

**MRR 06/2012**



## **Research Report**

# **Predicting Malaysian Road Fatalities for Year 2020**



Rohayu Sarani  
Sharifah Allyana Syed Mohamed Rahim  
Jamilah Mohd Marjan, PhD  
Wong Shaw Voon, PhD

**Research Report**

# **Predicting Malaysian Road Fatalities for Year 2020**

Rohayu Sarani  
Sharifah Allyana Syed Mohamed Rahim  
Jamilah Mohd Marjan, PhD  
Wong Shaw Voon, PhD

Published by:

**Malaysian Institute of Road Safety Research (MIROS)**

Lot 125-135, Jalan TKS 1, Taman Kajang Sentral,  
43000 Kajang, Selangor Darul Ehsan, Malaysia.

Perpustakaan Negara Malaysia

Cataloguing-in-Publication Data

Predicting Malaysian road fatalities for year 2020 / Rohayu Sarani ... [et al.]

Bibliography: p. 26

ISBN 978-967-5967-32-0

1. Traffic accident investigation--Malaysia--Forecasting.

2. Traffic safety--Malaysia. I. Rohayu Sarani.

363.125209595

For citation purposes

Rohayu S, Sharifah Allyana SMR, Jamilah MM & SV Wong (2012), Predicting Malaysian Road Fatalities for Year 2020, MRR 06/2012, Kuala Lumpur: Malaysian Institute of Road Safety Research.

Printed by: MIROS

Typeface : Calibri

Size : 11 pt

**DISCLAIMER**

None of the materials provided in this report may be used, reproduced or transmitted, in any form or by any means, electronic or mechanical, including recording or the use of any information storage and retrieval system, without written permission from MIROS. Any conclusion and opinions in this report may be subject to reevaluation in the event of any forthcoming additional information or investigations.

# Contents

	<b>Page</b>
<b>Acknowledgements</b>	<b>vii</b>
<b>Executive Summary</b>	<b>ix</b>
<b>1.0 Introduction</b>	<b>1</b>
1.1 Road Safety Development in Malaysia	1
1.2 Objective of the Study	4
<b>2.0 Literature Review</b>	<b>5</b>
2.1 Fatality Models	5
2.2 Fatality Models Used in Malaysia	6
<b>3.0 Methodology</b>	<b>10</b>
3.1 Data	10
3.2 Box-Jenkins Methodology and Multivariate Model	11
3.3 Generalised Linear Model	13
3.4 Poisson Regression Model	13
3.5 Over-dispersion in Poisson	14
<b>4.0 Analysis</b>	<b>16</b>
4.1 Scatter Plot	16
4.2 ARIMA Model	18
4.3 Model Identification	18
4.4 Model Estimates	20
4.5 Model Validation	20
4.6 Generalised Linear Model	21
<b>5.0 Result</b>	<b>23</b>
<b>6.0 Conclusion and Recommendation</b>	<b>25</b>
<b>References</b>	<b>26</b>



## Reviewer

Prof Law Heng Chin  
Deputy Dean (Mathematics)  
Division of Industry & Community Network  
School of Mathematical Sciences  
Universiti Sains Malaysia  
11800 USM, Penang, Malaysia



## Acknowledgements

This study would not have materialized without the support and motivation from the former Director-Generals of MIROS, Prof Ir Radin Umar Radin Sohadi and Prof Ahmad Farhan Mohd Sadullah, and also our colleagues at Accident Database and Analysis Unit (ADSA), MIROS. Greatest appreciation goes to our reviewer, Prof Law Heng Chin of Universiti Sains Malaysia for her constructive comments and recommendations. Thank you to Dr Law Teik Hua from Universiti Putra Malaysia for his valuable inputs. Finally, special thanks to ADSA Research Assistants for their help in completing this project.



## Executive Summary

This report highlights some fatality models developed in Malaysia in predicting road fatalities for the year 2020. Section 1.0 elaborates on the fatalities scenario in Malaysia from the 1970s to 2010. In the early 70s when Malaysia focused on agriculture as the main economic activity, fewer fatalities were recorded. Road fatalities increased in the 80s where an average of 4% annual increment in fatalities was observed. In the 90s, fatalities continued, becoming a bigger problem for the country with an average increment of 5% annually. An integrated approach to reducing road crashes started aggressively in the late 90s and the effect of the efforts were seen from 2000 to 2009, where annual fatalities were reduced to a lower rate of 2%.

There are basically three approaches to setting road safety targets, as mentioned in Section 2.0. Aspirational target is the easiest and has been adopted in some states in the US. Another example of aspirational target is Zero Fatality, implemented in New Zealand. Model-based targets, on the other hand, require the establishment of strong relationships between fatalities and the influencing factors, which are usually hard to establish. The third approach is extrapolation and evidence-led judgment where a baseline forecast is created based on future travel patterns and the number of fatalities is extrapolated. Potential countermeasures are then added to the model, and possible fatalities reduction calculated, and target-setting was then determined. Basically, Malaysia depends on fatality models and evidence-led judgment in setting a fatalities target. The Linear model was started in the early 90s by Aminuddin, and further improved by Rehan. Radin in 1998 used the Quasi Poisson model in developing his fatality models with more exposure variables. Law *et al.* in 2005 developed a new model based on the Gompertz growth model, taking into account the increment in population sizes and vehicle growth.

Section 3.0 covers the use of the ARIMA model in forecasting road fatalities. The ARIMA model was considered after data showed the existence of auto correlation. The independent variables such as population, cumulated vehicles registered, and road length were found to be highly correlated with fatalities, as expected. However, they are also correlated with each other and hence fail the assumption of multivariate modelling. Therefore, this paper continues to use ARIMA to predict fatalities for year 2020. In the analysis segment, Section 4.0 shows analysis for Generalised Linear Model (GLM)

## Predicting Malaysian Road Fatalities for Year 2020

following Poisson and Negative Binomial. ARIMA modeling was shown and based on the model fit; and managed to explain almost 98% variation in the data. Results showed that ARIMA performed better than the other two models. Based on the ARIMA model, predicted fatalities for the year 2015 is 8,760 and 10,716 for the year 2020.

## 1.0 Introduction

### 1.1 Road Safety Development in Malaysia

In 2004, road traffic injuries were the ninth most frequent cause of death and WHO predicted that by 2030, they would become the fifth most frequent (World Health Organization 2009). Over 1.2 million people die on the road each year, with 20 to 50 million suffering from non-fatal injuries. Malaysia, as one of the developing countries, has acknowledged road safety as a critical problem that should be addressed. In the 80s, Malaysia recorded an average of 4% annual increment of road traffic fatalities, followed by 5% in the 90s. However, the figure increased at a slower rate of 2% in more recent years (2000–2009). Even though the trend for the last 30 years is moving in a positive direction, where smaller increments have been recorded over the last nine-year period, the real number of fatalities is increasing. In 2010 alone, 6,872 fatalities were recorded (Royal Malaysian Police 2009), a 1.8% increase from the previous year (2009: 6,745). Figure 1 depicts Malaysian road traffic fatalities. Looking at it, one can see that there have been several turning points, indicated by oval shapes, throughout the years.

## Predicting Malaysian Road Fatalities for Year 2020

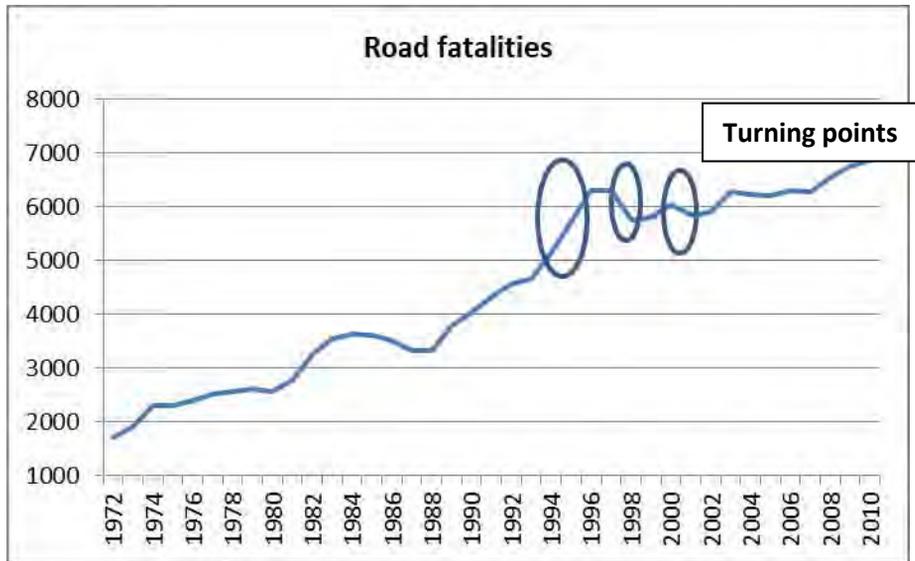


Figure 1 Malaysia road fatalities

There was a big jump in 1992–1996, where an increment of 11% in fatalities was recorded for two consecutive years (1993–1994 and 1994–1995). During the reduction period, a significant reduction in the year 1997–1998 (9% reduction) was observed. Then another reduction was observed in year 2000–2001 (3% reduction), and year 2003–2004 (1% reduction).

Various efforts were on stage to offset death figures. In the early 70s, the first motorcycle lane was built along Federal Highway, Route 2. A 16-kilometre stretch from the city of Kuala Lumpur to Subang International Airport was designed to separate motorcyclists from other vehicles, to reduce conflict and accidents. In 1993, the stretch was extended to Shah Alam and Klang under improvement projects of four-lane dual carriageway PLUS highway (Umar, MG and BL 1995). Generally, separating motorcycles from other vehicles has reduced motorcycle accidents by 34% (Umar, MG and BL 1995).

In 1989, The Road Safety Cabinet Committee was formed, with the responsibility of formulating a national road safety target. Pursuant to the target, a planning matrix was developed to categorise the actions to be taken and agencies responsible for carrying

out the countermeasures. In the 90s, early use of customised Microcomputer Accident Analysis Program (MAAP), licensed by Transport Research Laboratory (TRL), enabled Malaysia to assess blackspot analysis and conduct necessary treatment to the affected roads. Other than infrastructure treatment, there were efforts to change user behaviour, especially motorcyclists. In 1997, the Road Safety Research Centre (RSRC) under the Faculty of Engineering of Universiti Putra Malaysia was mandated to conduct research on motorcycle safety. Among the programmes implemented for motorcyclists were the exposure control programme, conspicuity programme, road engineering and injury control programme.

The efforts continue with more collaboration from other parties. Since the year 2001, an integrated Ops Sikap has been conducted during festive seasons to offset death toll. There are two types of Ops Sikap: a full one is conducted twice a year, with continuous enforcement for 14 days. This is during Hari Raya Aidilfitri and Chinese New Year. On the other hand, Ops Sikap for Deepavali and Christmas celebration is conducted for four to seven days. During festive seasons, where exposure is high, traffic police were stationed at various spots along roadsides for increased enforcement. Aside from traffic police, there were enforcement officers from other agencies such as Road Transport Department to ensure drivers adhered to traffic rules.

In 1996, the Malaysian Government established a five-year national road safety target with a target of 30% accident death reduction by year 2000. Various initiatives and interventions from national to community level were carried out to achieve the target. This included conspicuity programme, road enhancement programme, road user behavioural change, and accident prevention and reduction programme. In 2000, the reported accident deaths was 6,035; 5% lower compared to the predicted number of 6,389 deaths (Radin Umar 1998).

In efforts to be on par with developed countries, the Malaysian Road Safety Plan 2006–2010 (Road Safety Department 2006) was introduced. The plan outlined the following targets:

## Predicting Malaysian Road Fatalities for Year 2020

- a) reduction of 52.4% of deaths per 10,000 vehicles from 4.2 in 2005 to 2.0 in 2010;
- b) to reduce deaths per 100,000 population from the existing 23 (2005) to 10 deaths; and
- c) to reduce deaths per billion Vehicle Kilometre Travelled (VKT) from current 18 (2005) to 10 deaths.

In 2010, deaths per 10,000 vehicles stood at 3.4 and deaths per 100,000 people was 23.73. The government established the Malaysian Institute of Road Safety Research (MIROS) in 2007 under the 9th Malaysia Plan (RMK-9) to carry out research and identify scientifically-driven initiatives as the critical success factors of the safety investments in Malaysia. To help with the initiatives, the Road Safety Department (RSD) and MIROS were set up to specifically plan, coordinate, implement research and evaluate the safety interventions in the country.

### 1.2 Objective of the Study

Road safety target can be set arbitrarily, clearly stating the reduction in the number of fatalities a country would like to have in a certain year. It can also be established based on the percentage reduction a country would like to have for the next 10 years, for example. Malaysia established her road safety target based on her road fatality models, developed by Malaysian road safety experts.

This project attempts to develop new fatality model for Malaysia. There are two methods of modelling fatalities in Malaysia that will be considered as below.

- a) ARIMA Time Series model.
- b) Generalised Linear Model:
  - i) Radin's model 1998, predicted for year 2020; and
  - ii) Generalised Linear Model.

There are three models in total to be considered. The best model will be selected to predict year 2020 fatalities.

## 2.0 Literature Review

### 2.1 Fatality Models

Globally, there are various models used in forecasting road fatalities. Forecasting road fatalities is an important step in setting road safety target. Marsden in his paper (Marsden and Bonsall 2006) provided a summary of approaches of target-setting, basically categorised into three categories: 1) model-based, 2) extrapolation and evidence-led judgment, and 3) aspirational approaches. Young-Jun Kweon (Kweon 2010) summarised models used in developing data-driven reduction targets for highway safety plans. He noted that model-based was the most rigorous in setting targets. The need to understand the relationship between traffic safety and the influencing factors, most of the time failed to establish the real relationship, and hence was misleading.

As cited in Kweon (2010), Broughton *et al.* (2000) used a second approach, where he generated baseline forecasts based on future travel scenarios and extrapolation of crash victims, and applied potential countermeasures that would be most likely to be implemented, to the forecast to produce the 2010 reduction target. This is similar to Radin (Radin Umar 1998), where he estimated the exposure (vehicle, population and road length) for the forecasted year, and planned for a 30% reduction target by including potential road safety efforts, to come out with his target.

Zero fatality targets such as those implemented in New Zealand and some states in the US are an example of aspirational targets (Kweon 2010). These countries just set their targets but how the numbers are derived is not documented in publicly accessible documents.

The time series approach has been used in road safety fields, but mostly to evaluate road safety interventions. For example, Forester, McNown and Singell (1984) evaluated speed limit changes, and Garbacz and Kelly (1987) used log linear time series model to

## Predicting Malaysian Road Fatalities for Year 2020

evaluate the safety impact of vehicle inspection. In forecasting, Raeside and White (2004) used monthly data on fatal and seriously injured victims in traffic crashes in Great Britain. However, the limitation of the data makes their forecasts less reliable as they used eight years' data to forecast for another 10 years ahead. Raeside (2004) then used annual data from 1970 to 2002, employing autoregressive error model with lagged dependent variables, and forecast fatalities in 2010.

Road fatality is count data, and hence the normal approach of modelling is by using Poisson or Negative Binomial Generalised Linear Model. For data aggregated over time, the normal procedure is to use Integer-Value Autoregressive (INAR) Poisson model. However, Quddus in his paper (Quddus 2008), proved that if the data is aggregated annually, ARIMA models performed better than the INAR (Integer-Value Autoregressive) Poisson models. He used annual traffic fatalities in Great Britain and compared between INAR and ARIMA models. From his findings, ARIMA outperformed INAR in modelling traffic fatalities in Britain.

## 2.2 Fatality Models Used in Malaysia

Various fatality models have been developed for Malaysian road accident deaths, and used in the national road safety plan. Early works of modelling road death in Malaysia started with the development of a simple linear model by Aminuddin (1990). In the mid-nineties, Rehan (Mohamed Rehan 1995) proposed an improved model, similar to Smeed's, as follows:

$$Death = 0.08193 (population \times number\ of\ vehicle)^{0.335}$$

From the above model, Rehan projected 5,067 deaths in the year 2000, with estimated exposures of a population of 23 million and 10 million vehicles by the year 2000.

However, the increase in road length especially after the opening of a new expressway<sup>1</sup> in 1994, together with the increasing number of vehicles, created the need to revise the model. Based on the actual figures of population and number of vehicles, Rehan's model forecasted a much lower figure in comparison with the actual deaths.

Multivariate time series modelling was applied to estimate fatalities in Malaysia. Various studies have used log linear modelling following Poisson distribution; due to the nature of accident data, which falls under count data. This was necessary after considering the exponential growth in exposure variables. Radin and Hamid (1998) proposed that the rate of infrastructure growth in both roads and highways are highly correlated with  $r=0.95$ .

In light of improving the model, Radin (1998) added another two explanatory variables: road length and the effect of standardised accident data (Table 1). Additional data coverage to Sabah and Sarawak in 1981 exhibited a quantum leap in amount of data collected and hence influences the data interpretation.

---

<sup>1</sup> The North-South Expressway starts at Bukit Kayu Hitam in the north and ends in Johor Bahru in the south, linking all major cities on the west coast of Peninsular Malaysia between Thailand and Singapore. In addition, two other adjoining expressways complete the total 847.7 km of inter-urban toll highways referred to collectively as the North-South Expressway under PLUS' concession, namely the New Klang Valley Expressway and Federal Highway Route 2. Constructed in phases over a period of seven years, the North-South Expressway was officially opened on 8 September 1994 by the Prime Minister of Malaysia, YAB Dato' Seri Dr. Mahathir Mohamad, signalling the coming of age of Malaysia's road transportation system (*Source:* <http://www.plus.com.my/>).

## Predicting Malaysian Road Fatalities for Year 2020

**Table 1** Traffic exposures in Radin's model (Radin Umar 1998)

Variable	Details	Unit
Death	All death within 30 days due to road accident as reported to Royal Malaysian Police	Number of people killed
Population (P)	Population figure as reported by Statistics Department, Malaysia	In million
Vehicle (V)	Number of registered vehicles from Vehicles Registration Department	In million
Road (R)	Road length from Public Works Department	In thousand kilometre
Record system	Changes in accident recording system	0: Peninsular Malaysia only 1: Peninsular, Sabah and Sarawak

Some of the earlier studies (Radin *et al.* 1996; Homel 1994; Maycock and Summersgill 1994) also found that the shape of this model is suitable for explaining accident count although it requires some corrections caused by over-dispersion. To overcome the problem of over-dispersion, Radin used the 'quasi-likelihood' method proposed by McCullagh and Nedler (1983). Radin (Radin Umar 1998) developed an exponential model to explain fatalities in Malaysia as follows:

$$Death = 2289 (e^{0.00007 \text{vehicle} \cdot \text{population} \cdot \text{road}}) (e^{0.2073 \text{system}})$$

Using the above model, a projected death figure based on the business as usual (BAU) approach for year 2000 was established. The model predicted some 9,127 deaths in year 2000 if traffic exposure continued to increase at the projected rate of 23.2 million population, 10 million registered vehicles and 72,400 km of road length. As such, to offset the forecasted figure of 9,127 deaths, BAU approach should be changed. Intervention programmes were formulated and government commitment was obtained in reducing the forecasted deaths. Nevertheless, The Road Safety Cabinet Committee decided to maintain the earlier 30% reduction target from the base year approach to the interventional approach for year 2000.

Law *et al.* (2005) predicted road accident deaths in the year 2010 by making the projection for the vehicle ownership rate in the year 2010. The study used Gompertz growth model and Autoregressive Integrated Moving Average (ARIMA) to project the vehicle ownership model and to predict the road accident death rate respectively. The study used the number of registered vehicles, per-capita Gross Domestic Product and population from 1976 to 2001 as the dependent variable to develop the prediction model. The study revealed that vehicle ownership saturation occurs when per-capita GDP reaches 0.066 million annually and it was predicted that the ownership per population would be equal to 0.4409 by the year 2010. With this the road accident death rate is projected to decrease to 4.22 per 10,000 vehicles in 2010, at an average decline rate of 2.14% per annum. In 2010, Malaysia’s fatality rate stood at 3.2 fatalities per 10,000 vehicles. Table 2 below summarises previous fatality models developed in Malaysia.

Table 2 Previous fatality models in Malaysia

Author / Year	Variable	Model	Predicted fatalities and actual death
Aminuddin / 1990	Death	Linear model	4,950 <sup>p</sup> (Yr 2000) 6,035 <sup>a</sup>
Rehan / 1995	Population, number of vehicles	Death = 0.08193 (population X number of vehicle) <sup>0.335</sup>	5,067 <sup>p</sup> (Yr 2000) 6,035 <sup>a</sup>
Radin / 1998 (Radin Umar 1998)	Population, number of vehicles, road length, effect of standardised accident data	Death = 2289 (exp <sup>0.00007veh.pop.road</sup> ) (exp <sup>0.2073 data system</sup> )	9,127 <sup>p</sup> (Yr 2000) Plan for 30% reduction (6,389) 6,035 <sup>a</sup>
Law <i>et al.</i> / 2005 (Law, Radin Umar and Wong 2005)	Vehicle ownership rate	Gompertz Growth ARIMA model	4 <sup>p</sup> deaths/ 10,000 vehicles (Yr 2010) 3.98 <sup>a</sup> deaths/ 10,000 vehicles (Yr 2006)

<sup>p</sup> denotes predicted value; <sup>a</sup> denotes actual value

## 3.0 Methodology

### 3.1 Data

Road traffic fatality is based on the official road death fatalities figure reported by the Royal Malaysian Police (PDRM). This annual figure is obtained from the Annual Road Traffic Report. In Malaysia, fatalities are defined as any person who died within 30 days as a result of accident (Royal Malaysian Police 2009). Data is available from 1972 to 2010 (39 observations) and contains all road traffic fatalities from all 14 states in Malaysia.

The number of registered vehicles is taken from the Road Transport Department, and consists of cumulative registered vehicles for all types of vehicles. Each year, all commercial and private vehicles need to be registered with the Road Transport Department (RTD) when they renew their insurance.

Population and Gross Domestic Product (GDP per capita) figures are taken from the Department of Statistics Malaysia. For 2010, the estimated figure was taken from the Department of Statistics, Economic Planning Unit. Estimates for 2011–2015 were based on 10th Malaysian Plan by Economic Planning Unit. It is estimated that Malaysia will have a population of 29.8 million in 2015.

Road length (in kilometre) was taken from the Ministry of Works, Malaysia. The statistics used cover all Federal Roads built and maintained by the Department of Works, Malaysia. Data from 1972 was used based on historical data available, through year 2010.

### 3.2 Box-Jenkins Methodology and Multivariate Model

When data is collected over time, time series data analysis is more appropriate in forecasting future observations. There are many time series models that cater for different data type and structure, depending on the nature of data. Among the popular time series models is the Box-Jenkins approach, which is synonymous with ARIMA modelling. ARIMA modelling is usually applied to time series analysis, forecasting and control. The term ARIMA is a combination of Auto Regressive (AR) Integrated (I) Moving Average (MA) models. There are three stages in developing ARIMA models as depicted in the charts (Mohd Alias 2007) in Figure 2. Further explanation is available in the Analysis section.

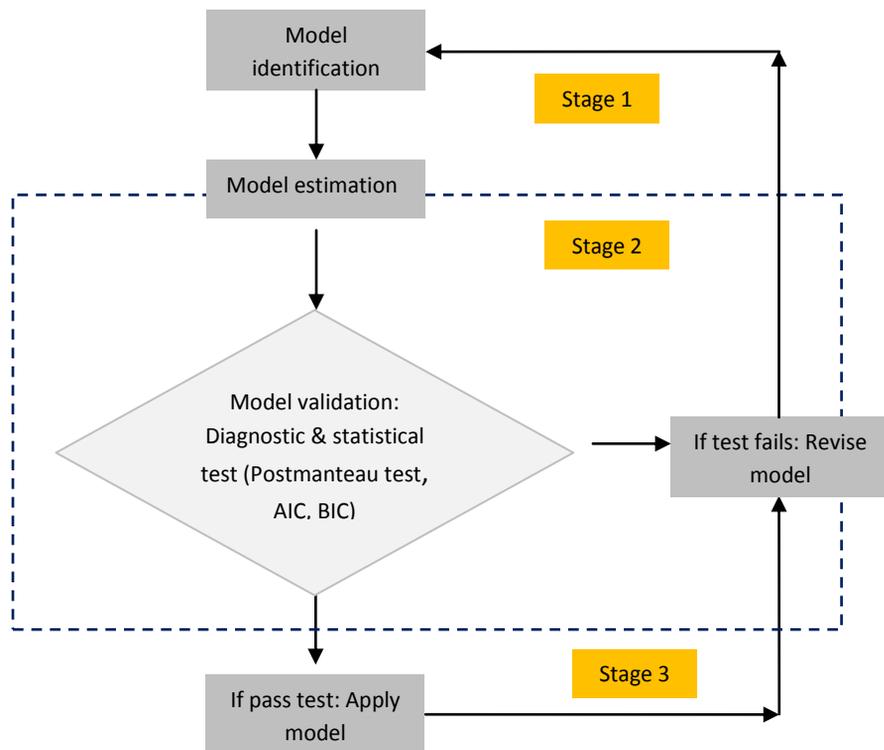


Figure 2 Stages in ARIMA modelling (Mohd Alias 2007)

## Predicting Malaysian Road Fatalities for Year 2020

A non-seasonal ARIMA model is classified as an "ARIMA (p,d,q)" model, where:

- p is the number of autoregressive terms,
- d is the number of non-seasonal differences, and
- q is the number of lagged forecast errors in the prediction equation.

ARIMA uses past values and past errors to detect patterns and predict future values. A simple ARIMA (0,1,1) with constant is expressed as below:

$$(1 - B)Y_t = (1 - \theta_1 B)e_t \quad (\text{Equation 1})$$

Where B is backward shift operator, which is used as follows:

$$BY_t = Y_{t-1}$$

Expert Modeller in SPSS v20 was employed in getting the best time series model that suits the data. It considered various models and picked the best model based on data itself. The acronym ARIMA stands for "Auto-Regressive Integrated Moving Average." Lags of the differenced series appearing in the forecasting equation are called "auto-regressive" terms; lags of the forecast errors are called "moving average" terms; and a time series which needs to be differenced to be made stationary is said to be an "integrated" version of a stationary series.

Other than ARIMA modelling, this study also considered multivariate count models, namely Generalised Linear Model (Poisson and Negative Binomial). This was due to the nature of fatalities data which is usually referred to as count data (data that takes only round numbers and positive values). Normal regression is not appropriate as it needs the assumption of normality. There are some data considerations that need to be tested before one can apply multivariate models. Variable selection plays an important role in multivariate models. If the wrong variables are selected, misspecification of the model might occur, leading to inaccurate and less reliable results. Multicollinearity problem, which can be detected through correlations analysis, is another threat to multivariate models.

### 3.3 Generalised Linear Model

In analysing count data, where death is part of it, normal linear regression analysis is not an appropriate tool to use. Under normal regression analysis, the dependent variable is assumed to be continuous data, which means that it can take any value of decimal, positive and negative. In this study, the dependent variable is the number of people who died on the road each year. It is considered count data where the number of deaths only takes positive and discrete value. An early paper (Miaou and Lum 1993) discussed some statistical issue and concluded that either the Poisson model or negative binomial model would be suitable to model accident count. However, where over-dispersion (where the mean is not equal to the variance) exists, negative binomial is a better model.

### 3.4 Poisson Regression Model

The Poisson regression model is often referred to as a benchmark model for modelling count data. Most of the count data in road safety modelling uses Poisson model as it suits the statistical properties of count data and is flexible enough to be reparameterised into other forms of distributional functions (Cameron and Trivedi 1998). In this model, the response variable is assumed to be independent and follows a Poisson distribution. It specifies that each observed count  $y_i$  is drawn from Poisson distribution with conditional mean of  $\mu_i$ , given vector  $X_i$  for case  $i$ . Thus the density function of  $y_i$  can be expressed as;

$$f(y_i|X_i) = \frac{e^{-\mu_i} \mu_i^{y_i}}{y_i!}, \text{ for } y = 0, 1, 2, \dots \quad (1)$$

where  $\mu_i = \exp(X_i' \beta)$ . In order to develop a Poisson model,  $\mu_i$  is expressed as a function of some explanatory variables through a log link function in the following form:

## Predicting Malaysian Road Fatalities for Year 2020

$$\ln \mu_i = X' \beta$$

$$\ln \mu_i = \alpha + \beta_1 x_1 + \dots + \beta_k x_k \quad (2)$$

or  $\mu_i = \exp(X' \beta)$

$$\mu_i = \exp(\alpha + \beta_1 x_1 + \dots + \beta_k x_k)$$

$$\mu_i = \exp^\alpha \left[ \left( \exp \right)^{(\alpha + \beta_1 x_1 + \dots + \beta_k x_k)} \right]$$

Given the independent observations assumption, with density function (1), the regression parameters  $\beta$  can be estimated using the maximum likelihood method based on maximum likelihood function of

$$\log_e L(\beta) = \sum \{ Y_i X_i' \beta - \exp(X_i' \beta) - \ln(Y_i!) \} \quad (3)$$

### 3.5 Over-dispersion in Poisson

The common problem when dealing with Poisson regression model is the problem of over-dispersion. The Poisson regression model should fulfil the assumption of equality of variance and mean, which is hardly met. If this assumption is not valid, the standard error estimated will be biased and the model will produce incorrect test statistics (Shankar 1995; Miaou 1994). When the variance of count data exceeds the mean,  $Var(Y_i) > E(Y_i)$  a feature of "over-dispersion" will occur. Otherwise, the Poisson model implies a property of so-called "equi-dispersion." When over-dispersion occurs, the Poisson maximum likelihood estimator obtained will be incorrect (Cameron and Trivedi 1998; 2005).

To overcome over-dispersion, several alternatives have been developed, such as Quasi Poisson estimation procedure. Wedderburn (1974) developed a quasi-likelihood estimation technique to estimate parameters under GLM model. This technique corrects for over-dispersion by multiplying the standard error of the model with the dispersion parameter value of  $\phi$ . McCullagh and Nelder (1989) suggested estimating the  $\phi$  as a ratio of the deviance or the Pearson Chi-Square to its degree of freedom. This treatment will not change parameter estimates or intercept but their standard errors will be corrected giving wider confidence interval and higher p-value.

## 4.0 Analysis

### 4.1 Scatter Plot

Before further analysis is conducted, it is essential to run scatter plot of each series involved against the dependent variable. Scatter plot is important to see visually the relationship and the direction of two variables.

Figure 3 shows scatter plot analysis for the independent variables and the dependent variable. It also shows how each independent variable reacts to another independent variable. From the scatter plot analysis, one can see that not all variables are linearly correlated with the dependent variable, fatalities. However, scatter plot only gives an early indication of how variables interact with each other, without any proper conclusive readings. Therefore, to confirm the relationship, bivariate correlation analysis is conducted (Table 3).

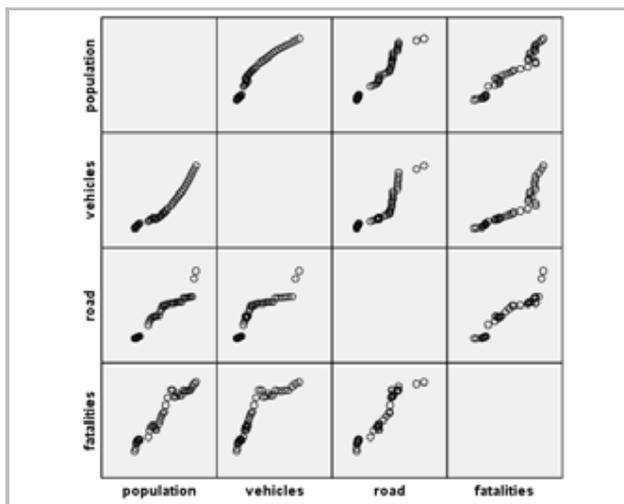


Figure 3 Scatter plot of variables

Through correlation analysis, it can be concluded that these independent variables are highly correlated with traffic fatalities, but at the same time, they are also highly correlated with each other. Hence, this leads to a multi-collinearity problem which, if ignored, will lead to invalid interpretation of parameter estimates. Therefore, in this study, univariate time series model will be used in forecasting for road traffic fatalities in 2020.

**Table 3** Correlations analysis (Dependent variable is fatalities)

		population	vehicles	fatalities	road	VPR
population	Pearson correlation	1	.962**	.968**	.968**	.917**
	Sig. (2-tailed)		.000	.000	.000	.000
	N	35	35	35	35	35
vehicles	Pearson correlation	.962**	1	.922**	.871**	.988**
	Sig. (2-tailed)	.000		.000	.000	.000
	N	35	35	35	35	35
fatalities	Pearson correlation	.968**	.922**	1	.950**	.868**
	Sig. (2-tailed)	.000	.000		.000	.000
	N	35	35	35	35	35
road	Pearson correlation	.968**	.871**	.950**	1	.803**
	Sig. (2-tailed)	.000	.000	.000		.000
	N	35	35	35	35	35
VPR	Pearson correlation	.917**	.988**	.868**	.803**	1
	Sig. (2-tailed)	.000	.000	.000	.000	
	N	35	35	35	35	35

\*Denotes significant correlation at alpha 0.05

## 4.2 ARIMA Model

Auto-Regressive Integrated Moving Average (ARIMA) was developed by Box and Jenkins, and has been widely used in road safety research. Quddus (2008), in his recent papers claimed that ARIMA can be used in long series of data, even if the data is count data like road fatalities. In developing the ARIMA model, data from 1972 to 2006 (36 observations) was used. Then the model developed was used to forecast for year 2007 to 2020 (14 observations). This is quite a long forecast horizon; hence the accuracy of predicted values will lessen as the horizon expands.

## 4.3 Model Identification

The hardest part in Box-Jenkins methodology is to identify which model really describes the data. The identification of the correct model starts with inspecting the autocorrelation coefficients (ACs) and Partial autocorrelation coefficients (PACs). These two measure the degree of interdependence among the observations in the series. In other words, the ACs and PACs measure the degree of interdependence between current and past fatalities figures. The ACs and PACs are reflected in the sample Autocorrelation Function (ACF) and Partial Auto correlation Function (PACF) below. Based on ACF and PACF, the most suitable model is identified and further analysis on model fit is identified.

From Figure 4, the sample ACF values are large and decline rather slowly to zero, therefore it can be concluded that the original series of road fatalities is not stationary. The PACF correlogram shows that there is a significant large spike at Lag 1, followed by other smaller spikes. These suggest that the original fatalities series can be made stationary after performing the first difference. After inspecting ACF and PACF of original series, it is concluded that no seasonal variation exists in the series. Therefore a non-seasonal ARIMA should be considered.

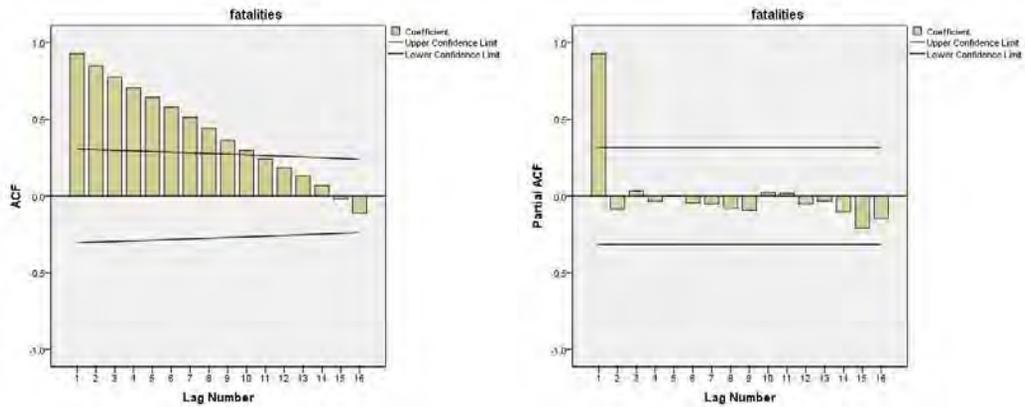


Figure 4 ACF and PACF for fatalities

First differencing plot is as shown in Figure 5. The ACF and PACF suggest that ARIMA (p,d,q) model should be used. The fatalities series is now stationary after the first difference. However, the next step to identify which ARIMA (p,d,q) that suits the data best, is more rigorous. A list of models to be considered should be checked carefully and tested for model fit and error rate.

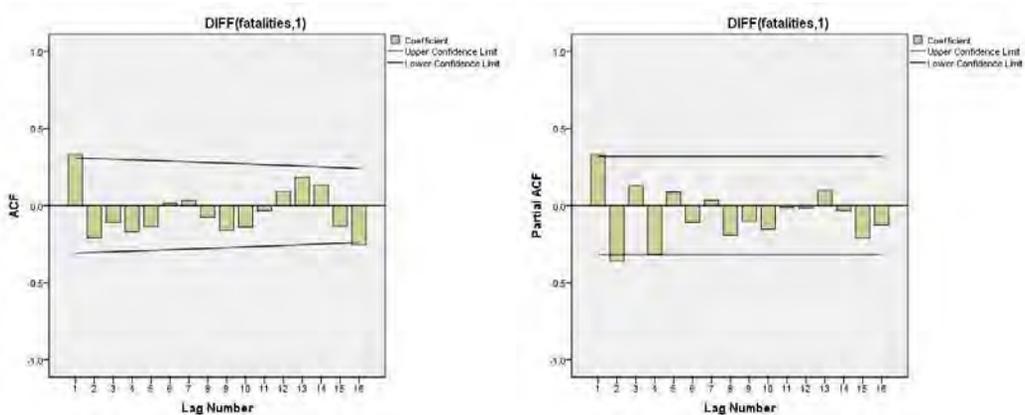


Figure 5 ACF and PACF for first difference of fatalities series

## 4.4 Model Estimates

SPSS version 20.0 has the capability of selecting which model is the best through its Expert Modeller function. Table 4 below illustrates the final ARIMA model selected to represent Malaysian road fatalities.

Table 4 Parameter estimates for ARIMA (0,1,1)

Model	Parameter	Estimate	Standard error	t-statistic	Sig.
Fatalities ARIMA (0,1,1)	Constant	0.036	0.016	2.226	0.033
	Natural log difference	1			
	MA Lag 1	-0.901	0.117	-7.712	0.000

The best ARIMA model produced from the data is ARIMA (0,1,1). The ARIMA(0,1,1) model *with* constant has the prediction equation:

$$\widehat{Y}_t = \mu + Y_{(t-1)} - \theta e_{(t-1)}$$

Where  $\widehat{Y}_t$  is the predicted fatalities at current time, t,  $Y_{(t-1)}$ , is fatalities a year before, and  $\theta = -0.901$ .

## 4.5 Model Validation

For model fit, R-squared value for ARIMA (0,1,1) is 0.981; in other words, the model suits the data. Ljung Box statistics, which provide an indication of whether the model is correctly specified, Hypothesis is accepted, as the Ljung Box statistics = 0.938 is more than alpha = 0.05. This indicates that residuals of ARIMA models have no auto correlation (Table 5).

Table 5 Goodness of fit ARIMA (0,1,1)

Model	Number of predictors	Model fit statistics		Ljung-Box Q(18)			Number of outliers
		Stationary R-squared	R-squared	Statistics	DF	Sig.	
fatalities-Model_1	0	.252	.981	9.075	17	.938	0

## 4.6 Generalised Linear Model

### i) Radin’s model

This model has been used by Radin (Radin Umar 1998) to set up Malaysian road fatalities target for year 2000. Radin developed his model in 1998 and his model is as below:

$$\text{Death} = 2289 (\exp^{0.00007\text{veh.pop.road}}) (\exp^{0.2073 \text{ data system}})$$

In order to make predictions for future observations (up to year 2020), the same model will be used. A series of independent variables used in the model needs to be forecasted beforehand. Extrapolation method using average yearly growth is used to forecast annual figures for vehicle, population and road length.

VPR is estimated using mean annual growth for the last 10 years (2001–2010). Annual growth for population is 2% annually, number of registered vehicles is 7% annually, and road length is estimated to be 1% annually. By using exponential model, the estimated fatalities tend to increase rapidly as time increases. VPR is assumed to be exponential, but the magnitude of increment in real value (yearly figures) is less than the predicted value in the model. Hence, this model provides the worst case scenario, as the increment from yearly fatalities for forecasted figure is almost 14%, much higher than annual of 0.5% average fatalities growth for year 2000–2010.

**ii) Generalised linear model – updating Radin’s model**

When Radin developed his model, he used data up to year 1998. When new data is available, the model needs to be updated. Using the same approach, the same method is used, but with longer time series (up to year 2006), GLM was performed. Due to the nature of data, new models created using the longer series and some variables that were significant before, were no longer significant.

From the result below (Table 6), Negative Binomial should be used as the log likelihood value is smaller than the Poisson model. The Poisson model has serious over-dispersion problem; scaled deviance is far from 1. However, Negative Binomial also showed that only intercept is significant. In other words, VPR (vehicle x population x road) and data do not have any significant impact on fatalities. If we use the above Negative Binomial model, which only depends on the mean value =  $\exp(7.745) = 2310$ , it will not give any better result, as it fails to incorporate mean changes over time.

**Table 6** Result from GLM procedure

Model	Variables used	B	Alpha value	Deviance	Log likelihood
Poisson	Intercept	7.746	0.000	96.205	-1716.174
	Data	0.489	0.000		
	VPR	0.0000225	0.000		
Negative Binomial	Intercept	7.745	0.000	0.023	-324.838
	Data	0.46	0.313		
	VPR	0.000025	0.292		

## 5.0 Result

From the above analysis, Poisson and Negative Binomial GLM seem to be inferior to the ARIMA model. The best model for predicting Malaysian road fatalities is ARIMA. ARIMA takes into account the previous observation and the past errors to observe patterns and make predictions. Based on the ARIMA model, it is predicted that the number of fatalities will continue to increase up to 8,760 (year 2015) and 10,716 (year 2020).

Figure 6 illustrates the original fatalities series as compared to the ARIMA model. It shows that ARIMA closely follows the series at the beginning of the forecast. As the forecast horizon gets bigger (more than five observations ahead), the forecasted value elevates as expected. To overcome this, as new data is added to the series, new ARIMA models should be developed and new forecasts will be generated.

### Predicting Malaysian Road Fatalities for Year 2020

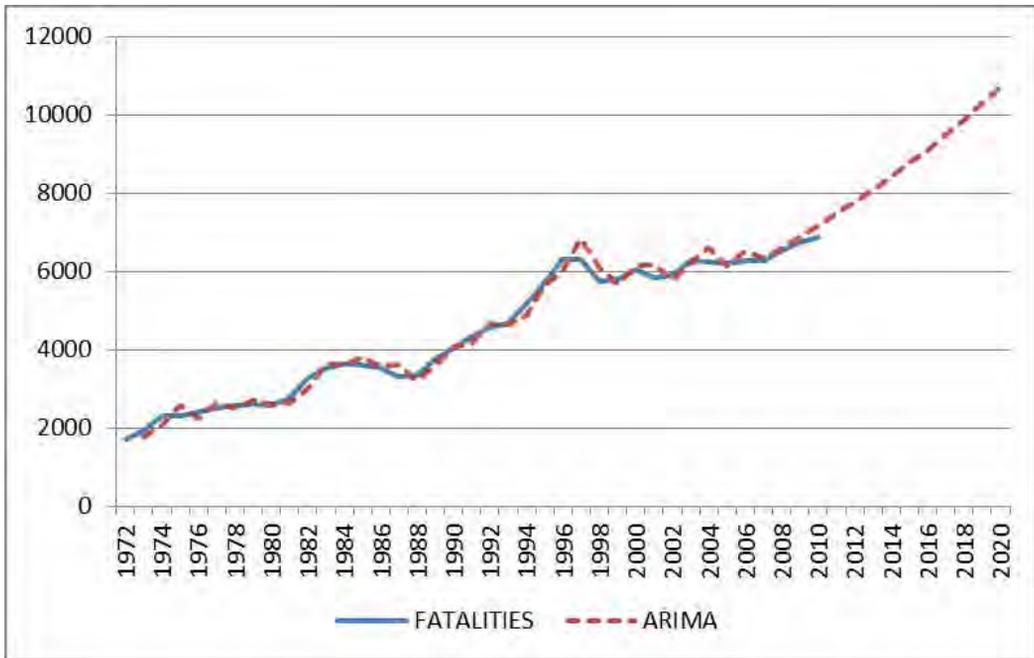


Figure 6 Fatality forecast for Malaysia

## 6.0 Conclusion and Recommendation

Target-setting in road safety is not an easy task. While many countries in the world use aspirational targets, Malaysia has been using an empirical method in setting her target. One of the important steps in target setting is predicting future fatalities. Road fatality is count data in nature, and usually the Poisson GLM will be used. Where over-dispersion exists, however, the Negative Binomial model is preferable.

There are many independent variables that can be considered in developing the GLM model. Among variables that have been used are population, vehicles on the road, road length, health status, GDP and many more. However, Mohan (Mohan 2011) indicates that there is no evidence that income level would necessarily affect fatality and vice versa. The GLM model using longer time series for Malaysia also showed that population, road length and number of vehicles on the road do not have a significant effect on Malaysian fatality.

Further study was conducted by using the fatalities series itself to produce a better model. The time series ARIMA model has been considered the best alternative to GLM in forecasting future fatalities especially if the data, such as annual fatalities, for example, is aggregated over time (Quddus 2008; Kweon 2010). ARIMA (0,1,1) is considered the best model for Malaysian fatalities series. Results showed that R-squared value is 0.981, really close to 1, indicating that the model is able to explain 98% of variation in the data. Ljung-Box statistics also showed that the error of the model is not correlated. Based on the ARIMA model produced, it is expected that the number of fatalities will increase to 8,760 (year 2015) and 10,716 (year 2020). More efforts and effective interventions should be created creatively to prevent fatal road crashes, to bring down the fatalities.

## References

- Aminuddin, A (1990), *Road safety in Malaysia*, Proceedings Road Safety Seminar, Kuala Lumpur: National Road Safety Council: 144–157.
- Broughton, J, Allsop, R, Lynam, D & McMahon, C (2000), *The numerical context for setting national casualty reduction target*, Crowthorne, United Kingdom: Transport Research Laboratory.
- Cameron, AC & Trivedi, PK (1998), *Regression analysis of count data*, Cambridge University Press.
- Dupont, Ea (2007), *Multilevel modeling and time series analysis in traffic research - methodology deliverable D7.4 of the EU FP6 pproject SafetyNet*, SafetyNet.
- Forester, T, McNown, R & Singell, L (1984), A cost-benefit analysis of the 55 mph speed limit, *Southern Economic Journal*, 50(3): 631–641.
- Garbacz, C & Kelly, J (1987), Automobile safety inspection: new econometric and benefit/cost estimates, *Applied Economics*, 19: 763–771.
- Homel, R (1994), Drink-driving law enforcement and the legal blood alcohol limit in New South Wales, *Accident Analysis and Prevention*, 26(2): 147–155.
- Kweon, Y-J (2010), Data-driven reduction targets for a highway safety plan, *Transport Policy*, 17: 230–239.
- Law, TH, Radin Umar, RS & Wong, SV (2005), The Malaysian government's road accident death reduction target for year 2010, *IATSS Research*, 29(1): 42–49.
- Marsden, G & Bonsall, P (2006), Performance targets in transportation policy, *Transport Policy*, 13(3): 191–203.

- Maycock, G & Summersgill, I (1994), *Methods for investigating the relationship between accidents*, Road User Behavior and Road Design Standards, Swov.
- McCullagh, P & Nelder, JA (1983), *Generalized linear models*, Chapman and Hall.
- McCullagh, P & Nelder, JA (1989), *Generalized linear models, 2nd Edition*, Chapman and Hall.
- Miaou, S & Lum, H (1993), Modelling vehicle accidents and highway geometric design relationships, *Accident Analysis and Prevention*, 25(6): 689–709.
- Miaou, S (1994), The relationship between truck accidents and geometric design of road sections: poisson versus negative binomial regressions, *Accident Analysis and Prevention*, 26(4): 471–482.
- Mohamed Rehan, K (1995), A macro analysis of road accident trends in Malaysia, *Journal of Eastern Asia Society for Transportation Studies*, 1(3): 941–950.
- Mohan, D (2011), Analysis of road traffic fatality data for Asia, *Journal of Eastern Asia Society for Transportation Studies*, 9: 1786–1794.
- Mohd Alias, L (2007), *Introductory business forecasting: a practical approach*, Shah Alam: University Publication Centre (UPENA).
- Quddus, M (2008), Time series count data model: an empirical application to traffic accidents, *Accident Analysis and Prevention*, 40: 1732–1741.
- Radin Umar RS, Mackay, MG & Hills, BL (1996), Modelling of conspicuity-related motorcycle accidents in Seremban and Shah Alam, Malaysia, *Accident Analysis and Prevention*, 28(3): 325–332.
- Radin Umar RS (1998), Model kematian jalan raya di Malaysia: unjuran tahun 2000, *Pertanika Journal of Science & Technology*, 6(2).
- Raeside, R (2004), Predicting and monitoring casualty numbers in Great Britain, *Journal of Transportation and Statistics*, 7(1): 61–68.
- Raeside, R & White, D (2004), Predicting casualty numbers in Great Britain, *83rd Annual Meeting of the Transportation Research Board*, Washington DC.

## Predicting Malaysian Road Fatalities for Year 2020

Road Safety Department (2006), *Road safety plan of Malaysia 2006-2010*, Putrajaya: Ministry of Transport, Malaysia.

Royal Malaysian Police (2009), *Annual road accident statistics*, Bukit Aman, Kuala Lumpur: Traffic Division.

Shankar, V, Mannering, F & Barfield, W (1995), Effect of roadway geometries and enviromental factors on rural freeway accident frequencies, *Accident Analysis and Prevention*, 27(3): 371–389.

Umar, R, MG, M & BL, H (1995), Preliminary analysis of exclusive motorcycle lanes along Federal Highway F02, Shah Alam, Malaysia, *IATSS Research*,19.

Wedderburn, RWM (1974), Quasi-likelihood functions, generalized linear models, and the gauss-newton method, *Biometrika*, 61(3): 439–447.

World Health Organization (2009), *Global status report on road safety: time for action*, Gevena: World Health Organization.







Research Report

# Predicting Malaysian Road Fatalities for Year 2020

Designed by: Publications Unit, MIROS



**Malaysian Institute of Road Safety Research**

Lot 125-135, Jalan TKS 1, Taman Kajang Sentral  
43000 Kajang, Selangor Darul Ehsan

**Tel** +603 8924 9200 **Fax** + 603 8733 2005

**Website** [www.miros.gov.my](http://www.miros.gov.my) **Email** [dg@miros.gov.my](mailto:dg@miros.gov.my)

ISBN 978-967-5967-32-0



9 789675 967320