BIC is 1387.776 for the Poisson INGARCH, while the NB-INGARCH has an AIC of 1184.195 and BIC of 1195.727, as previously reported. Nonetheless, the majority of the model parameters have a negative binomial distribution, which results in better AIC scores but worse BIC values. The Negative Binomial distribution outperforms both criteria and is consequently the better model.

4.5.2 Model diagnostics

Model diagnostics involve checking how well the model fits. The residual analysis will be used.

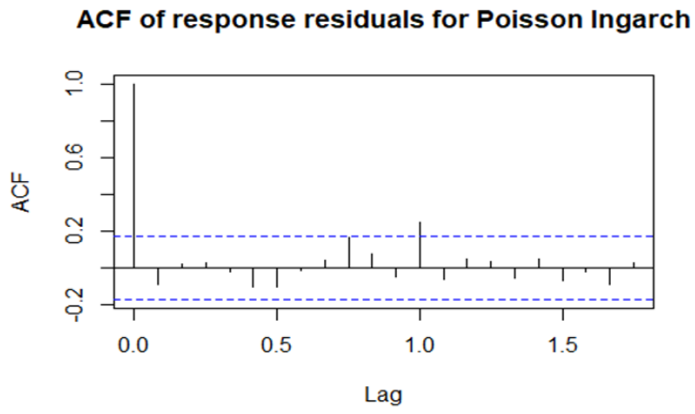

Figure 4.5: ACF for Poisson INGARCH residual

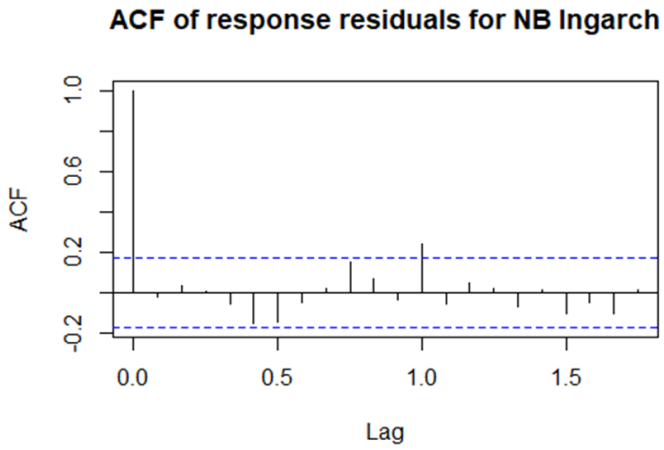

Figure 4.6: ACF for NB-INGARCH residual.

Figures 4.5 and 4.6 show ACF plots with adequate findings to indicate residual independence, despite the presence of a spike in the ACF that exceeds the confidence interval. Nonetheless, given that this isn't a theoretically compromised data set, the results are encouraging.

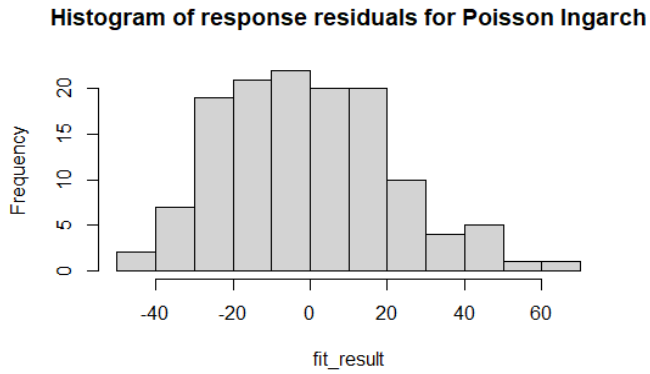

Figure 4.7: Histogram for Poisson INGARCH residual.

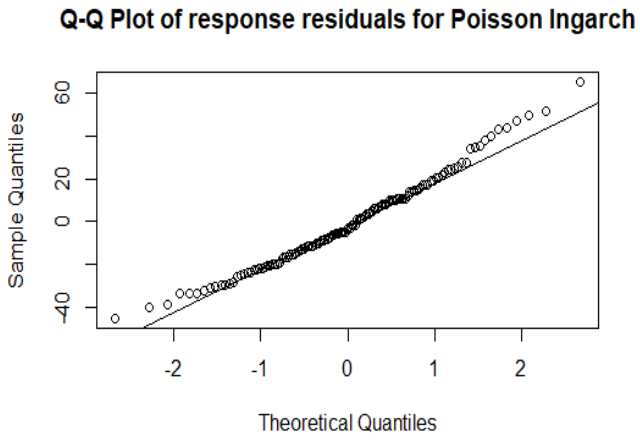

Figure 4.8: Q-Q plot for Poisson INGARCH residual.

Figure 4.7 reveals histogram to be normally distributed, whereas the Q-Q plot in Figure 4.8 appears to be normally distributed since the points sit on the line, despite the fact that the points at the bottom of the line are distant from the line. This does not imply that we should reject the normality of error terms in this model.

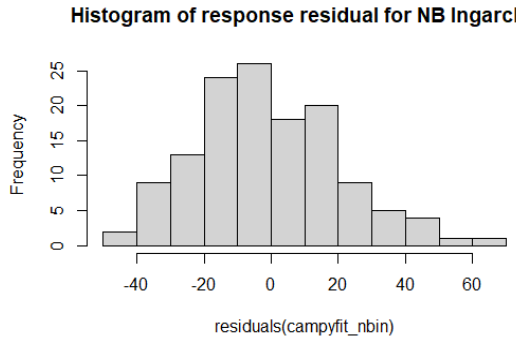

Figure 4.9: Histogram for NB-INGARCH residual

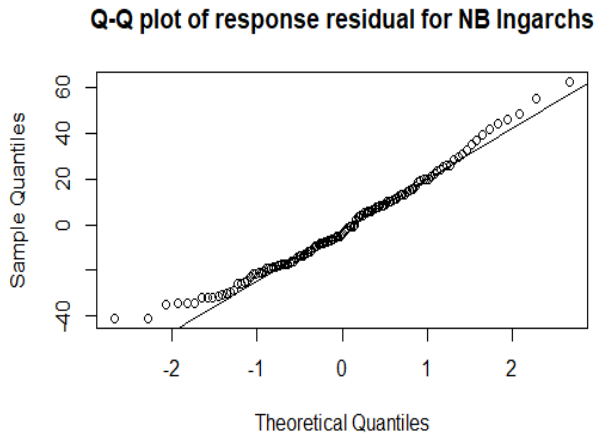

Figure 4.10: Q-Q plot for NB-INGARCH residual

Figure 4.9 shows that the histogram is normally distributed, whereas Figure 4.10 shows that the Q-Q plot is normally distributed since the points lie on the line, despite the fact that the points at the bottom of the line are away from the line. This does not imply that we should reject the normality of error terms in this model.

Table 4.26: The Langrange multiplier (LM) test on model A

Model	t-stats	DF	p-value
Poisson INGARCH (1,1)	0.67404	1	0.4116
NB-INGARCH (1,1)	1.26	1	0.2616

Table 4.26 reveals that there are no ARCH effects among the lags. The p-values of Poisson INGARCH, 0.4116, and NB-INGARCH, 0.2616, are greater than the significance level of 0.05, while the test statistics of Poisson INGARCH and NB-INGARCH are 0.674 and 1.26, respectively. As a result, the best model is the one with the highest test statistic value. Therefore, it indicates that the NB-INGARCH is the better model.

4.5.3 Model assessment

The model is assessed through evaluating the prediction error on new data. The accuracy measure known as root mean square error (RMSE) will be used to compare these time series models.

Table 4.27: Model assessment for model A

Link	Normalised squared error score(NRMSE)	Squared error score (MSE)
Log-NB	0.977	455.254
Log-Poisson	3.877	455.254

Table 4.27 presents the link function of both Poisson INGARCH and NB-INGARCH. The normalised root mean squared error (NRMSE) of Poisson INGARCH is 3.877, while NB-INGARCH is 0.977. Additionally, the mean squared error (MSE) for Poisson INGARCH is 455.24, while NB-INGARCH is 455.24, both of which are equal. As a result, the better model employing the relationship is the one with a lower MSE and NRMSE value, which gives us the better model. The results suggest that a Negative Binomial distribution yields better outcomes. As a result, NB-INGARCH is the better model.

4.6 ARIMA in MP

We present the research analysis of autoregressive integrated moving average (ARIMA) models in this section. The analysis for MP is presented first. The ACF plot in Figure 4.3 shows a few spikes after the first lag, and two of them exceed the confidence interval, indicating a significant relationship between the value itself and the values corresponding to its first lag. There is a significant drop-off following the first lag. There are a couple of spikes after the first lag on the PACF plot, and three spikes that exceed the confidence interval, indicating a significant correlation between the value itself and the values corresponding to its first, second, and third lags. There is a significant drop after the third lag.

4.6.1 Model identification

In this section, we estimate the performance of various competing models with the hope of choosing the best one. The one with the lowest BIC and AIC should be our choice. We will follow the most commonly used of these approaches, which is to start with the most complex model, and see which model can be dropped. The specific goal here is to obtain some idea of the values of p, d, and q needed in the general linear ARIMA model and to obtain initial estimates for the parameters.

Table 4.28: Model summary

Model	ARIMA(p,d,q)(P,D,Q)
A	ARIMA $(1,1,1) \times (1,1,1)_{12}$
B	ARIMA $(1,1,2) \times (1,1,1)_{12}$
C	ARIMA$(1,1,3) \times (1,1,1)_{12}$

The ARIMA models that best fit the data based on observing the ACF and PACF of the first differencing data are constructed using the ACF and PACF mentioned in Figure 4.3: The ACF and PACF were examined, and the three ARIMA models presented in Table 4.28 were summarised.

Table 4.29: Fit statistics for model A

Fit statistics	Mean
Stationary R-squared	0.538
R-squared	0.346
RMSE	21.871
MAPE	16.545
MAE	17.601
Normalised BIC	6.371

The results in Table 4.29 show that the value of R-squared is 0.345, which accounts for approximately 35% of the variation in death due to injury in MP. The mean absolute error (MAE) value is 17.601, showing a 17.601 average error between forecasts and actuals, while the root mean squared error (RMSE)

is 21.871, suggesting a 21.871 weighted average error between forecasts and actuals. The mean absolute percentage error (MAPE) value is 16.545, which means that on average, the forecast is off by 16.5%.

Table 4.30: Parameter estimation for model A

Parameters	Estimate	SE	t-stats	p-value
Constant	0.126	0.190	0.663	0.509
AR(1)	-0.031	0.136	-0.231	0.817
MA(1)	0.675	0.104	6.473	0.000
SAR(1)	-0.026	0.134	-0.195	0.846
SMA(1)	0.855	0.174	4.922	0.000

Table 4.30 shows a parameter estimate for the ARIMA $(1,1,1)\times(1,1,1)_{12}$ model. It also displays the estimate's p-value. The model contains four parameters: AR(1), MA(1), seasonal autoregressive (SAR)(1), and seasonal moving average (SMA)(1). The constant estimate, with a p-value of 0.509, is a function of the mean term and is not statistically significant at the 5% level. The MA(1), with a reordered p-value of less than 0.05, and the SMA(1), with a recorded p-value of less than 0.05, are both significant because their p-values are less than alpha 5%. The AR(1) had a p-value of 0.817 and the SAR(1) had a p-value of 0.846, both of which are not significant at the 5% level of significance.

Table 4.31: Fit statistics for model B

Fit statistics	Mean
Stationary R-squared	0.540
R-squared	0.349
RMSE	21.924
MAPE	16.477
MAE	17.549
Normalised BIC	6.416

Table 4.31 shows that the value of R-squared is 0.349, which accounts for approximately 35% of the variation in MP injury mortality. The mean absolute error (MAE) value is 17.549, showing a 17.549 average error between forecasts

and actuals, while the root mean squared error (RMSE) is 21.924, suggesting a 21.924 weighted average error between forecasts and actuals. The mean absolute percentage error (MAPE) value is 16.477, which means that on average, the forecast is off by 16.5%.

Table 4.32: Parameter estimation for model B

Parameters	Estimate	SE	t-stats	p-value
Constant	0.126	0.190	0.665	0.507
AR(1)	-0.964	0.113	-8.509	0.000
MA(1)	-0.304	30.516	-0.010	0.992
MA(2)	0.696	21.238	0.033	0.974
SAR(1)	-0.024	0.137	-0.175	0.862
SMA(1)	0.841	0.172	4.900	0.000

The parameter estimates for the ARIMA $(1,1,2) \times (1,1,1)_{12}$ models are shown in Table 4.32. The table also presents the p-value for the estimates. The model consists of five parameters: AR(1), MA(1), MA(2), SAR(1), and SMA(1). The constant estimate, with a p-value of 0.507, is a function of the mean term and is not statistically significant at the 5% level. The AR(1) had a p-value of less than 0.05, and the SMA(1) had a p-value of less than 0.05, both of which are significant because their p-values are less than alpha 5%. The p-values for the MA(1), MA(2), and SAR(1) were 0.992, 0.974, and 0.862, respectively, which are not significant at the 5% level of significance.

Table 4.33: Fit statistics for model C

Fit statistics	Mean
Stationary R-squared	0.539
R-squared	0.347
RMSE	22.050
MAPE	16.501
MAE	17.549
Normalised BIC	6.468

Table 4.33 shows the value of R-squared is 0.347, which accounts for approximately 35% of the variation in MP injury mortality. The mean absolute er-

ror (MAE) value is 17.549, showing a 17.549 average error between forecasts and actuals, while the root mean squared error (RMSE) is 22.050, suggesting a 22.050 weighted average error between forecasts and actuals. The mean absolute percentage error (MAPE) value is 16.501, which means that on average, the forecast is off by 16.5%.

Table 4.34: Parameter estimation for ARIMA model C

Parameters	Estimate	SE	t-stats	p-value
Constant	0.167	0.186	0.898	0.371
AR(1)	-0.967	0.142	-6.815	0.000
MA(1)	-0.286	69.205	-0.004	0.997
MA(2)	0.698	49.400	0.014	0.989
MA(3)	-0.016	1.113	-0.015	0.988
SAR(1)	0.018	0.140	0.128	0.898
SMA(1)	0.882	0.215	4.107	0.000

The results in Table 4.34 show parameter estimates for the ARIMA $(1,1,3) \times (1,1,1)_{12}$ model. It also presents the p-value for the estimates. The model consists of six parameters: AR(1), MA(1), MA(2), MA(3), SAR(1), and SMA(1) (1). The constant estimate, with a p-value of 0.371, is a function of the mean term and is not statistically significant at the 5% level. The AR(1), with a p-value less than 0.05, and the SMA(1), with a p-value less than 0.05, are both significant because their p-values are less than alpha 5%. The p-values for the MA(1), MA(2), MA(3), and SAR(1) were 0.997, 0.989, 0.988, and 0.898, respectively, which are not significant at the 5% level of significance.

4.6.2 Model diagnostics

Model diagnostics involve checking how well the model fits. The residual analysis will be used.

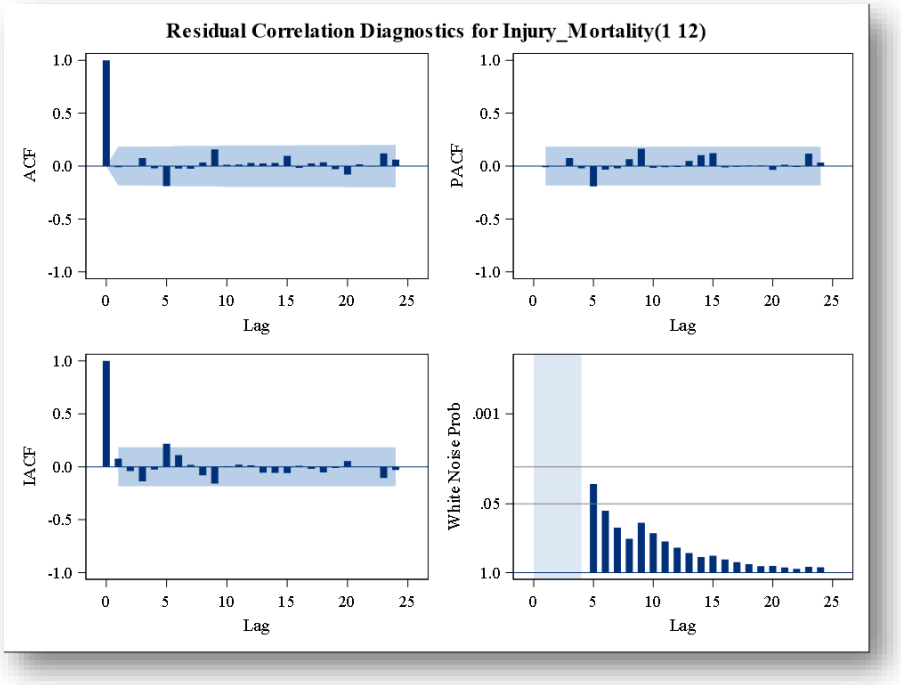

Figure 4.11: Residual ACF and PACF

Figure 4.11 shows residual ACF and PACF with a significant peak at lag 5, non-significant peaks, and all values fall within the upper and lower confidence ranges, indicating that the model is white noise, even though there is autocorrelation in the residuals.

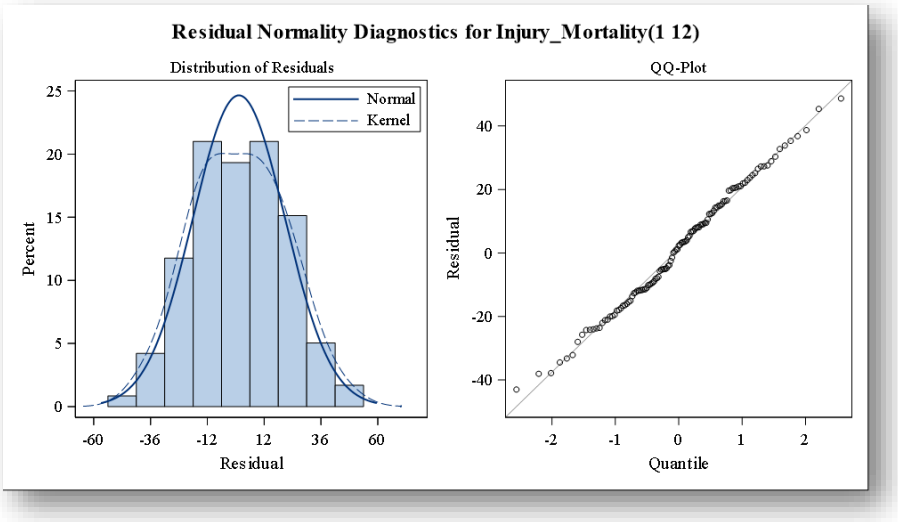

Figure 4.12: Residual Histogram and Q-Q plot

The results in Figure 4.12 reveal the histogram data to be normally distributed, while on the same Q-Q plot in Figure 4.12, the points appear to be normally distributed because they lie on the line, even though at the bottom of the line, points are away from the line.

Table 4.35: Ljung Box test for model A

Model	t-stats	Df	p-value
ARIMA $(1,1,1) \times (1,1,1)_{12}$	11.632	14	0.636

The Ljung Box test $Q* = 11.632$ with a probability of 0.636 is shown in Table 4.35, showing that the estimated model is uncorrelated and may be adequate. Since the p-value of 0.636 is greater than the 5% threshold of significance, the model is not significant at the 5% level of significance.

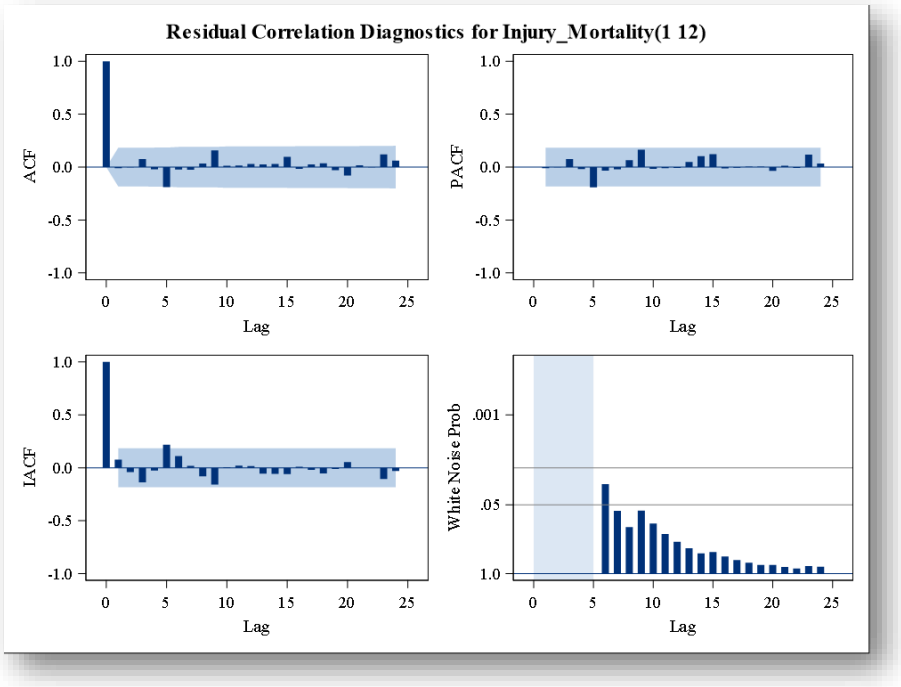

Figure 4.13: Residual ACF and PACF

Figure 4.13 shows residual ACF and PACF that have a significant peak at lag 5 and non-significant peaks, and all their values lie within the upper and lower confidence intervals, indicating that the model is white noise, even though there is autocorrelation in the residuals.

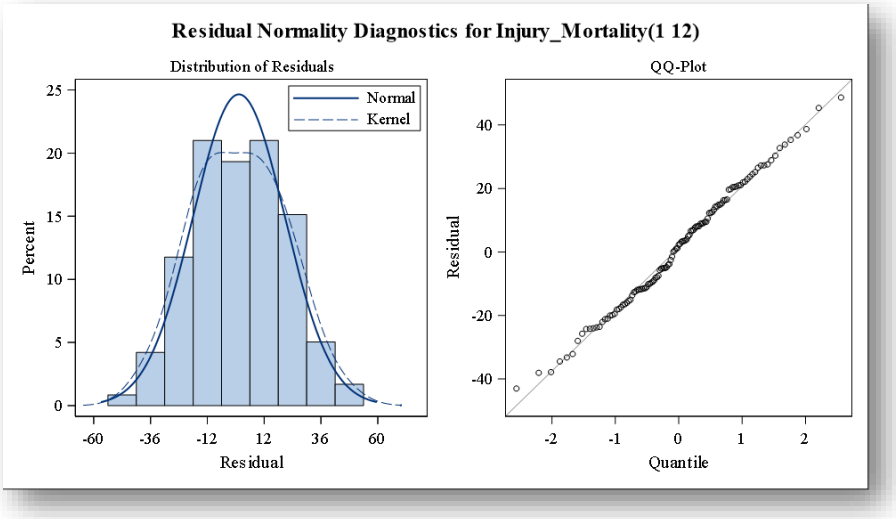

Figure 4.14: Residual Histogram and Q-Q plot

The histogram data in Figure 4.14 shows to be normally distributed, and the Q-Q plot in the same figure also appears to be normally distributed because the points lie on the line even though the points at the bottom of the line are away from the line.

Table 4.36: Ljung Box test for model B

Model	t-stats	Df	p-value
ARIMA $(1,1,2) \times (1,1,1)_{12}$	11.113	13	0.601

Table 4.36 reveals that the Ljung Box test $Q* = 11.113$ with a probability of 0.601 indicates that the estimated model is uncorrelated and may be adequate. The model is not significant at the 5% level of significance because the p-value of 0.601 is greater than the 5% level of significance.

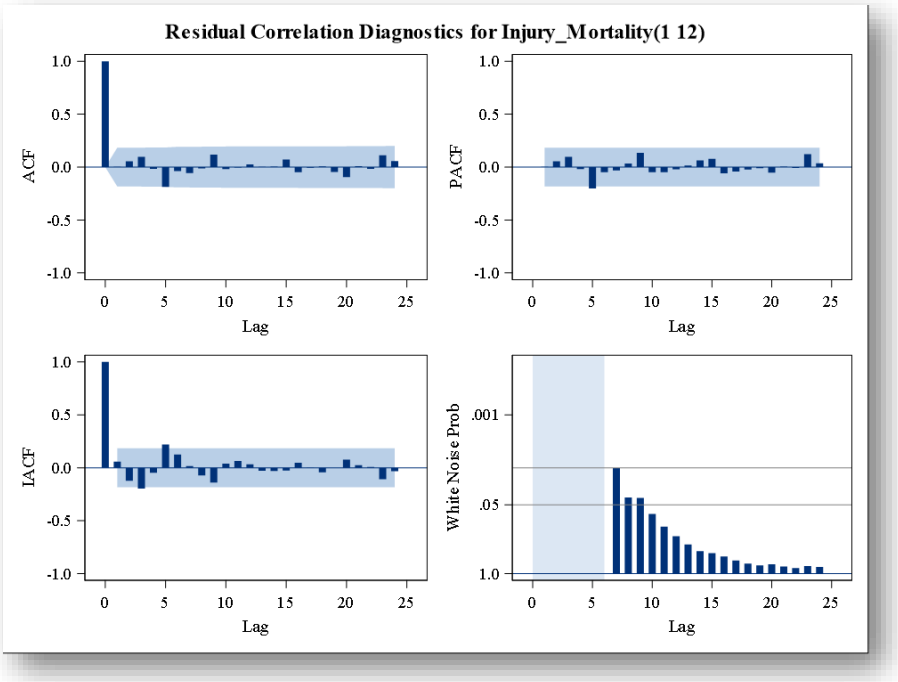

Figure 4.15: Residual ACF and PACF

The results in Figure 4.15 show residual ACF and PACF that have a significant peak at lag 5 and non-significant peaks, and all their values lie within the upper and lower confidence intervals, indicating that the model is white noise, even though there is autocorrelation in the residuals.

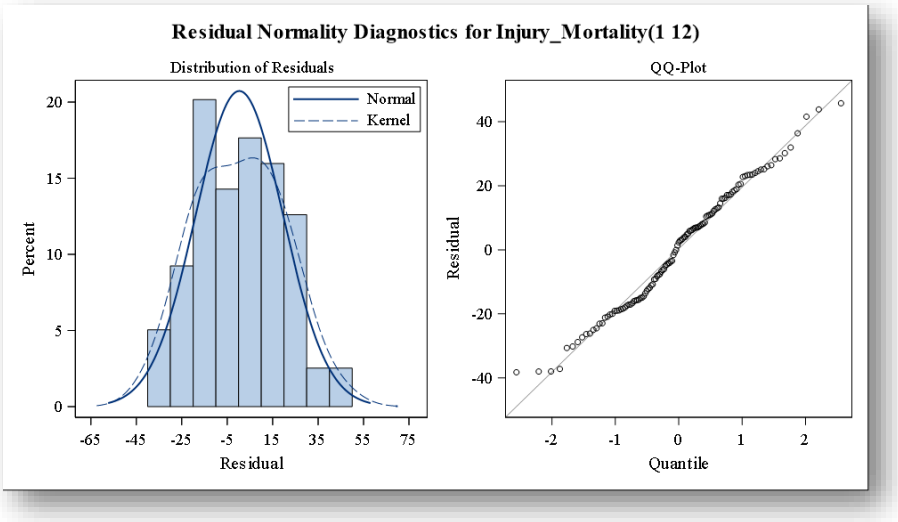

Figure 4.16: Residual Histogram and Q-Q plot

The histogram data in Figure 4.16 appears to be normally distributed, and the Q-Q plot in the same figure also appears to be normally distributed because the points lie on the line even though the points at the bottom of the line are away from the line.

Table 4.37: Ljung Box test for model C

Model	t-stats	Df	p-value
ARIMA $(1,1,3) \times (1,1,1)_{12}$	11.350	12	0.499

The results in Table 4.37 reveal the Ljung Box test $Q* = 11.350$ with probability of 0.499, indicating that the estimated model is uncorrelated, and it may be adequate. Since the p-value of 0.499 is greater than the 5% threshold of significance, the model is not significant at the 5% level of significance.

4.6.3 Model assessment

The model is assessed through evaluating the prediction error on new data. The accuracy measure known as root mean square error (RMSE) will be used to compare these time series models.

Table 4.38: Model assessment

Model	BIC	RMSE
ARIMA $(1,1,1) \times (1,1,1)_{12}$	6.371	21.871
ARIMA $(1,1,2) \times (1,1,1)_{12}$	6.416	21.924
ARIMA $(1,1,3) \times (1,1,1)_{12}$	6.468	22.050

The results in Table 4.38 reveal a model assessment of MP monthly deaths due to injury. The best model selection is based on normalised Bayesian information criteria (BIC) and root mean square error (RMSE). The lowest of BIC and RMSE is used for model selection. ARIMA $(1,1,1) \times (1,1,1)_{12}$ is the best model and has a minimum value of 6.371 and 21.871 for BIC and RMSE, respectively.

4.7 Model assessment in MP

The model is assessed through evaluating the prediction error on new data. The accuracy measure known as root mean square error (RMSE) will be used to compare these time series models.

4.7.1 Comparing the NB-INGARCH and ARIMA models

Table 4.39: Model assessment

Model	RMSE
ARIMA $(1,1,1) \times (1,1,1)_{12}$	21.871
NB-INGARCH$(1,1)$	112.976

Table 4.39 displays the findings of two models based on MP monthly injury mortality data. The best model is chosen using the root mean square error (RMSE).

The best model is the one with the lowest RMSE value. The NB-INGARCH (1,1) model has an RMSE of 112.976, but the ARIMA $(1,1,1) \times (1,1,1)_{12}$ model has an RMSE of 21.871, making it the most suitable model for MP monthly injury mortality data.

4.8 Injury mortality model in MP

The equation of ARIMA $(1,1,1) \times (1,1,1)_{12}$ model is given by

$$\phi_1(B)(1-B)\Phi_1(B^{12})(1-B^{12})x_t = \theta_1(B)\Theta_1 B^{12} z_t \tag{4.1}$$

$$(1-\phi_1 B)(1-B)(1-\Phi_1 B^{12})(1-B^{12})x_t = (1-\theta_1 B)(1-\Theta_1 B^{12})z_t$$

$$(1+0.031B)(1-B)(1+0.026B^{12})(1-B^{12})x_t = (1-0.675B)(1-0.855B^{12})z_t$$

$$\therefore x_t = 0.969x_{t-1} + 0.031x_{t-2} + 0.974x_{t-12} - 0.944x_{t-13} - 0.030x_{t-14} + 0.026x_{t-24} - 0.0252x_{t-25} - 0.0008x_{t-26} + z_t - 0.675z_{t-1} - 0.855z_{t-12} + 0.5771z_{t-13}$$

4.9 Forecasting MP injury mortality

Once we have the final ARIMA model, we are now ready to make predictions on the future time points. We also visualise the trends.

4.9.1 MP monthly injury mortality forecast

Table 4.40: Forecast with 95% confidence limits

Date	Time	Forecast	Lower Limit	Upper Limit
Jan-19	133	108.9254	70.1173	147.7336
Feb-19	134	112.1336	71.6923	152.5749
Mar-19	135	117.697	75.4429	159.9511
Apr-19	136	107.5042	63.5196	151.4887
May-19	137	104.4013	58.7516	150.0511
Jun-19	138	110.4783	63.2221	157.7346
Jul-19	139	126.3687	77.5588	175.1786
Aug-19	140	132.1524	81.8367	182.468
Sep-19	141	125.3733	73.5957	177.1509
Oct-19	142	130.4094	77.2100	183.6088
Nov-19	143	115.7680	61.1838	170.3522
Dec-19	144	153.8829	97.9483	209.8176
Jan-20	145	114.8414	56.4284	173.2544
Feb-20	146	118.3944	58.4257	178.3631
Mar-20	147	123.9526	62.4448	185.4605
Apr-20	148	113.9446	50.9360	176.9533
May-20	149	110.9213	46.4466	175.3959
Jun-20	150	117.4350	51.5270	183.343
Jul-20	151	132.4485	65.1376	199.7593
Aug-20	152	139.1159	70.4309	207.801
Sep-20	153	132.4855	62.4532	202.5178
Oct-20	154	136.6668	65.3127	208.0209
Nov-20	155	122.3256	49.6738	194.9774
Dec-20	156	161.2123	87.2855	235.1391

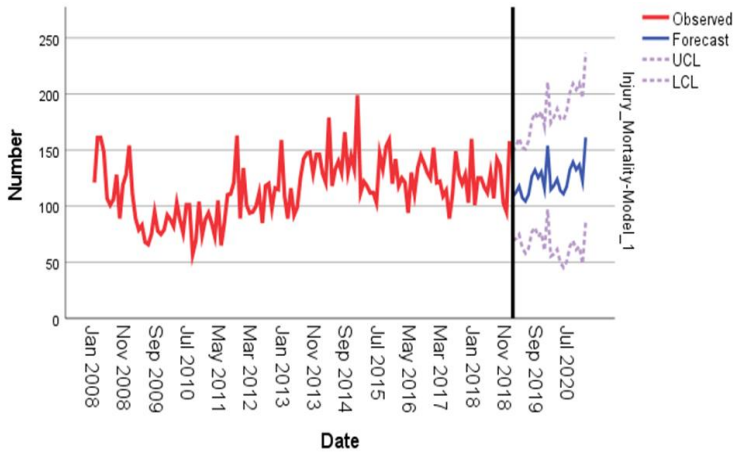

Figure 4.17: Time series plot for forecast

The forecasted results for injury mortality from January 2019 to December 2020 are shown in Table 4.40 and Figure 4.17, along with monthly injury mortality estimates with 95% confidence intervals over two years. On a monthly basis, the MP forecasts show a decreasing trend from year to year. Furthermore, the ARIMA$(1,1,1)\times(1,1,1)_{12}$ model suggests a decrease in injury mortality in 2020 compared to 2019.

4.10 Time series analysis in GP

In this section, we analyse the time series data for GP to identify the plausible model and underlying pattern and make predictions.

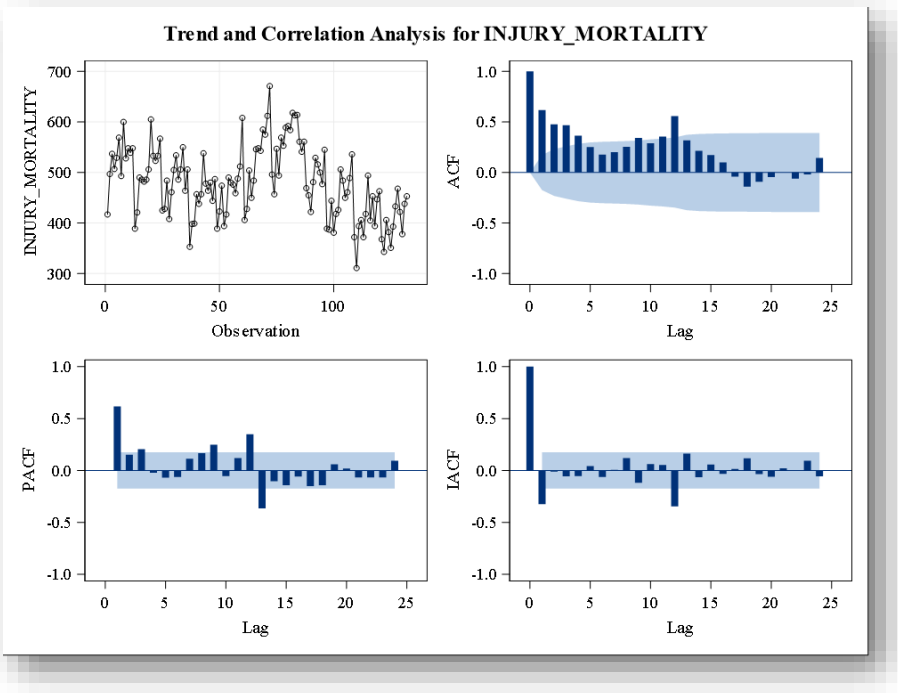

Figure 4.18: Trend and correlation plots.

The results in Figure 4.18 reveal a time series plot that indicates the monthly injury mortality cases, which have many trends and do not seem to be stationary. There is an upward and downward movement, which shows seasonality. The autocorrelation function (ACF) plot shows that the data is not stationary, and the plot also indicates that there is a seasonal pattern. The plot shows a non-seasonal significant peak at lag 1 until lags 4, 9, and 11, and a seasonal significant peak at lag 12. Furthermore, the partial autocorrelation function (PACF) shows that the data is not stationary. The plot indicates that there is a seasonal pattern. The plot also shows a non-seasonal significant peak at lags 1, 3, 8, 9, and 13, and a seasonal significant peak at lag 12.

Table 4.41: Tests for stationarity

Name	t-Stat	p-value
ADF	-3.477	0.047
PP	-59.727	$<$0.010
KPSS	0.55993	0.028

The ADF and PP tests statistics with p-values of 0.047 and less than 0.01, respectively, are shown in Table 4.41; the null hypothesis of non-stationarity of injury mortality is then rejected at the 5% level of significance. With a p-value of 0.028 for the KPSS test statistic, the results show that we reject the null hypothesis of stationarity at a 5% level of significance. Finally, the ADF and PP results show that the injury mortality cases are stationary; however, the KPSS test showed that the injury mortality cases are non-stationary, indicating that this data requires differencing.

Table 4.42: Test for normality

Name	t-Stat	p-value
JB	1.5487	0.461
SW	0.99127	0.583

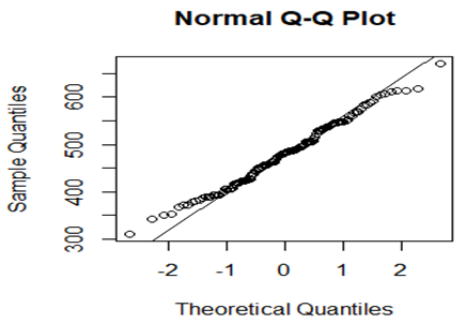

Figure 4.19: Normal Q-Q plot.

The findings in Table 4.42 show that there is evidence of platykurtic behaviour in injury mortality. The p-values were 0.461 and 0.583 for JB and SW, respec-

tively. Using the JB and SW p-values from Table 4.42, we clearly fail to reject the null hypothesis of Gaussian distribution at the 5% level of significance. This demonstrates that injury mortality cases are distributed normally. Furthermore, the Q-Q plot in Figure 4.19 demonstrates that injury mortality cases depart from the normal distribution at both ends. The JB and SW test statistics presented in Table 4.42 do not support the suggestion of non-normality.

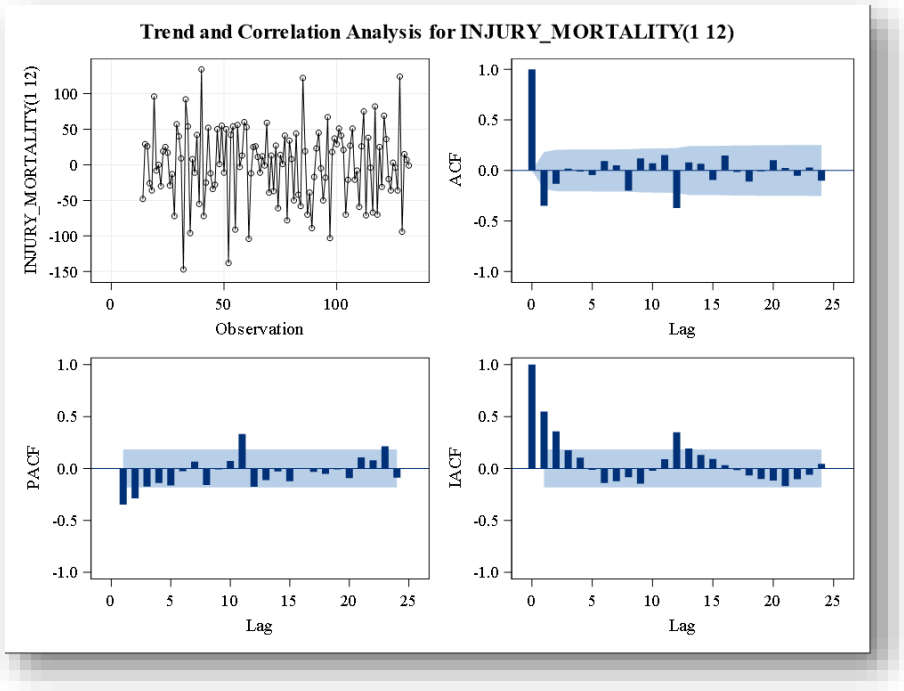

Figure 4.20: Trend and correlation for first differencing.

Figure 4.20 reveals that the time series plot is stationary after first-order differencing, and because the data is monthly, there is seasonality. The ACF plot shows a non-seasonal significant peak at lag 1, an annual significant peak at lag 8, and a peak at lag 12. The PACF plot of first-order differencing monthly injury mortality data shows a non-seasonal significant peak at lags 1, 2, 3, and 11 and a seasonal significant peak at lags 12 and 24.

Table 4.43: Tests for stationarity for first differencing

Name	t-Stat	p-value
ADF	-6.2075	$<$0.010
PP	-146.51	$<$0.010
KPSS	0.038289	$>$0.100

The results in Table 4.43 show that the p-values for the ADF and PP test statistics are less than 0.01; thus, the null hypothesis of non-stationarity of injury mortality is rejected as significant at the 5% level. The results show that we fail to reject the null hypothesis of stationarity at the 5% level of significance for KPSS test statistics with a p-value greater than 0.10. Finally, the ADF, PP, and KPSS tests show that the injury mortality cases are stationary after the first difference.

Table 4.44: Test for normality first differencing

Name	t-Stat	p-value
JB	25.929	$<$0.010
SW	0.93967	$<$0.010

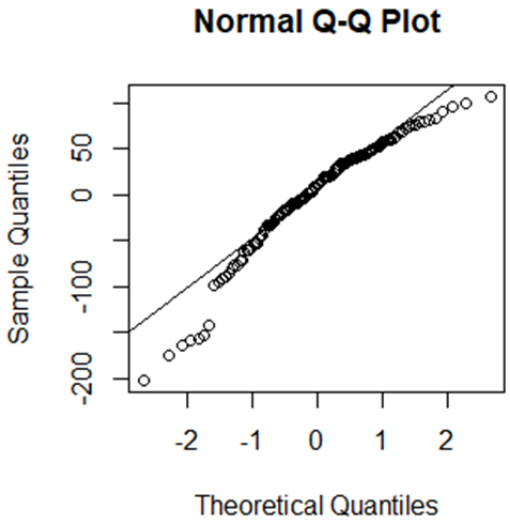

Figure 4.21: Normal Q-Q plot.

The results in Table 4.44 reveal that there is an indication of platykurtic behaviour in the injury mortality. The JB and SW p-values are less than 0.01. We reject the null hypothesis of a Gaussian distribution at the 5% level of significance using JB and SW p-values from Table 4.44. This confirms that the injury mortality cases are not normally distributed. Furthermore, the Q-Q plot in Figure 4.21 shows that injury mortality cases deviate from the normal distribution at both ends of the distribution. The indication of normality is not supported by the JB and SW test statistics reported in Table 4.44.

4.11 Poisson and NB-INGARCH models in GP

In this section, we present the research analysis on both Poisson and NB-INGARCH models. We present an analysis for GP. In Figure 4.20, the ACF plot shows a few peaks after the first lag, and there are three peaks that exceed the confidence interval, indicating a significant correlation between the value itself and the values corresponding to its first lag. After the first delay there is a noticeable drop. In the PACF plot, there are a few peaks after the first lag, and there are five peaks that exceed the confidence interval, indicating a significant correlation between the value itself and the values corresponding to its first, second, and third lags. After the third lag there is a significant drop.

4.11.1 Model identification

In this section, we estimate the performance of various competing models with the hope of choosing the best one. The one with the lowest BIC and AIC should be our choice. We will follow the most commonly used of these approaches, which is to start with the most complex model, and see which model can be dropped. The following models are constructed using the ACF and PACF mentioned in Figure 4.20: INGARCH models that best fit the data based on the first difference. The ACF and PACF were examined, and the following three models

were summarised.

Table 4.45: Poisson and NB-INGARCH models

Model	AR	MA	Link	distribution	AIC	BIC
A	1	1	Ident	Poisson	1923.939	1932.587
A	1	1	Ident	Neg.Bin	1444.431	1455.962
A	1	1	log	Poisson	1844.367	1853.015
A	1	1	log	Neg.Bin	1431.971	1443.502
B	1	2	Ident	Poisson	2355.884	2364.533
B	1	2	Ident	Neg.Bin	1498.021	1509.552
B	1	2	log	Poisson	1830.350	1838.999
B	1	2	log	Neg.Bin	1429.042	1440.573
C	2	1	Ident	Poisson	2377.047	2385.696
C	2	1	Ident	Neg.Bin	1500.125	1511.656
C	2	1	log	Poisson	2037.426	2046.074
C	2	1	log	Neg.Bin	1460.925	1472.456

The results in Table 4.45 reveal that the logarithmic link function models present better results across the table for the criterion. The logarithmic function is the best link function since it minimises the loss of information and improves the results in the criterion compared to the identity link function. Nevertheless, the majority of the model variation shows a Negative Binomial distribution that presents better scores in the AIC and worse results in the BIC of 1429.042 and 1440.573, respectively, compared to Poisson scores in the AIC and worse results in the BIC of 1830.350 and 1838.999, respectively. Model B performs slightly better across both criteria and is thus the preferred model.

Table 4.46: Parameter estimation on Poisson INGARCH

Parameters	Estimate	SE	CI(lower)	CI(upper)
(Intercept)	0.951	0.5680	0.1956	1.335
beta(1)	0.512	0.0311	0.4510	0.573
alpha(2)	0.334	0.0453	0.2450	0.422

$$AIC = 1830.350, BIC = 1838.999$$

Table 4.47: Parameter estimation on NB-INGARCH

Parameters	Estimate	SE	CI(lower)	CI(upper)
(Intercept)	0.95148	0.4672	0.0357	1.867
beta(1)	0.51235	0.0736	0.3681	0.657
alpha(2)	0.33362	0.1073	0.1232	0.544
sigmasq	0.00975			

$AIC = 1429.042, BIC = 1440.573$

The results in Tables 4.46 and 4.47 show an AIC of 1830.350 and a BIC of 1838.999 under Poisson INGARCH and an AIC of 1429.042 and a BIC of 1440.573 under NB-INGARCH. Nonetheless, the majority of the model parameters exhibit a Negative Binomial distribution, showing better scores on the AIC criterion and worse scores on the BIC. The negative binomial distribution performs slightly better on both criteria and is therefore the model of choice.

4.11.2 Model diagnostics

Model diagnostics involve checking how well the model fits. The residual analysis will be used.

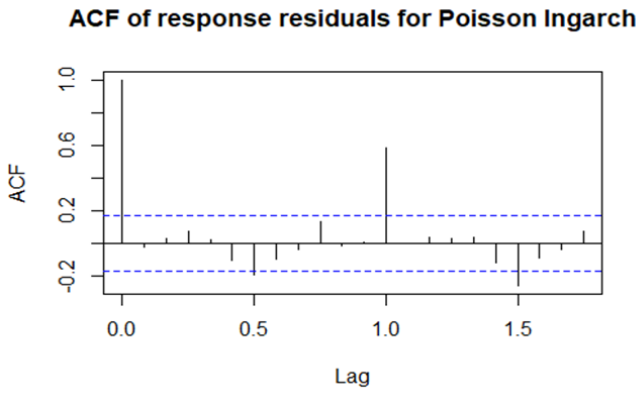

Figure 4.22: ACF for Poisson INGARCH residual.

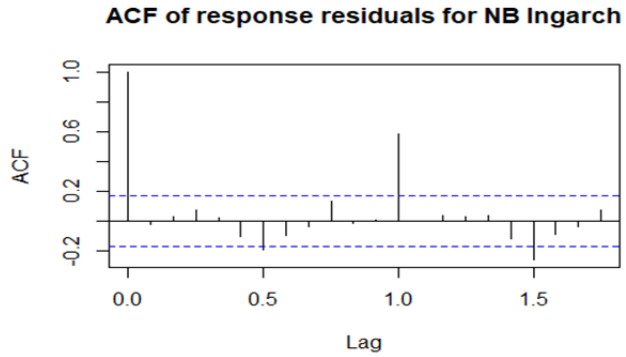

Figure 4.23: ACF for NB-INGARCH residual

Figures 4.22 and 4.23 show ACF plots with decent results to support residual independence despite the presence of a peak that exceeds the confidence interval in the ACF. However, given that this is not a theoretically compromised dataset, the results are encouraging.

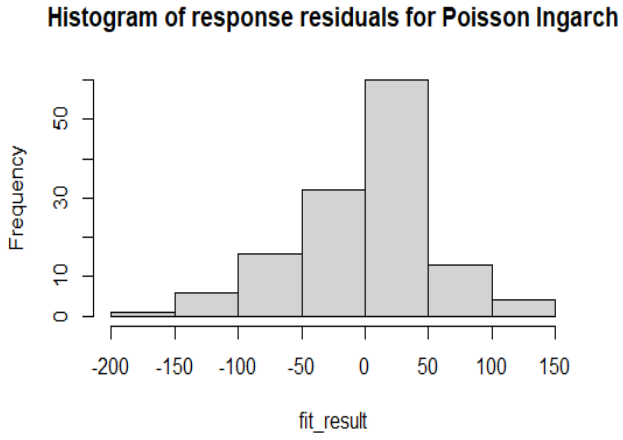

Figure 4.24: Histogram for Poisson INGARCH residual

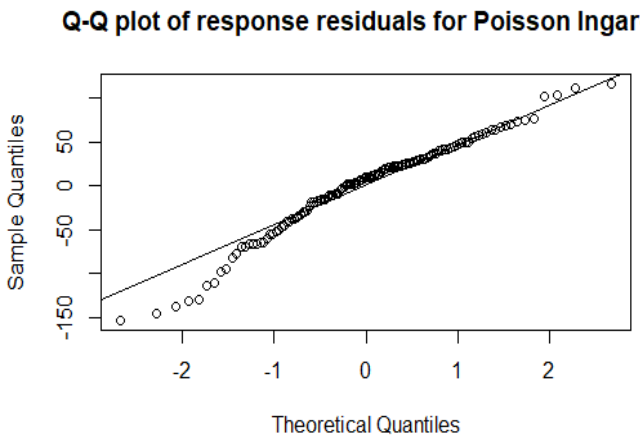

Figure 4.25: Q-Q plot for Poisson INGARCH residual

Figure 4.24 reveals that the histogram data is skewed to the left, while the Q-Q plot in Figure 4.25 appears to be normally distributed because the points are on the line, even though the points at the bottom of the line are off the line. This does not imply that we should reject the normality of error terms in this model.

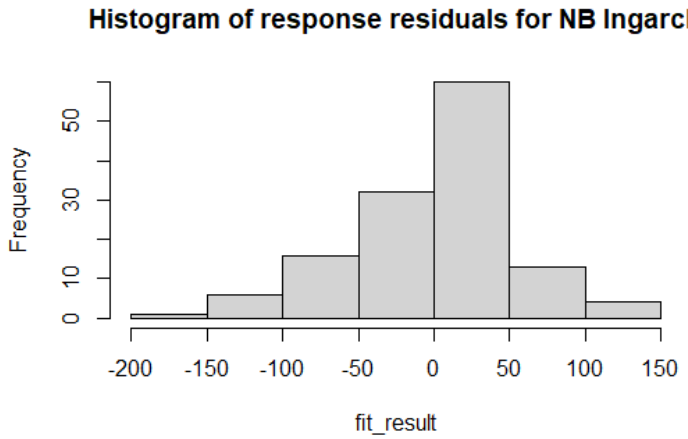

Figure 4.26: Histogram for NB-INGARCH residual.

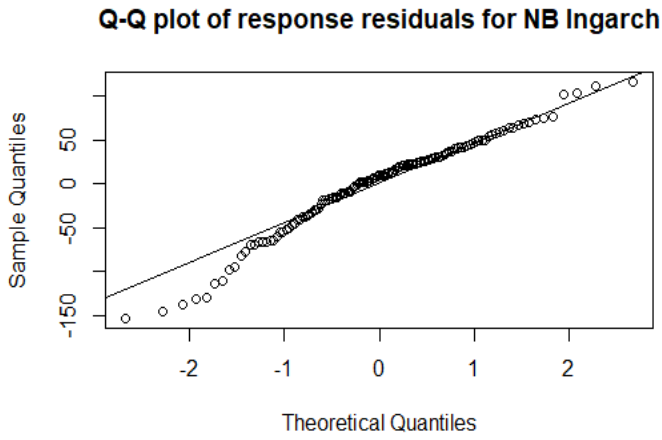

Figure 4.27: Q-Q plot for NB-INGARCH residual.

The histogram data in Figure 4.26 is skewed to the left, while the Q-Q plot in Figure 4.27 appears to be normally distributed because the points fall on the line, even though the points at the bottom of the line are off the line. This does not imply that we should reject the normality of error terms in this model.

Table 4.48: The Langrange multiplier (LM) test on model B

Model	t-stats	DF	p-value
Poisson INGARCH(1,2)	3.1688	2	0.2051
NB-INGARCH (1,2)	3.1688	2	0.2051

Table 4.48 shows that there are no ARCH effects among the lags. Since the Poisson p-values of INGARCH, which is 0.2051; and NB-INGARCH, which is 0.2051, are larger than the significance level of 0.05, we also observe the test statistics of Poisson INGARCH of 3.1688 and NB-INGARCH of 3.1688. As a result, the model with the highest test statistics value is the best. Consequently, both models appear to be better.

4.11.3 Model assessment

The model is assessed through evaluating the prediction error on new data. The accuracy measure known as root mean square error (RMSE) will be used to compare these time series models.

Table 4.49: Model assessment

Link	Normalised squared error score (NRMSE)	Squared error score (MSE)
Log-NB	0.978	2796.065
Log-Poisson	5.658	2796.065

Table 4.49 presents the link function of both Poisson INGARCH and NB-INGARCH. The normalised root mean squared error (NRMSE) of Poisson INGARCH is 5.658, while that of NB-INGARCH is 0.978. Moreover, the mean squared error (MSE) for Poisson INGARCH is 2796.065, while NB-INGARCH was 2796.065. As a result, the better model using the link is the one with a lower MSE and NRMSE values, which gives us the better model. The results suggest that a Negative Binomial distribution yields better results. As a result, NB-INGARCH is the better model.

4.12 ARIMA in GP

The research analysis of autoregressive integrated moving average (ARIMA) models is presented in this section. We present a GP analysis. The ACF plot in Figure 4.20 includes a few of spikes following the first lag, and three of them exceed the confidence interval, indicating a significant correlation between the value itself and the values corresponding to its first lag. There is a significant drop-off after the first lag. There are a few of spikes after the first lag, and five spikes that exceed the confidence interval, indicating a significant correlation between the value itself and the values corresponding to its first, second, and third lags. There is a significant drop-off after the third lag.

4.12.1 Model identification

In this section, we estimate the performance of various competing models with the hope of choosing the best one. The same approach that was used in the section on MP previously is being applied here.

Table 4.50: Model summary

Model	ARIMA(p,d,q)(P,D,Q)
A	ARIMA $(1,1,1) \times (1,1,1)_{12}$
B	ARIMA $(1,1,2) \times (1,1,1)_{12}$
C	ARIMA$(1,1,3) \times (1,1,1)_{12}$

The following models are constructed using the ACF and PACF mentioned in Figure 4.20. ARIMA models that best fit the data based on observing the ACF and PACF of the first differencing data. The ACF and PACF were examined, and the following three ARIMA models were summarised in Table 4.50.

Table 4.51: Fit statistics for model A

Fit statistics	Mean
Stationary R-squared	0.448
R-squared	0.685
RMSE	40.025
MAPE	6.471
MAE	30.243
Normalised BIC	7.580

The results in Table 4.51 show that the value of R-squared is 0.685, which explains about 69% of the variability in deaths from injuries among GP. The RMSE value is 40.025, suggesting a 40.025 weighted average error between forecasts and actuals, while the MAE is 30.243, showing a 30.243 average error between forecasts and actuals. The MAPE value is 6.471, which means that on average, the forecast is off by 6.5%.

Table 4.52: Parameter estimation for model A

Parameters	estimate	SE	t-stats	p-value
Constant	-0.142	0.476	-0.297	0.767
AR(1)	0.090	0.152	0.592	0.555
MA(1)	0.672	0.116	5.805	0.000
SAR(1)	0.128	0.152	0.839	0.403
SMA(1)	0.833	0.166	5.009	0.000

Table 4.52 shows a parameter estimate for the ARIMA $(1,1,1)\times(1,1,1)_{12}$ model. The p-value for the estimate is also shown. The model contains four parameters: AR(1), MA(1), SAR(1) and SMA(1). The constant estimate, with a p-value of 0.767, is a function of the mean and not significant at the 5% significance level. The MA(1) with a p-value less than 0.05 and the SMA(1) recorded p-values less than 0.05, both of which are significant because their p-values are less than alpha 5%. The AR(1) recorded a p-value of 0.555 and the SAR(1) recorded a p-value of 0.403, which are not significant at the 5% significance level.

Table 4.53: Fit statistics for model B

Fit statistics	Mean
Stationary R-squared	0.448
R-squared	0.685
RMSE	40.194
MAPE	6.470
MAE	30.230
Normalised BIC	7.628

The results in Table 4.53 reveal that the value of R-squared is recorded at 0.685, which explains about 69% of the variation in death due to injury in GP. The RMSE value is 40.194, suggesting a 40.194 weighted average error between forecasts and actuals, while the MAE is 30.230, showing a 30.230 average error between forecasts and actuals. The MAPE value is 6.470, which means that on average, the forecast is off by 6.5%.

Table 4.54: Parameter estimation for model B

Parameters	Estimate	SE	t-stats	p-value
Constant	-0.139	0.481	-0.289	0.773
AR(1)	-0.065	1.625	-0.040	0.968
MA(1)	0.516	1.621	0.319	0.751
MA(2)	0.099	0.997	0.099	0.921
SAR(1)	0.129	0.154	0.841	0.402
SMA(1)	0.832	0.167	4.987	0.000

The parameter estimates for the ARIMA $(1,1,2) \times (1,1,1)_{12}$ model models are shown in Table 4.54. It also shows the p-value for the estimates. The model contains five parameters: AR(1), MA(1), MA(2), SAR(1), and SMA(1). The constant estimate, with a p-value of 0.773, is a function of the mean term and is not statistically significant at the 5% significance level. SMA(1) had a p-value less than 0.05, which is significant because their p-values were less than alpha 5%. The p-values for the AR(1), MA(1), MA(2), and SAR(1) were 0.968, 0.751, 0.921, and 0.402, respectively, which are not significant at the 5% level of significance.

Table 4.55: Fit statistics for model C

Fit statistics	Mean
Stationary R-squared	0.452
R-squared	0.688
RMSE	40.237
MAPE	6.378
MAE	29.778
Normalized BIC	7.671

Table 4.55 reveals that the value of R-squared is recorded at 0.688, which explains about 69% of the variation in death due to injury in GP. The RMSE value is 40.237, suggesting a 40.237 weighted average error between forecasts and actuals, while the MAE is 29.778, showing a 29.778 average error between forecasts and actuals. The MAPE value is 6.378, which means that on average, the forecast is off by 6.4%.

Table 4.56: Parameter estimation for model C

Parameters	Estimate	SE	t-stats	p-value
Constant	-0.100	0.605	-0.166	0.869
AR(1)	0.856	0.305	2.804	0.006
MA(1)	1.447	0.310	4.668	0.000
MA(2)	-0.416	0.257	-1.617	0.109
MA(3)	-0.099	0.097	-1.017	0.312
SAR(1)	0.109	0.156	0.693	0.490
SMA(1)	0.813	0.156	5.220	0.000

Table 4.56 shows the parameter estimates for the ARIMA $(1,1,3) \times (1,1,1)_{12}$ model. It also shows the p-value for estimates. There are six parameters in the model: AR(1), MA(1), MA(2), MA(3), SAR(1), and SMA(1). The constant estimate with a p-value of 0.869 is the function of the mean term and is not significant at the 5% significance level. The AR(1), MA(1), and SMA(1) recorded p-values of 0.006, less than 0.05, which are significant since their p-values are smaller than alpha 5%. The MA(2), MA(3), and SAR(1) recorded p-values of 0.109, 0.312, and 0.490, respectively, which are not significant at the 5% level of significance.

4.12.2 Model diagnostics

Model diagnostics involve checking how well the model fits. The residual analysis will be used.

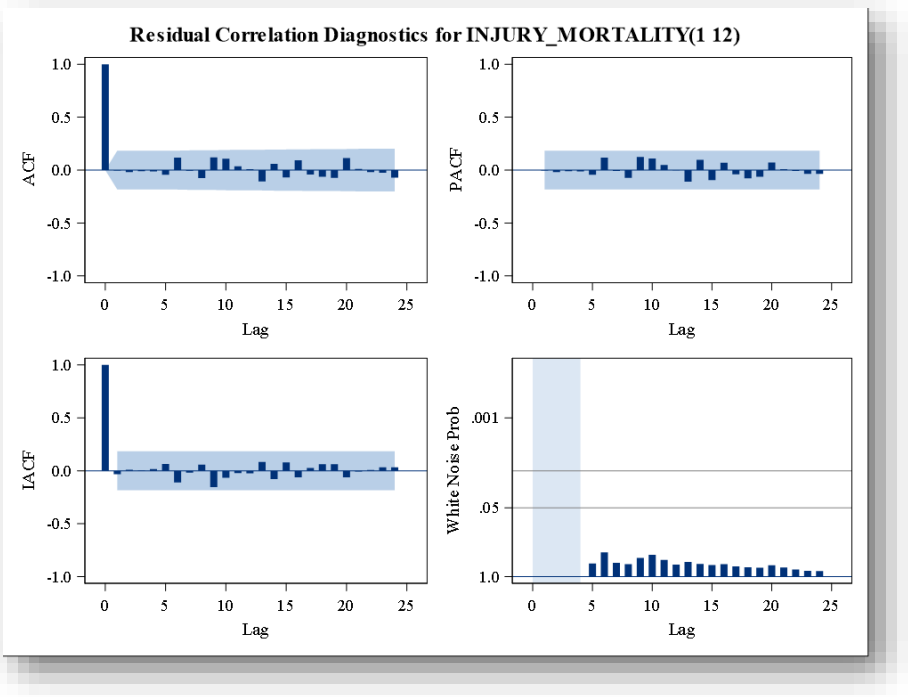

Figure 4.28: Residual ACF and PACF

The results in Figure 4.28 show ACF and PACF residuals that have non-significant peaks, all of whose values lie within the upper and lower confidence interval and therefore the model is white noise. There is no autocorrelation in the residuals.

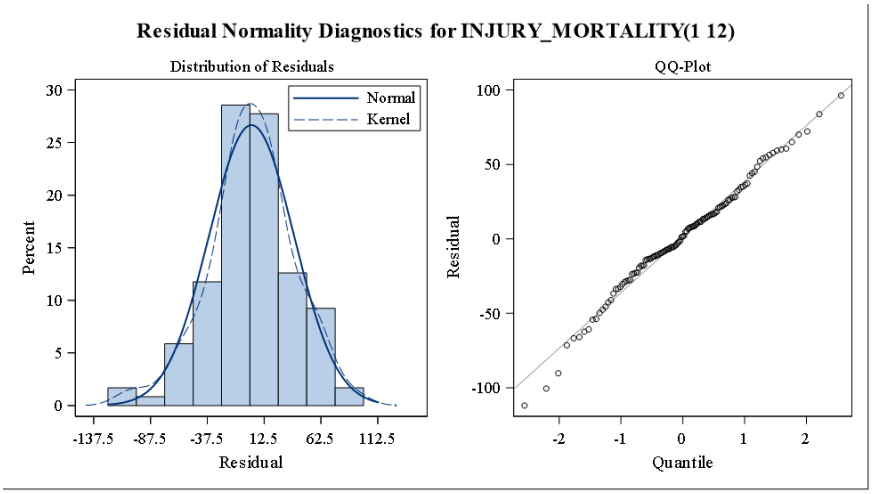

Figure 4.29: Residual Histogram and Q-Q plot

The results in Figure 4.29 show that the histogram data is normally distributed. In the same Figure 4.29, the Q-Q plot appears to be normally distributed because the points lie on the line, although there are points off the line at the bottom of the line. This does not imply that we should reject the normality of error terms in this model.

Table 4.57: Ljung Box test for model A

Model	t-stats	Df	p-value
ARIMA $(1,1,1) \times (1,1,1)_{12}$	10.823	14	0.700

Table 4.57 reveal the Ljung Box test $Q* = 10.823$ with a probability of 0.700, indicating that the estimated model is uncorrelated and may be adequate. The model is not significant at 5% level of significance since the p-value of 0.700 is greater than the level of significance of 5%.

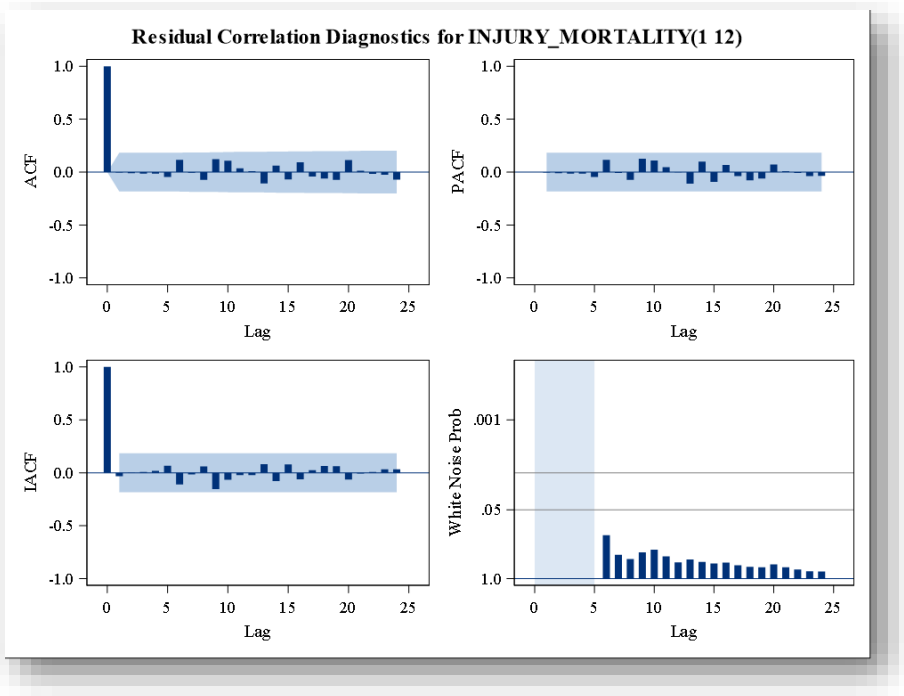

Figure 4.30: Residual ACF and PACF

The results in Figure 4.30 show ACF and PACF residuals that have non-significant peaks, all of whose values lie within the upper and lower confidence intervals, and therefore the model is white noise. There is no autocorrelation in the residuals.

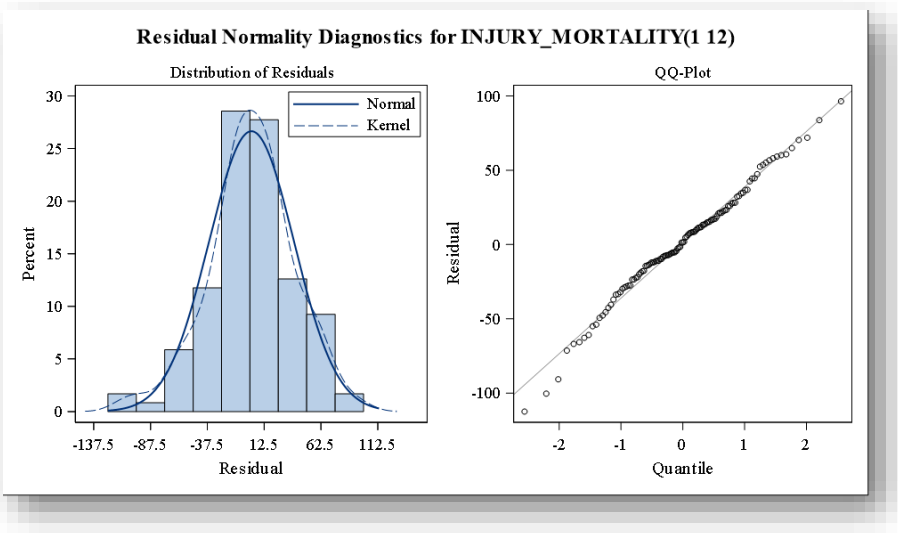

Figure 4.31: Residual Histogram and Q-Q plot

Figure 4.31 shows that the histogram data is normally distributed. The Q-Q plot in Figure 4.31appears to be normally distributed since the points are on the line, even though there are points away from the line near the bottom of the line.

Table 4.58: Ljung Box test for model B

Model	t-stats	Df	p-value
ARIMA $(1,1,2) \times (1,1,1)_{12}$	10.765	13	0.630

The results in Table 4.58 show the Ljung Box test $Q* = 10.765$ with a probability of 0.630, indicating that the estimated model is uncorrelated and may be appropriate. The model is not significant at the 5% significance level because the p-value of 0.700 is greater than the 5% significance level. This does not mean that we should reject the normality of error terms in this model.

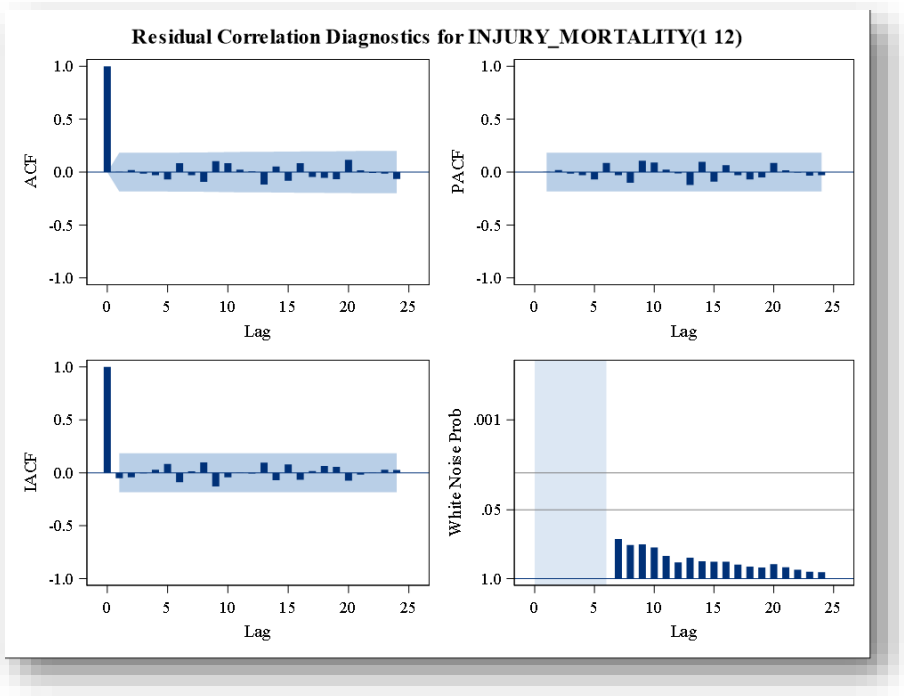

Figure 4.32: Residual ACF and PACF

The Figure 4.32 shows ACF and PACF residual which have non-significant peaks and all its values lie within the upper and lower confidence interval and therefore the model is a white noise.

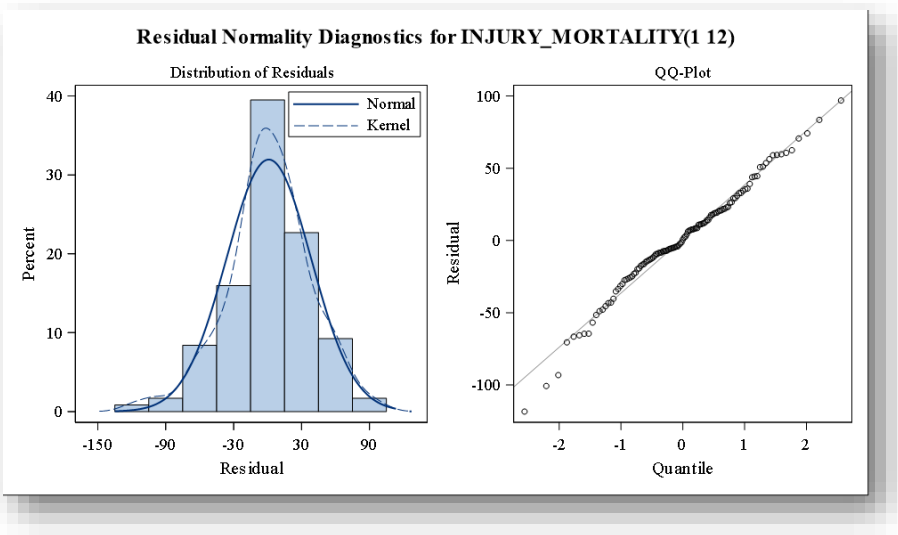

Figure 4.33: Residual Histogram and Q-Q plot

The results in Figure 4.33 show that the histogram data is normally distributed, while in the same Figure 4.33 the Q-Q plot appears to be normally distributed because the points are on the line, even though the points at the bottom of the line are off the line.

Table 4.59: Ljung Box test for model C

Model	t-stats	Df	p-value
ARIMA $(1,1,3) \times (1,1,1)_{12}$	9.903	12	0.624

The results in Table 4.59 show the Ljung Box test $Q* = 9.903$ with a probability of 0.624, indicating that the estimated model is uncorrelated and may be appropriate. The model is not significant at the 5% significance level because the p-value of 0.624 is greater than a 5% significance level.

4.12.3 Model assessment

The model is assessed through evaluating the prediction error on new data. The accuracy measure RMSE will be used to compare these time series models.

Table 4.60: Model assessment

Model	BIC	RMSE
ARIMA $(1,1,1) \times (1,1,1)_{12}$	7.580	40.025
ARIMA $(1,1,2) \times (1,1,1)_{12}$	7.628	40.194
ARIMA $(1,1,3) \times (1,1,1)_{12}$	7.671	40.237

Table 4.60 shows the results of a model assessment of GP monthly deaths from injuries. The best model selection is based on the normalised BIC and the RMSE. The lowest of BIC and RMSE is used for model selection. The ARIMA $(1,1,1) \times (1,1,1)_{12}$ model is the best model and has a minimum score of 7.580 and 40.025 for BIC and RMSE, respectively.

4.13 Model assessment in GP

The model is assessed through evaluating the prediction error on new data. The accuracy measure RMSE will be used to compare these time series models.

4.13.1 Comparing the NB-INGARCH and ARIMA Models

Table 4.61: Model assessment

Model	RMSE
ARIMA $(1,1,1) \times (1,1,1)_{12}$	40.025
NB-INGARCH$(1,2)$	469.44

Table 4.61 shows the findings of two models based on GP monthly injury mortality data. The best model is chosen using the RMSE. The best model is the one with the lowest RMSE value. The RMSE for the NB-INGARCH (1,1) model was 469.44, whereas the RMSE for the ARIMA $(1,1,1) \times (1,1,1)_{12}$ model was 40.025, making it the best model for GP monthly injury mortality data.

4.14 Injury mortality model in GP

The equation of ARIMA $(1,1,1) \times (1,1,1)_{12}$ model is given by

$$\phi_1(B)(1-B)\Phi_1(B^{12})(1-B^{12})x_t = \theta_1(B)\Theta_1 B^{12} z_t \tag{4.2}$$

$$(1-\phi_1 B)(1-B)(\Phi_1 B^{12})(1-B^{12})x_t = (1-\theta_1 B)(1-\Theta_1 B^{12})z_t$$

$$(1+0.090B)(1-B)(1+0.128B^{12})(1-B^{12})x_t = (1-0.672B)(1-0.833B^{12})z_t$$

$$\therefore x_t = 1.090x_{t-1} - 0.090x_{t-2} + 1.128x_{t-12} - 1.230x_{t-13} + 0.102x_{t-14} - 0.128x_{t-24} + 0.140x_{t-25} - 0.012x_{t-26} + z_t - 0.672z_{t-1} - 0.833z_{t-12} + 0.560z_{t-13}$$

4.15 Forecasting GP injury mortality

Once we have the final ARIMA model, we are now ready to make predictions on the future time points. We also visualise the trends.

4.15.1 GP monthly injury mortality forecast

Table 4.62: Forecast with 95% confidence limits

Date	Time	Forecast	Lower Limit	Upper Limit
Jan-19	133	326.6962	252.0641	401.3283
Feb-19	134	317.5735	236.6963	398.4507
Mar-19	135	375.9863	290.6337	461.339
Apr-19	136	340.8089	251.3093	430.3086
May-19	137	340.8284	247.3746	434.2823
Jun-19	138	368.4878	271.2412	465.7344
Jul-19	139	407.7643	306.8674	508.6612
Aug-19	140	412.4477	308.0281	516.8673
Sep-19	141	394.1221	286.2948	501.9494
Oct-19	142	379.6706	268.5400	490.8011
Nov-19	143	402.1555	287.8171	516.4939
Dec-19	144	434.1733	316.7146	551.632
Jan-20	145	299.4376	172.2082	426.6669
Feb-20	146	291.8360	159.5895	424.0825
Mar-20	147	349.5045	212.7216	486.2875
Apr-20	148	312.7717	171.6234	453.92
May-20	149	316.633	171.2526	462.0135
Jun-20	150	342.3356	192.843	491.8283
Jul-20	151	381.3962	227.9014	534.891
Aug-20	152	382.0806	224.6855	539.4757
Sep-20	153	367.1694	205.9683	528.3705
Oct-20	154	356.3719	191.4527	521.2912
Nov-20	155	373.9375	205.382	542.4929
Dec-20	156	408.0074	235.8926	580.1223

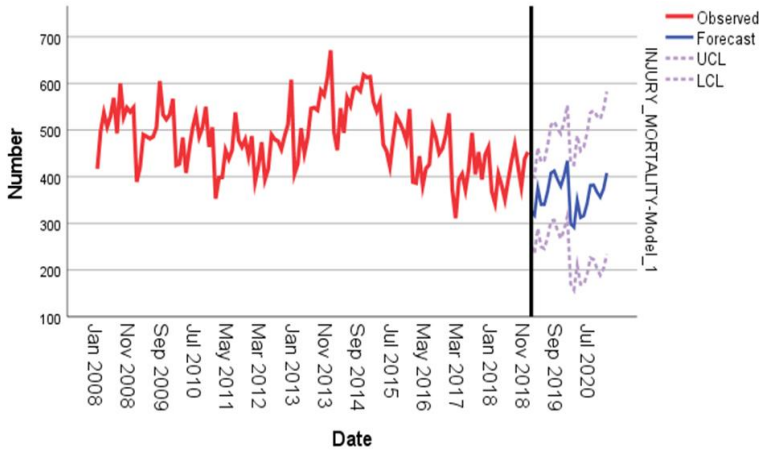

Figure 4.34: Time series plot for forecast

The results in Table 4.62 and Figure 4.34 show the forecasted data for injury mortality from January 2019 to December 2020. The GP forecasts show a decreasing trend from year to year on monthly basics. Furthermore, the forecasted results of the ARIMA$(1,1,1) \times (1,1,1)_{12}$ model show a decrease in injury mortality in 2020 as compared to 2019.

4.16 Discussion

The first thing we did in chapter four was to check the factors that contribute to injury mortality in MP and found that transport accidents followed by homicide were the highest, while in GP province, homicide followed by transport accidents were the highest. This is consistent with Matzopoulos et al. (2015) and WHO (2021) on MP regarding transport accidents and homicide as contributing factors to injury mortality. This result of MP could be because people are travelling from rural places to urban places. While the GP Pillay-van Wyk et al. (2016) study found that homicide was the leading cause of injury mortality, followed by transport accidents, these results could be because more people are

moving into urban areas. Moreover, as MP is a rural province while GP is an urban province according to Swart et al. (2012), there is a little variation between urban and rural regions in terms of age-group, month, and gender, but no difference in injury mortality. The only difference is in the causes of injury mortality. This observation is in line with Prinsloo et al. (2016). We also checked whether the series is stationary or not. The ADF, KPSS, and PP tests revealed that the series is stationary at the 5% level of significance after first differencing. The fitting of candidate distributions revealed that NB-INGARCH $(1, 1)$ and ARIMA $(1, 1, 1) \times (1, 1, 1)_{12}$ had appropriate distributions in MP, while NB-INGARCH $(1, 2)$ and ARIMA $(1, 1, 1) \times (1, 1, 1)_{12}$ had appropriate distributions in GP. The best model was chosen using the RMSE. The best model was the one with the lowest RMSE value. The RMSE revealed that ARIMA $(1, 1, 1) \times (1, 1, 1)_{12}$ were the appropriate distributions to model the series in both MP and GP. There is a trend of decreasing injury mortality from year to year, although there also appears to be a seasonal pattern with a cycle of less than two years. In addition, the variance of the data decreases over time. The reason why ARIMA models perform better could be that they represent stationary as well as non-stationary time series, while NB-INGARCH models are better at mainly capturing volatility in time series data (Bhardwaj et al., 2014).

Chapter 5

Conclusion

5.1 Introduction

The chapter presents the conclusions and recommendations based on the findings about monthly injury mortality. The chapter is focused on drawing conclusions from previous chapters. The chapter ends with some recommendations and future research directions.

5.2 Conclusions

The objectives of this study were stated in the first chapter. In terms of the first objective, descriptive statistics was used and the results revealed higher rates of injury mortality in Mpumalanga are driven by the high transport accidents followed by homicide, while in Gauteng the high injury mortality rates are driven by the high homicide in injury mortality followed by transport ac-

cidents. The second objective revealed that there are indeed differences for the injury mortality in both Mpumalanga and Gauteng. The findings highlight a significant higher likelihood for homicide in Gauteng which is an urban place compared to Mpumalanga which is rural. Transport accidents were also found to be lower in Gauteng. After transforming the data into stationary form, the third objective was assessed through modelling the data pattern. The findings highlight the fitted models NB-INGARCH (1,1), Poisson INGARCH (1,1) and ARIMA $(1,1,1) \times (1,1,1)_{12}$ in Mpumalanga, while in Gauteng the models fitted were NB-INGARCH(1,2), Poisson INGARCH(1,2) and ARIMA $(1,1,1) \times (1,1,1)_{12}$. In the fourth objective, the better model between Poisson INGARCH (1,1) and NB-INGARCH (1,1) was NB-INGARCH(1,1) based on the AIC and BIC criteria. However, comparing it to the ARIMA model by using RMSE, the ARIMA $(1,1,1) \times (1,1,1)_{12}$ was found to be the best to forecast Mpumalanga, while in Gauteng the better model between Poisson INGARCH(1,2) and NB-INGARCH(1,2) was NB-INGARCH(1,2) based on the values of the AIC and BIC. Furthermore, comparing it to the ARIMA model using RMSE, the ARIMA $(1,1,1) \times (1,1,1)_{12}$ was found to be the best for forecasting. Therefore, Mpumalanga and Gauteng monthly injury mortality model is the ARIMA $(1,1,1) \times (1,1,1)_{12}$. The forecast results of 24 months show that there will be a decrease in injury mortality in the last objective, but also that there are months like December, October and August where December has higher injury than the other two months.

5.3 Further discussions and recommendations

Injury is a truly global health issue with massive societal and economic consequences. The study has revealed some evidence demonstrating the importance of keeping track of South Africa's injury mortality rates in order to inform the 2030 Sustainable Development Goals (SDGs) for reducing injuries (Assembly,

2015). The World Health Organization (WHO) expressed the need for country-specific statistics broken down by age, gender, race, ethnicity, and other characteristics in order to track socioeconomically disadvantaged people within countries and enhancing international comparisons on injury statistics (Organization, 2016). In South Africa, reliable injury mortality rates are difficult to obtain, therefore these findings add value and inform South Africa's contribution to monitoring SDGs for injuries; and inform government and public health actors about the magnitude and public health relevance of specific factors related to injuries.

The high homicide and transport accidents mortality for both Mpumalanga and Gauteng indicate a need for urgent intervation in order to achieve the SDGs by 2030. In particular, the SDGs related to:

- "eliminate all forms of violence against all women and girls in public and private spheres, including trafficking and sexual and other types of exploitations"
- "To significantly reduce all forms of violence and related death rates everywhere".

In Mpumalanga, the rate of transport accidents was high, followed by homicide. The current strategies appear to be ineffective in reducing large-scale transport accidents. Policymakers should be concerned about the increased relative risk of transport accidents among people aged 15 to 34. This could be due to an increase in the number of motor vehicle and motorcycle drivers, as well as increased automobile ownership and the presence of youth driver communities. A Graduated Driver Licensing (GDL) system could benefit the youth driver population in both Mpumalanga and Gauteng. Evidence from the GDL system in New Zealand, the United States, and Australia suggests that it reduces the risk of death (Bates et al., 2014).

This system gradually exposes a new driver to complex driving situations while keeping them safe. The *first* phase of the GDL system consists of a learner phase, which provides new drivers with practical driving experience under supervision in a lower risk situation. The *provisional* license phase limits exposure to potentially dangerous situations, such as restrictions on driving at night, driving with passengers, driving after consuming alcohol, using a mobile phone, and vehicle power restrictions. Before a new driver can be issued a full, *open* phase license, he or she must pass an exit test. South Africa's current licensing system consists of a leaner phase and a full, open license phase, and the restrictions noted in the provisional phase of a GDL system could benefit if applied. Existing intervention such as drink driving campaigns should be enforced, in both Mpumalanga rural region and Gauteng urban region.

When compared to Mpumalanga, Gauteng has a higher homicide risk. Homicide prevention can only be addressed through interventions that have a broader social impact on violence prevention. This will necessitate an increase in government funding for social protection in both Gauteng and Mpumalanga. Security for old age, disability, housing, and unemployment, in particular (Rogers and Pridemore, 2013), will aid in mitigating the effects of societal inequality. According to Organization et al. (2010), regulating alcohol sales and raising alcohol prices may prevent all forms of violence, while improving drinking environments prevent youth violence. Attempts to reform South African policies on the sale, use, and advertising of alcohol have been met with stiff opposition from the liquor industry (Parry et al., 2014).

According to Krisch et al. (2015), violence can be reduced by investing in urban planning and focusing on hotspots through urban upgrading in fast-growing cities that are most prone to violence. Violence Prevention through Urban Upgrading (VPUU) was implemented in communities in the Western Cape of

South Africa, and it involved the modification and upgrading of public space as well as the monitoring of liquor outlets (Matzopoulos and Myers, 2014).

To monitor trends, more studies on injury mortality data are needed on a regular basis. This will necessitate the integration of data sources from forensic pathology, police, and forensic chemistry laboratories. A nationally representative survey of this scope can be expensive, necessitating support and resources from government or private funding sources. Also, the implementation of surveillance systems will present similar difficulties. While surveys can be completed in a specific time frame, maintaining ongoing surveillance systems will be more difficult. Surveys are thus more likely to be accurate and complete than surveillance systems in a South African setting.

Improved data quality is required for optimal resource allocation to high-risk groups. StatsSA data will need to be strengthened, which will necessitate a change in the death notification form. Medical certification training for doctors is also important, as well as complete death registration. This will reduce the proportion of misclassified deaths within vital registration data.

Studies on seasonal autoregressive integrated moving average (SARIMA) model to predict extreme seasonal injury mortality are proposed for future research directions.

References

ABIO, A., BOVET, P., DIDON, J., BÄRNIGHAUSEN, T., SHAIKH, M. A., POSTI, J. P., AND WILSON, M. L. (2020). Trends in Mortality from External Causes in the Republic of Seychelles Between 1989 and 2018. *Scientific Reports*, **10** (1), 1–11.

ABRAHAM, B. AND LEDOLTER, J. (2009). *Statistical Methods for Forecasting*, volume 234. John Wiley & Sons.

AHMAD, A. AND FRANCQ, C. (2016). Poisson QMLE of Count Time Series Models. *Journal of Time Series Analysis*, **37** (3), 291–314.

AKAIKE, H. (1973). Information Theory and an Extension of the Maximum Likelihood Principle,[w:] Proceedings of the 2nd International Symposium on Information, bn Petrow, f. *Czaki, Akademiai Kiado, Budapest*.

ANTENEH, A. AND ENDRIS, B. S. (2020). Injury Related Adult Deaths in Addis Ababa, Ethiopia: Analysis of Data from Verbal Autopsy. *BMC Public Health*, **20** (1), 1–8.

ARMITAGE, P., BERRY, G., AND MATTHEWS, J. N. S. (2008). *Statistical Methods in Medical Research*. John Wiley & Sons.

ASSEMBLY, G. (2015). Resolution adopted by the general assembly on 11 september 2015. Technical report, A/RES/69/315 15 September 2015. New York: United Nations.

ATKINSON, D. (2014). Rural-Urban Linkages: South Africa Case Study. *Territorial Cohesion for Development Program, Rimisp, Santiago*.

BALLESTEROS, M. F., WILLIAMS, D. D., MACK, K. A., SIMON, T. R., AND SLEET, D. A. (2018). The Epidemiology of Unintentional and Violence-related Injury Morbidity and Mortality Among Children and Adolescents in the United States. *International Journal of Environmental Research and Public Health*, **15** (4), 616.

BANK, W. (2011). *World Development Report 2011: Conflict, Security, and Development*. The World Bank.

BATES, L. J., ALLEN, S., ARMSTRONG, K., WATSON, B., KING, M. J., AND DAVEY, J. (2014). Graduated driver licensing: An international review. *Sultan Qaboos university medical journal*, **14** (4), e432.

BERK, R. AND MACDONALD, J. M. (2008). Overdispersion and poisson regression. *Journal of Quantitative Criminology*, **24**, 269–284.

BHARDWAJ, S., PAUL, R. K., SINGH, D., AND SINGH, K. (2014). An empirical investigation of arima and garch models in agricultural price forecasting. *Economic Affairs*, **59** (3), 415.

BOLLERSLEV, T. (1986). Generalized Autoregressive Conditional Heteroskedasticity. *Journal of Econometrics*, **31** (3), 307–327.

BRYSIEWICZ, P. (2001). Trauma in South Africa. *International Journal of Trauma Nursing*, **7** (4), 129–132.

CF, O. (2015). Transforming Our World: The 2030 Agenda for Sustainable Development. *United Nations: New York, NY, USA*.

CHRISTOU, V. AND FOKIANOS, K. (2014). Quasi-likelihood inference for negative binomial time series models. *Journal of Time Series Analysis*, **35** (1), 55–78.

COOK, R. D. AND WEISBERG, S. (1982). *Residuals and Influence in Regression*. New York: Chapman and Hall.

CURTIN, S. C., HERON, M., MINIÑO, A. M., AND WARNER, M. (2018). Recent Increases in Injury Mortality Among Children and Adolescents Aged 10-19 Years in the United States: 1999-2016. *National Vital Statistics Reports: From the Centers for Disease Control and Prevention, National Center for Health Statistics, National Vital Statistics System*, **67** (4), 1–16.

DE MYTTENAERE, A., GOLDEN, B., LE GRAND, B., AND ROSSI, F. (2016). Mean Absolute Percentage Error for Regression Models. *Neurocomputing*, **192**, 38–48.

dNATIONAL DEPARTMENT OF HEALTH, S. A. (2015). Strategic Plan 2015-2020.

ENGLE, R. F. (1982). Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation. *Econometrica: Journal of the Econometric Society*, 987–1007.

FERLAND, R., LATOUR, A., AND ORAICHI, D. (2006). Integer-Valued GARCH Process. *Journal of Time Series Analysis*, **27** (6), 923–942.

FOKIANOS, K. (2012). Count Time Series Models. *In Handbook of Statistics*, volume 30. Elsevier, pp. 315–347.

FOKIANOS, K., RAHBEK, A., AND TJØSTHEIM, D. (2009). Poisson Autoregression. *Journal of the American Statistical Association*, **104** (488), 1430–1439.

FREEMAN, M. F. AND TUKEY, J. W. (1950). Transformations Related to the Angular and the Square Root. *The Annals of Mathematical Statistics*, 607–611.

FULLER, W. A. (2009). *Introduction to Statistical Time Series*, volume 428. John Wiley & Sons.

GANTCHEV, G., SWART, L., LAHER, H., AND SEEDAT, M. (2015). Urban and Rural Differences in Child Injury Deaths in South Africa: A One-year review.

GARDNER, W., MULVEY, E. P., AND SHAW, E. C. (1995). Regression Analyses of Counts and Rates: Poisson, Overdispersed Poisson, and Negative Binomial Models. *Psychological Bulletin*, **118** (3), 392.

GARRIB, A., HERBST, A. J., HOSEGOOD, V., AND NEWELL, M.-L. (2011). Injury Mortality in Rural South Africa 2000–2007: Rates and Associated Factors. *Tropical Medicine & International Health*, **16** (4), 439–446.

GHASEMI, A. AND ZAHEDIASL, S. (2012). Normality Tests for Statistical Analysis: A Guide for Non-Statisticians. *International Journal of Endocrinology and Metabolism*, **10** (2), 486.

GOOSEN, J., BOWLEY, D. M., DEGIANNIS, E., AND PLANI, F. (2003). Trauma Care Systems in South Africa. *Injury*, **34** (9), 704–708.

HAIGHT, F. A. (1967). Handbook of the Poisson Distribution. Technical report.

HENLEY, G., HARRISON, J., AND AVEFUA, S. (2019). Trends in Injury Deaths, Australia, 1999–00 to 2016–17.

HERON, M. P. (2019). Deaths: Leading Causes for 2017.

HILBE, J. M. (2011). *Negative Binomial Regression*. Cambridge University Press.

HOLDER, Y. (2001). Injury Surveillance Guidelines: World Health Organization.

JAMES, S. L., CASTLE, C. D., DINGELS, Z. V., FOX, J. T., HAMILTON, E. B., LIU, Z., ROBERTS, N. L., SYLTE, D. O., HENRY, N. J., LEGRAND, K. E., ET AL. (2020). Global Injury Morbidity and Mortality from 1990 to 2017: Results from the Global Burden of Disease Study 2017. *Injury Prevention*, **26** (Supp 1), i96–i114.

KANE, E. J. (1948). Economic statistics and econometrics.

KMET, L. AND MACARTHUR, C. (2006). Urban–Rural Differences in Motor Vehicle Crash Fatality and Hospitalization Rates Among Children and Youth. *Accident Analysis & Prevention*, **38** (1), 122–127.

KRISCH, M., EISNER, M., MIKTON, C., AND BUTCHART, A. (2015). Global strategies to reduce violence by 50% in 30 years: Findings from the who and university of cambridge global violence reduction conference 2014.

KUTNER, M. H., NACHTSHEIM, C. J., NETER, J., AND LI, W. (2005). Applied Linear Statistical Models, 2005. *McGraw Hill Irwin, New York. NY*, 409.

LEE, L.-F. (1986). Specification Test for Poisson Regression Models. *International Economic Review*, 689–706.

LEILEI, D., PENGPENG, Y., HAAGSMA, J. A., YE, J., YUAN, W., YULIANG, E., XIAO, D., XIN, G., CUIRONG, J., LINHONG, W., ET AL. (2019). The Burden of Injury in China, 1990–2017: Findings from the Global Burden of Disease Study 2017. *The Lancet Public Health*, **4** (9), e449–e461.

LI, Y., PU, M., WANG, Y., FENG, T., AND JIANG, C. (2020). Analysis of the Reduction in Injury Mortality Disparity Between Urban and Rural Areas in Developing China from 2010 to 2016. *BMC Public Health*, **20** (1), 1–12.

LJUNG, G. M. AND BOX, G. E. (1978). On a Measure of Lack of Fit in Time Series Models. *Biometrika*, **65** (2), 297–303.

MATZOPOULOS, R. AND MYERS, J. E. (2014). The western cape government's new integrated provincial violence prevention policy framework: successes and challenges. *Aggression and Violent Behavior*, **19** (6), 649–654.

MATZOPOULOS, R., PRINSLOO, M., BRADSHAW, D., AND ABRAHAMS, N. (2019). Reducing Homicide Through Policy Interventions: The Case of Gun Control. *South African Medical Journal*, **109** (11b), 63–68.

MATZOPOULOS, R., PRINSLOO, M., WYK, V. P.-V., GWEBUSHE, N., MATHEWS, S., MARTIN, L. J., LAUBSCHER, R., ABRAHAMS, N., MSEMBURI, W., LOMBARD, C., ET AL. (2015). Injury-related Mortality in South Africa: A Retrospective Descriptive Study of Postmortem Investigations. *Bulletin of the World Health Organization*, **93**, 303–313.

MCCULLAGH, P. AND NELDER, J. (1989). Generalized Linear Models II.

MEEL, B. (2017). Incidence of Unnatural Deaths in Transkei Sub-region of South Africa (1996–2015). *South African Family Practice*, **59** (4), 138–142.

MOY, E., GARCIA, M. C., BASTIAN, B., ROSSEN, L. M., INGRAM, D. D., FAUL, M., MASSETTI, G. M., THOMAS, C. C., HONG, Y., YOON, P. W., ET AL. (2017). Leading Causes of Death in Nonmetropolitan and Metropolitan Areas—United States, 1999–2014. *MMWR Surveillance Summaries*, **66** (1), 1.

MURPHY, S. L., XU, J., KOCHANEK, K. D., AND ARIAS, E. (2018). Mortality in the United States, 2017.

MÜTZE, T., GLIMM, E., SCHMIDLI, H., AND FRIEDE, T. (2019). Group Sequential Designs for Negative Binomial Outcomes. *Statistical Methods in Medical Research*, **28** (8), 2326–2347.

MYUNG, I. J. (2003). Tutorial on maximum likelihood estimation. *Journal of mathematical Psychology*, **47** (1), 90–100.

NELDER, J. A. AND WEDDERBURN, R. W. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society: Series A (General)*, **135** (3), 370–384.

NEUMANN, M. H. (2011). Absolute Regularity and Ergodicity of Poisson Count Processes. *Bernoulli*, **17** (4), 1268–1284.

ORGANIZATION, W. H. (2016). *World health statistics 2016: monitoring health for the SDGs sustainable development goals*. World Health Organization.

ORGANIZATION, W. H. ET AL. (2010). *Injuries and violence: the facts*. World Health Organization.

PANKRATZ, A. (2009). *Forecasting with Univariate Box-Jenkins Models: Concepts and Cases*, volume 224. John Wiley & Sons.

PARRY, C., LONDON, L., AND MYERS, B. (2014). Delays in south africa's plans to ban alcohol advertising. *The Lancet*, **383** (9933), 1972.

PHILLIPS, P. C. AND PERRON, P. (1988). Testing for a Unit Root in Time Series Regression. *Biometrika*, **75** (2), 335–346.

PILLAY-VAN WYK, V., MSEMBURI, W., LAUBSCHER, R., DORRINGTON, R. E., GROENEWALD, P., GLASS, T., NOJILANA, B., JOUBERT, J. D., MATZOPOULOS, R., PRINSLOO, M., ET AL. (2016). Mortality Trends and Differentials in South Africa from 1997 to 2012: Second National Burden of Disease Study. *The Lancet Global Health*, **4** (9), e642–e653.

PRASAD, N. N. AND RAO, J. N. (1990). The Estimation of the Mean Squared Error of Small-Area Estimators. *Journal of the American statistical association*, **85** (409), 163–171.

PRINSLOO, M., MATZOPOULOS, R., LAUBSCHER, R., MYERS, J., AND BRADSHAW, D. (2016). Validating Homicide Rates in the Western Cape Province, South Africa: Findings from the 2009 Injury Mortality Survey. *South African Medical Journal*, **106** (2), 193–195.

RAATINIEMI, L., STEINVIK, T., LIISANANTTI, J., OHTONEN, P., MARTIKAINEN, M., ALAHUHTA, S., DEHLI, T., WISBORG, T., AND BAKKE, H. K. (2016). Fatal Injuries in Rural and Urban Areas in Northern Finland: A 5-year Retrospective Study. *Acta Anaesthesiologica Scandinavica*, **60** (5), 668–676.

REID, A. E., HENDRICKS, M. K., GROENEWALD, P., AND BRADSHAW, D. (2016). Where do Children Die and What are the Causes? under-5 Deaths in the Metro West Geographical Service Area of the Western Cape, South Africa, 2011. *South African Medical Journal*, **106** (4), 359–364.

ROBERTSON, L. S. (2015). *Injury Epidemiology*. Lulu. com.

ROGERS, M. L. AND PRIDEMORE, W. A. (2013). The effect of poverty and social protection on national homicide rates: Direct and moderating effects. *Social Science Research*, **42** (3), 584–595.

ROTH, G. A., ABATE, D., ABATE, K. H., ABAY, S. M., ABBAFATI, C., ABBASI, N., ABBASTABAR, H., ABD-ALLAH, F., ABDELA, J., ABDELALIM, A., ET AL. (2018). Global, Regional, and National Age-sex-specific Mortality for 282 Causes of Death in 195 Countries and Territories, 1980–2017: A Systematic Analysis for the Global Burden of Disease Study 2017. *The Lancet*, **392** (10159), 1736–1788.

SCHLOTTMANN, F., TYSON, A. F., CAIRNS, B. A., VARELA, C., AND CHARLES, A. G. (2017). Road Traffic Collisions in Malawi: Trends and Patterns of Mortality on Scene. *Malawi Medical Journal*, **29** (4), 301–305.

SHIN, Y. AND SCHMIDT, P. (1992). The KPSS Stationarity Test as a Unit Root Test. *Economics Letters*, **38** (4), 387–392.

STATSSA (2019). Mid-year Population Estimates—P0302. *Statistics South Africa*.

STATSSA (2021). Mortality and Causes of Death in South Africa: Findings from Death Notification 2018. https://www.statssa.gov.za/publications/p03093/p030932018.pdf [last access: 26 october 2021].

SUBRAMONEY, S. D., CHINHAMU, K., AND CHIFURIRA, R. (2021). Value at Risk Estimation Using GAS Models with Heavy Tailed Distributions for Cryptocurrencies. *International Journal of Finance & Banking Studies (2147-4486)*, **10** (4), 40–54.

SWART, L.-A., LAHER, H., SEEDAT, M., AND GANTCHEV, G. (2012). Urban and Rural Differences in Child Injury Deaths in South Africa: A One-year Review. *African Safety Promotion: A Journal of Injury and Violence Prevention*, **10** (2), 28–40.

THADEWALD, T. AND BÜNING, H. (2007). Jarque–Bera Test and its Competitors for Testing Normality–A Power Comparison. *Journal of Applied Statistics*, **34** (1), 87–105.

TYSON, A. F., VARELA, C., CAIRNS, B. A., AND CHARLES, A. G. (2015). Hospital Mortality following Trauma: An Analysis of a Hospital-based Injury Surveillance Registry in Sub-Saharan Africa. *Journal of Surgical Education*, **72** (4), e66–e72.

VELEZ, J. I. AND MORALES, J. C. C. (2015). A Modified QQ Plot for Large Sample Sizes. *Comunicaciones en Estadística*, **8** (2), 163–172.

WEISS, C. H. (2009). Modelling Time Series of Counts with Overdispersion. *Statistical Methods and Applications*, **18** (4), 507–519.

WHO (2019). Mortality. https://www.who.int/topics/mortality/en/ [last access: 15 january 2021].

WHO (2021). Injuries and Violence. https://www.who.int/news-room/fact-sheets/detail/injuries-and-violence [last access: 26 october 2021].

WILLMOTT, C. J. AND MATSUURA, K. (2005). Advantages of the Mean Absolute Error (MAE) over the Root Mean Square Error (RMSE) in Assessing Average Model Performance. *Climate Research*, **30** (1), 79–82.

WOOLF, B. (1957). The log likelihood ratio test (the g-test). *Annals of human genetics*, **21** (4), 397–409.

ZHU, F. (2011). A negative binomial integer-valued garch model. *Journal of Time Series Analysis*, **32** (1), 54–67.

ZIVOT, E. AND WANG, J. (2003). Rolling Analysis of Time Series. *In Modeling Financial Time Series with S-Plus®*. Springer, pp. 299–346.