Scientific Research Publishing

Current Urban Studies, 2024, 12, 107-122
https://www.scirp.org/journal/cus
ISSN Online: 2328-4919
ISSN Print: 2328-4900

Comparative Performance of Two Sets of Global Traffic Death Models Using Line of Equality (LOE) Statistical Approach

Nouri Gsouda, Utpal Dutta*, Xiaohui Zhong

Civil Architectural & Environmental Engineering, University of Detroit Mercy, Detroit, MI, USA
Email: nouri.gsouda@gmail.com, *Duttau@udmercy.edu, Zhongk@udmercy.edu

How to cite this paper: Gsouda, N., Dutta, U., & Zhong, X. H. (2024). Comparative Performance of Two Sets of Global Traffic Death Models Using Line of Equality (LOE) Statistical Approach. *Current Urban Studies, 12,* 107-122.
https://doi.org/10.4236/cus.2024.122006

Received: February 21, 2024
Accepted: May 25, 2024
Published: May 28, 2024

Abstract

Two sets of models were developed to predict global traffic deaths per million urban people by considering several socioeconomic factors. In one scenario, data were classified as low, middle, and high-income countries to create income-based models. In another scenario, data were divided into four clusters based on all attributes to develop cluster-based models. To identify minimum biased models, the Line of Equity (LOE) statistical approach was used. Two previously derived model sets were tested using a new set of socioeconomic information and traffic death data. The cluster-based models appeared to produce minimal bias while predicting traffic deaths per million urban people. This paper documents the journey of identification of minimum-biased models using the LOE approach.

Keywords

Cluster, Traffic Death, LOE, Income

1. Introduction

Traffic crashes remain one of the primary causes of mortality worldwide in both developed and developing countries, presenting widespread and devastating consequences on public health, the global economy, and poverty levels. According to the World Health Organization (WHO), approximately 1.35 million people die from traffic crashes worldwide; additionally, 20 to 50 million people suffer non-fatal injuries in traffic accidents yearly.

Figure 1 shows the number of traffic deaths and the death rate per 100,000 people between 2000 and 2016. Traffic crash injuries are considered the eighth-leading cause of death globally for people of all ages and are the primary

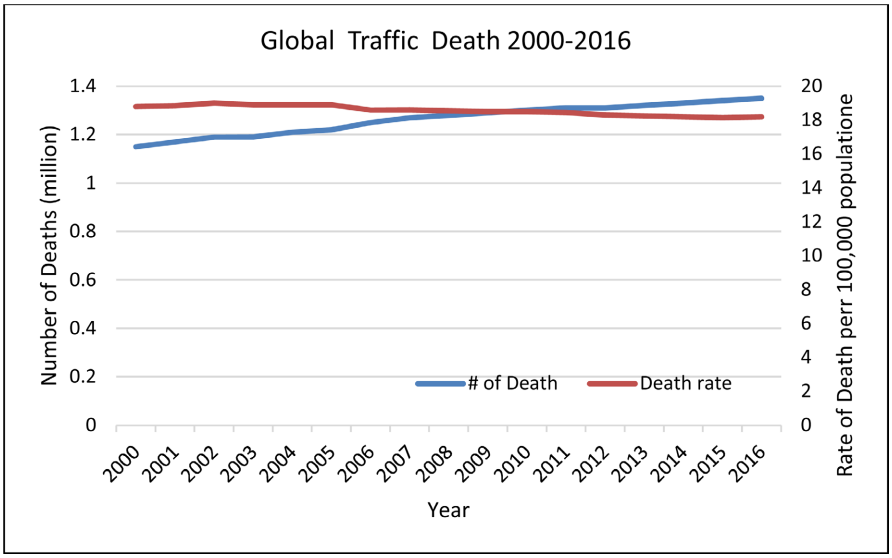

Figure 1. Number and rate of road traffic death per 100,000 population: 2000-2016 (WHO).

cause of death for young people aged 5 - 29 years (WHO, 2018). Furthermore, an estimated 0.9 million children aged 5 to 14 years died in 2018, about 2500 deaths in this age group a day (WHO, 2022). Thus, it is essential to study the relative factors contributing to traffic crashes within a global context.

To study the relation among global road traffic fatalities (RTF) and its related factors, three years (2007, 2010, 2013) of global traffic death and other related data were collected to develop models. World Health Organization (WHO) collects global traffic death data at regular intervals. The most recent global traffic death data currently available is from 2016. The collected global traffic death-related data were categorized into three levels based on each country's per capita gross national income level. GNI (WHO, 2018) Data were also collected from the World Bank, the United Nations, the United States Census Bureau, and other sources (UNESCO, 2015; IGI Global, 2018; UN, 2022; Economic Times, 2019; World Bank, 2022; KOF, 2019; US Census, 2024). Countries are assigned a low-, middle-, or high-income GNI based on 2016 income data. Countries with a GNI per capita of up to $1005, between $1006 to $12,235, and higher than $12,236 were designated as low-, middle-, and high-income level countries, respectively (WHO, 2018). Subsequently, countries were also grouped in clusters based on factors other than traffic death rates.

In this study, attempts were made to establish a relationship between traffic death rates and various socioeconomic factors. The factors considered are displayed in **Table 1**. Two sets of models were developed based on income classification and derived clusters. In the case of income-based approach, countries were classified based on their income levels, and three models were developed using step-wise registration analysis. The problem with this approach was while developing models for low-income countries with a lack of data availability for all considered variables. Additionally, some countries' statuses changed from

low to middle-income during 2007 and 2013, contributing to the low number of data available to fill the low-income category. Because there are many independent variables outside of income level that could have a large impact on the number of traffic crash fatalities, the cluster analysis approach was used to summarize the data in a meaningful way, creating small subsets based on shared resemblances and differences.

Clustering is used in unsupervised machine learning to group data points based on intrinsic characteristics, aiming to uncover hidden structures or relationships and divide a dataset into clusters. Clusters are formed based on the similarity or distance between data points. It is commonly measured using distance metrics like Euclidean distance or cosine similarity (Hastie et al., 2009).

There are various uses for clustering, including:

- Exploratory Data Analysis: Clustering helps understand data structure and distribution by identifying groups or clusters within the dataset. It can reveal patterns, trends, or outliers that may not appear at first glance.
- Customer segmentation in clustering helps businesses tailor marketing strategies, personalize recommendations, and effectively target specific customer segments.
- Using clustering techniques improves the retrieval of images and documents, the organization of large collections, and the identification of topics in texts.
- Clustering data points enables anomaly detection and identifies outliers or anomalies that may indicate unusual or important instances requiring further investigation.

Analyzing data with clustering algorithms such as K-means, hierarchical clustering, and DBSCAN helps analyze data by discovering meaningful patterns or groups. The choice depends on factors like data nature, the desired number of clusters, and the presence of any specific requirements or constraints. Clustering is considered as an effective tool for analyzing data and decision-making (Hastie et al., 2009).

For this study, K-means clustering was applied to derive clusters (Grubesic & Murray, 2001; Kumar & Toshniwal, 2015; Aggarwal & Reddy, 2013). Four clusters were derived and four clusters-based models were developed using step-wise regression analysis. In order to identify superior models among these two sets of models, the developed models were tested using the latest available global traffic death data of 2016. The model development process was presented in other publications (Dutta et al., 2022). This paper mainly documents the comparative performance of two sets of global traffic death models by subjecting them against a new set of data. The purpose of this approach is to identify less biased models using the Line of Equity (LOE) statistical approach. The Line of Equality (LOE) analysis is a method used in engineering and statistics to analyze the distribution of dependent variables within a population.

The concept of LOE is quite simple, representing a hypothetical line on a graph where everyone in the population would have equal sources. In other

words, it is the line where there is a perfect match between the predicted and observed variables. When conducting an LOE analysis, researchers typically plot the distribution of a predicted variable of a population on a graph, with the predicted variable on one axis and the observed variable on the other axis. The LOE is then drawn, usually as a straight diagonal line from the bottom left corner to the top right corner of the graph. By comparing the distribution of the observed data of the population with the Line of Equality, researchers can visually assess biases of the model. The part of the distribution curve that lies below the Line of Equality indicates that there is a zone of under-prediction. Conversely, the section of the curve that lies above the Line of Equality represents the zone of over-prediction. It is to be noted that a student was awarded a Ph.D. degree based on this modeling and validation work in 2023 (Nouri, 2023).

2. Method

2.1. Data Collection

Three years of global traffic death and other related data (2007, 2010, 2013) were collected from the World Health Organization (WHO), where global traffic death data were collected at regular intervals of 3 years. These data were used to develop models. At the point of time of this study, the latest global traffic death data available was for the year 2016, which became accessible in 2018. According to WHO, the global traffic death-related data were categorized into three levels based on a country's per capita gross national income level (GNI). As mentioned before, countries are assigned a low-, middle, or high-income GNI, based on 2016 income data. Countries with a GNI per capita of up to $1005, between $1006 to $12,235, and higher than $12,236, are designated as low-, middle-, and high-income level countries, respectively (WHO, 2018).

Since China, India, the United States, and Singapore represent outliers in population, number of crashes, and vehicle per population, they were removed from the database for this study. These countries can be studied separately if more detailed information is available. Data on some variables were collected directly and some were derived from the collected data as shown in Table 1. Descriptive statistics of income-based data is presented in Table 2.

Table 1. Various collected data.

Variables	Symbol	Direct	Cluster forming*
Number of Traffic Death per Million Urban Population	TRDPMUPOP	N	N
Number of Traffic Death per million Registered Vehicles	TRDHTRV	N	N
Number of Traffic Death/Population Density	TRDPOPDST	N	N
Number of Registered Vehicles/Population	NRVPOP	N	Y

Continued

Number of Registered Vehicles/Urban Population	NRVUPOP	N	Y
Income per Capita	Income	Y	Y
Education Index (0 - 1)	Edu_Index	Y	Y
Human Development Index (0 - 1)	HDI	Y	Y
Population density (people per sq. km of land area)	PoPD_SQKM	N	Y
Proportion of Urban Population	UrbPop_Per	N	Y
Social Globalization Index (0 - 100)	SoGl	Y	Y
Cultural Globalization Index (0 - 100)	CuGl	Y	Y
Alcohol Consumption (Liter/capita)	Alco_Consm	Y	Y
Average of Drivers Ages (years)	Avdr_Age	Y	Y

*The variables used to form clusters.

Table 2. Mean and standard deviation of variables by low-, middle-, and high-income countries.

Variables	Low-Income Countries		Middle-Income Countries		High-Income Countries	
	Mean	Std. Deviation	Mean	Std. Deviation	Mean	Std. Deviation
TRDPMUPOP	307.54	201.74	260.00	156.29	105.80	67.46
TRDHTRV	321.11	265	136.96	299.46	16.72	21.39
TRDPOPDST	23.6	20.68	161.59	512.26	68.5	315.4
NRVPOP	0.04	0.038	0.21	0.15	0.61	0.18
NRVUPOP	0.15	0.17	0.37	0.28	0.78	0.28
Income	548	277	5243	3006	36,457	19,379
Edu_Index	0.44	0.12	0.63	0.11	0.80	0.07
HDI	0.48	0.09	0.70	0.08	0.87	0.05
PoPD_SQKM	187	337	129	201	202	323
UrbPop_Per	27	9.15	58	17.13	78	13.01
SoGl	47.04	10.74	67.94	9.30	81.93	8.04
CuGl	45.93	15.35	61.04	14.37	79.39	15.79
Alco_Consm	2.51	2.92	4.73	3.47	9.05	3.49
Avdr_Age	36.2	1.31	40.21	3.11	45.21	3.66

2.2. Cluster Formation

Using the k-mean method, clusters were formed based on the variable indicated in Table 1. The following Table 3 is a summary of the clustering results.

Table 3. Mean and standard deviation of variables by clusters.

Variables	Cluster 1: Middle		Cluster 2: Low		Cluster 3: High		Cluster 4: Middle-High	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
TRDPMUPOP	269	150.70	314	214.84	84	44.40	182	85
TRDHTRV	116.32	177.50	380.84	499.34	9.18	3.95	34.77	25
TRDPOPDST	119.66	309.30	37.63	54.64	56.76	158.00	251.58	829.20
NRVPOP	0.19	0.13	0.05	0.054	0.69	0.12	0.39	0.13
NRVUPOP	0.33	0.24	0.16	0.18	0.89	0.24	0.60	0.29
Income	7,822	12,700	1,157	1080	43,038	17,352	11,961	6,481
Edu_Index	0.63	0.07	0.43	0.10	0.83	0.56	0.76	0.70
HDI	0.71	0.07	0.51	0.09	0.89	0.03	0.80	0.04
PoPD_SQKM	121.46	143.58	116.06	231.30	136.58	127.10	104.52	128.90
UrbPop_Per	60.83	17.46	31.34	11.61	77.72	11.3	68.24	15.60
SoGl	67.02	7.32	49.58	10.68	84.88	4.60	78.24	6.20
CuGl	59.51	13.23	44.66	15.3	85.64	8.00	73.92	9.10
Alco_Consm	3.55	2.47	2.61	2.76	9.97	2.19	9.57	3.10
Avdr_Age	39.14	2.19	36.48	1.32	46.68	1.60	44.76	2.30

3. Attributes of Clusters

After clusters were formed, various attributes of each cluster are summarized to make each group more representative.

Cluster 1—Middle

This cluster consists of 120 data points mainly from middle-income countries (L, M, H: 1, 111, 8). The average age of drivers is about 40 years, which is in the middle of all clusters. The income per capita of this group varies from $840 to $86,790, with an average of $7822, the second lowest among the four clusters. Since more than 80 percent of the population of this cluster has an income in the middle-income range, it can be considered as representative of middle-income countries as defined by WHO.

This population consumes an average of 3.5 liters of alcohol per person annually, which is significantly lower than those of clusters three (High) and four (Middle-High). The average education index is 0.63 on a scale of 0 - 1, compared to 0.43, 0.76, and 0.83 for clusters three (High), four (Middle-High), and two (Low) respectively. Sixty percent of the population lives in an urban area. One third of the urban population owns a vehicle, a rate twice that of cluster 2. The population density per square kilo-meter varies widely from 1.66 to 692 with an average of 120. The average social index is 0.67, which while lower than those of two other clusters (0.77 and 0.78 clusters 3 and 4), remains significantly higher than that of the cluster 2 (0.49).

Cluster 2—Low

This cluster consists of 45 data points (L, M, H: 28, 17, 0), with a maximum income of only $5170. The number of registered vehicles per population and urban population, human development, and social and cultural indices are at the lowest among all four clusters. The level of education is low (below 50th percentile). Average drivers are the youngest among the four clusters. Alcohol consumption is also the lowest among the four clusters. Less than one-third of the population resides in urban settings. More than 62 percent of the data points in this cluster are low-income countries, and the others are in the lower end of the middle-income countries. Thus, this cluster can be considered as representative of the low-income countries defined by WHO, named as Low. It is noticed that the traffic deaths per million urban population (TRDPMUPOP) is the highest among the four clusters, even though this variable was not considered when forming clusters.

Cluster 3—High

This cluster consists of 73 data points (L, M, H: 0, 0, 73), with the highest average per capita median income among all four clusters. The minimum income of this cluster population is $13,240 compared to $840 in cluster 1, $131 for Cluster 2, and $3570 for cluster 4 (Middle-High). The population of this cluster is well educated with an Education_Index of 0.83, the highest among the four clusters. In addition, the number of populations per SQ.km, percentage of people living in urban areas, number of registered vehicles per million urban population, and the average drivers' age are all the highest among the four clusters. This cluster can be considered representative of high-income countries. and thus, it is labeled as High. It is also noticed that this cluster has the lowest rate of traffic death per million of urban population.

Cluster 4—Middle-High

This cluster consists of 59 data points (L, M, H: 0, 34, 25). This cluster sits between clusters Middle and High many aspects, such as income, education index, HDI, and others. Cluster Middle-High has an average income more than $4000 higher than cluster Middle, but they consume as much as alcohol as cluster High. Car ownership (measured by NRVPOP and NRVUPOP) is much higher than those in clusters Middle and Low. Like other clusters, the urban population has higher car ownership than the overall population. The level of education as well as the average age of drivers are close to those of cluster High. The population density is the lowest among all clusters. About fifty-eight percent of the data points in this cluster consist of data points from middle-income countries and about forty-two percent of high-income countries. In terms of income, this cluster can be considered somewhere in the middle compared to the other clusters and is named Middle-High.

4. Traffic Death Models and Their Statistical Characteristics

Treating TRDMUPOP as a dependent variable, two sets of linear models were

developed using income-based and cluster-based data. Step-wise regression was used in the model development process considering 95 percent confidence level. These models are presented in Table 4 and Table 5. Detail model development process and selection was described in our earlier publication (Dutta et al., 2022).

4.1. Various Elements of Income Based Models

- The proportion of Urban Population is the only significant factor common to all three types of countries. The higher the Proportion of the Urban Population, the lower the traffic death rate when other factors are kept constant. Its effect is more significant in low-income countries.
- The number of Registered Vehicles/Population has been identified as a factor increasing the traffic death rates in low and middle-income countries, with more severe consequences in the low-income countries.

Table 4. Income based model parameters.

Dependent variable: TRDMUPOP	Income level Models		
Variables	**Low**	**Middle**	**High**
Constant	757	1433	2940
NRVPOP	14200 (0.000)	597 (0.000)	-
NRVUPOP	-	-	-
UrbPop_Per	−28.60 (0.000)	−8.05 (0.000)	−2.45 (0.000)
Income	-	-	-
SoGl	-	−6.67 (0.000)	-
CuGl	-	-	-
HDI	-	1590 (0.000)	−2490 (0.000)
Edu_Index	-	-	-
PoPD_SQKM	-	−0.31 (0.000)	−0.05 (0.000)
Alco_Consm	32.60 (0.004)	11.4 (0.01)	-
Avdr_Age	-	−37.40 (0.000)	−7.31 (0.000)
R Square	0.8	0.5	0.6
RMSE	96.96	113.87.	42.31
MEF	0.49	0.73	0.63

Note, ( ) represents p-value.

Table 5. Cluster based model parameters.

Dependent variable: TRDMUPOP	Cluster Based Models			
Variables	**Middle**	**Low**	**High**	**Middle-High**
Constant	2002.66	193.92	3579	3865
NRVPOP	963.76 (0.000)	-	-	-
NRVUPOP	-	871.50 (0.000)	-	111.00 (0.000)
UrbPop_Per	−6.97 (0.000)	−18.44 (0.000)	−3.22 (0.000)	−2.53 (0.013)
Income	-	1.15 (0.008)	-	-
SoGl	−8.94 (0.000)	-	-	−10.51 (0.000)
CuGl	-	-	-	-
HDI	-	-	−4086 (0.000)	−1925 (0.010)
Edu_Index	-	-	608.20 (0.004)	-
PoPD_SQKM	-	-	-	-
Alco_Consm	29.58 (0.000)	34.05 (0.007)	-	-
Avdr_Age	−24.78 (0.02)	-	-	−24.62 (0.000)
R Square	0.5	0.6	0.5	0.7
RMSE	107.04	128.31	31.93	47.608
MEF	0.71	0.60	0.72	0.56

Note, ( ) represents p-value.

- Population Density and Average Driver Age are negatively associated with traffic death in middle and high-income countries.
- Alcohol Consumption contributes to higher traffic death in low and middle-income countries.
- The Human Development Index has displayed conflicting influence in middle and low-income countries.
- Nonetheless, the predictability of these models is fairly weak because of moderate values R-squares and large RMSE values. They are not the ideal models.

Cluster based Models

The various attributes of the cluster-based models are presented in Table 5

(TRDMUPOP: Dependent variable).

4.2. Various Elements of Cluster Based Models

- Income impacts the Traffic Death Rate in the Low clusters only, showing how the rate increases as income increases for these two groups of countries.
- A higher proportion of the urban population is beneficial to all four clusters in lowering traffic death rates, especially in the Cluster Middle.
- Alcohol consumption significantly results in higher Traffic Death Rate for the Low and Middle clusters.
- Social and human development indexes are negatively associated with traffic death except in Cluster Low.
- A higher number of registered vehicles per million of the population are positively associated with Traffic Death Rates.

Some common observations between these two sets of models (Income based and Cluster based):

- The proportion of the urban population appeared to be a significant variable in both sets of models with a negative impact.
- The number of registered vehicles per million people appeared as a significant variable in Cluster Low and Cluster Middle, as well as in low and middle income countries
- The human development index emerged as a significant determinant in Cluster High and Cluster Middle-High, as well as Middle Income and High Income models. However, the Middle Income model shows a positive trend opposite of the others.
- Both sets of data identified that alcohol consumption as significantly positively related to Traffic Death Rates.

4.3. Statistical Attributes of Models

Various statistical attributes of the models are presented in Table 6.

These statistical attributes indicate that Cluster Middle-High and Cluster High models performed similar to the corresponding ones in the income-based models, but showed differences in low-income countries. This may be due to the relatively small sample size of data points used for the low-income countries. This may not be sufficient to distinguish their predictivities apart by just comparing the above statistics of these two sets of models. Thus, the models are tested using the latest available global traffic death data of 2016 to identify the ones with less bias. In this effort, the Line of Equity (LOE) statistical approach was employed.

5. Testing of the Derived Models against a New Set of Data to Identify the Less Biased Models

In order to determine the models with minimal bias, this study used goodness-of-fit statistics to evaluate each predictive model's performance, which involved comparing the actual and predicted values with the Line of Equality

Table 6. Statistics of developed models.

Income-based models	Low Income	Middle Income	High Income
R^2	0.8	0.5	0.6
Adjusted R^2	0.73	0.44	0.59
RMSE	96.96	113.87	42.31
MEF	0.49	0.73	0.63
N	30	165	112
Cluster-based models	**Low**	**Middle**	**High**	**Middle-High**
R^2	0.6	0.5	0.5	0.7
Adjusted R^2	0.60	0.47	0.45	0.65
RMSE	128.31	107.04	31.93	47.60
MEF	0.60	0.71	0.72	0.56
N	45	120	73	59

Table 7. Modified criteria for goodness-of-fit for LOE (Transportation Research Board, 2001).

Biasness	|1-slope|	|Intercept|	Goodness of fit	R^2	S_e/S_y
Unbiased	≤0.5	≤5	Excellent	≥0.9	≤ 0.35
Neutral	0.6 - 1	6 - 10	Good	0.70 - 0.89	0.36 - 0.55
			Fair	0.40 - 0.69	0.56 - 0.75
Biased	≥1	≥10	Poor	0.20 - 0.39	0.76 - 0.90
			Very poor	≤0.19	≥0.90

(*LOE*) proposed by NCHRP Report #465 (2001) (Transportation Research Board, 2001; Taha et al., 2015). First, a linear regression line between the predicted values and actual values called the Line of Equity (LOE) was created. Ideally, if the predicted values coincide with the actual values, then the slope should be 1, the intercept should be 0, and all points should lie in LOE. Therefore, the evaluation of the models can be done by evaluating the LOE with two parts, according to the criteria in **Table 7**: first to check the slope and intercept of LOE; second to check the goodness of points fitted to LOE. The first three columns address the biasness of the model as defined by the authors, and the last three columns assess the precision of the model (Transportation Research Board, 2001).

The remaining goodness of-fit statistical parameters were calculated using Equation (1) through Equation (4).

$$S_e = \sqrt{\frac{\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right)^2}{n-p}} \tag{1}$$

$$S_y = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - \bar{y}_i\right)^2}{n-1}} \tag{2}$$

$$R^2 = 1 - \frac{n-p}{n-1}\left(\frac{S_e}{S_y}\right)^2 \tag{3}$$

$$R^2_{adj} = 1 - \frac{n-1}{n-p}\left(1-R^2\right) = 1 - \left(\frac{S_e}{S_y}\right)^2 \tag{4}$$

where, S_e = standard error of estimate; S_y = standard deviation; y_i = actual value; $\hat{y}_i$ = predicted value; $\overline{y}_i$ = average of y_i; n = number of data points; p = number of regression coefficients, R^2 = coefficient of determination; and R^2_{adj} = adjusted coefficient of determination.

When 2016 data of traffic death and other factors were obtained, they were substituted to the models mentioned above. The predicted values of Traffic death rates and the corresponding actual death rates were plotted and LOE were created. The statistics of these LOEs for each model are presented in Table 8. Based on these statistics, the cluster models seem to be less biased and have better fit than the income-based models in general.

The conclusion that the cluster models are less biased and have a better fit than income-based models in general can be further illustrated visibly by Figure 2 and Figure 3.

Figure 2 depicts the Cluster-based goodness fit and biasness. For Cluster Middle, the data are not tightly fitted to the LOE which can also be seen by the moderate value of R^2 = 0.68. This model tends to underestimate the higher values of traffic death rates. For Cluster Low, the model shows excellence of fitting to the LOE, but overestimates the traffic death rates. The other two models perform similarly. It is noted that the overall traffic death rates for the countries in Cluster High are significantly lower than that the other groups.

Table 8. Statistics of LOE for models using 2016 data.

	Cluster based models				Income based models		
Model	**Middle**	**Low**	**High**	**Middle-High**	**Low**	**Middle**	**High**
|1-slope|	0.339	0.56	0.04	0.46	1.1	0.34	0.27
Intercept	10	−4.59	1.11	6.63	18.92	15.78	5.74
Biasness	Neutral	Unbiased	Unbiased	Neutral	Biased	Biased	Neutral
R^2	0.69	0.93	0.83	0.86	0.69	0.38	0.59
S_e/S_y	0.57	0.27	0.42	0.39	0.64	0.79	0.64
n*	31	13	16	16	10	50	28
Goodness of Fit	Fair	Excellent	Good	Good	Fair	Poor	Fair

*It is to be noted that the variable Edu_Index was a significant variable in Cluster High models, but did not appear in any Income-based models. For the validation data Edu-Index information was available for fewer countries. Therefore, the total sample size of Income-based models was larger than that of the cluster-based models.

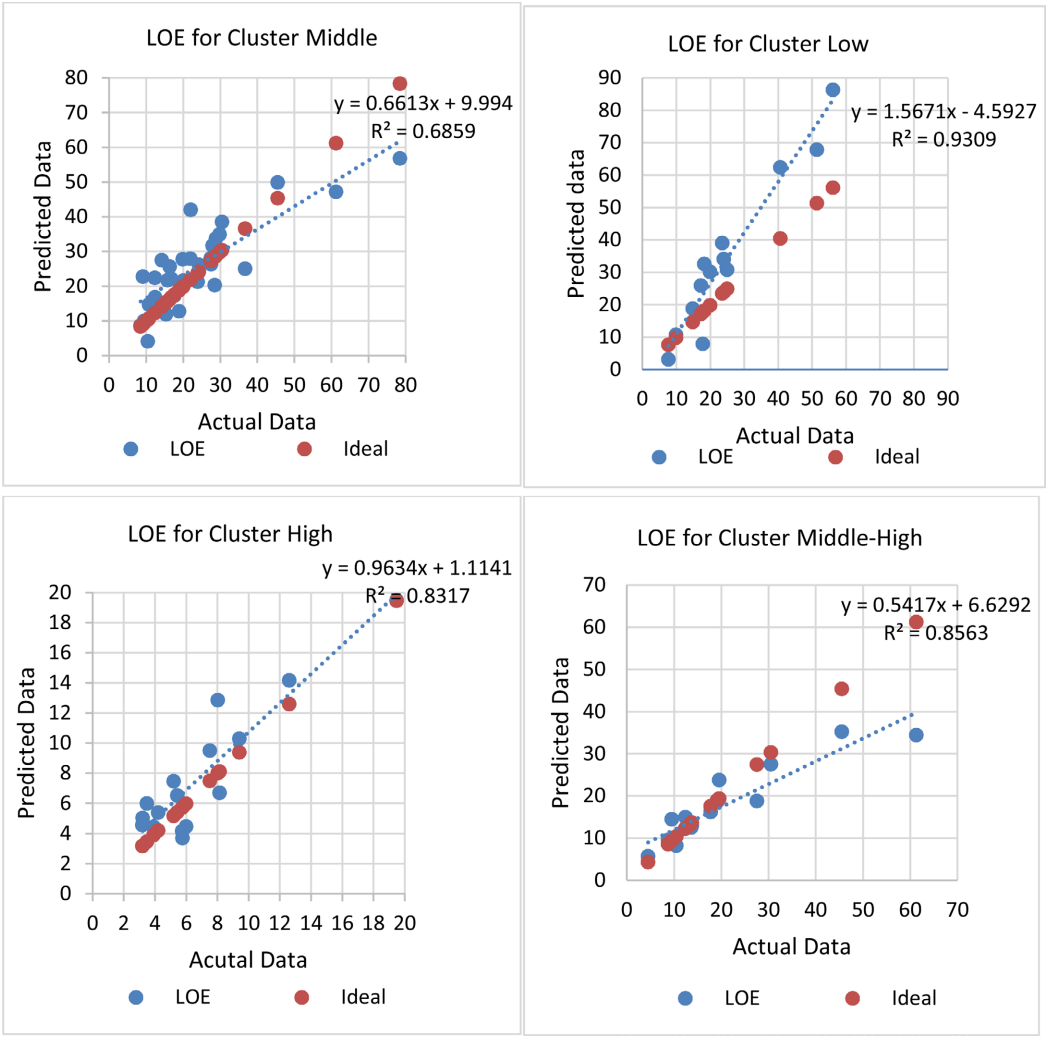

Figure 2. Actual and predicted value comparison Cluster-based approach validation 2016 data.

Figure 3 depicts the Income-based goodness fit and biasness. All of the models only have poor or fair fits to LOE with maximum R^2 of 0.63. It can be seen that for the Low-income countries and High-income countries, the predicted values are all about twice as high as the actual values. The models for Middle income countries have a very poor fit with predicted values scattered everywhere.

6. Findings of This Study

The degree of biasness of Income based and Cluster-based models were investigated using the LOE approach with a new set of data. It has been observed that:

- Cluster-based models are outperforming income-based models in terms of their accuracy in predicting the outcome of interest. Specifically, when comparing the predicted values from both types of models to the actual (observed) values, the cluster-based models display closer proximity to the original data than the income-based models.
- This suggests that the cluster-based models may be more reliable and efficient

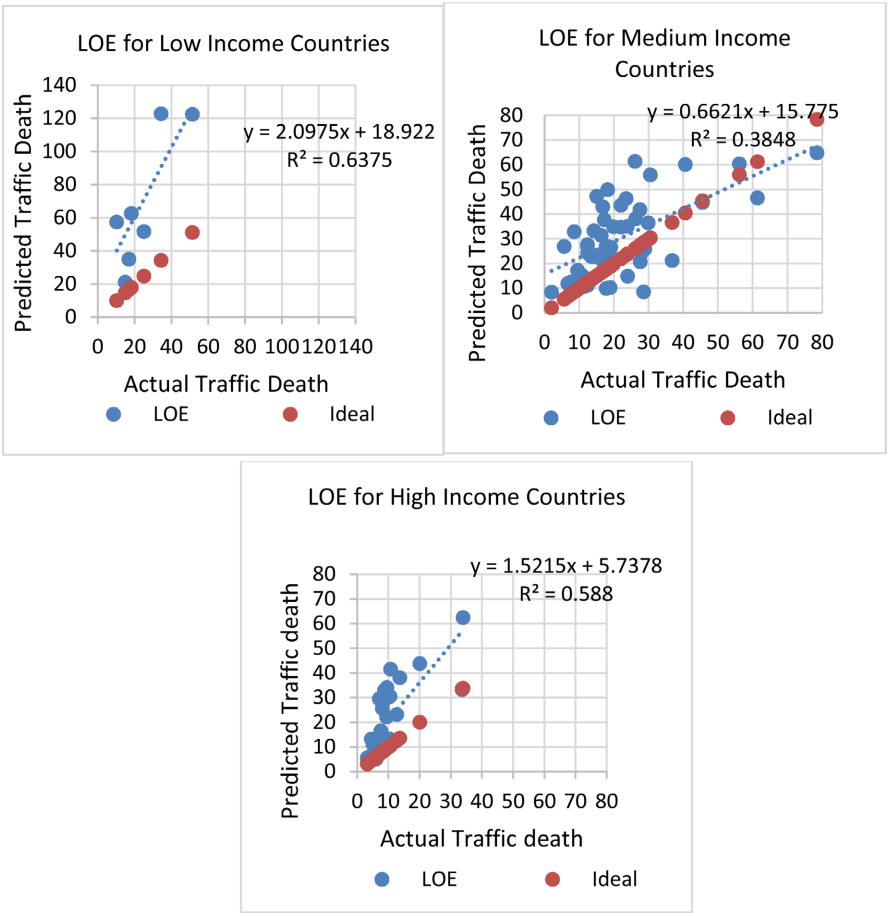

Figure 3. Actual and predicted value comparison income-based approach validation 2016 data.

than the income-based models in predicting the outcome.

- It is important to note that the reliability and efficiency of a model can depend on various factors, such as the quality and quantity of the data used to train the model, the complexity of the model, and the assumptions underlying the model
- Overall, it is encouraging to see that the cluster-based models are performing well on the latest data, which suggests that they may be a useful tool for predicting the outcome of interest in similar contexts.
- However, it's important to continue to evaluate and refine these models as new data becomes available, to ensure their ongoing reliability and accuracy.

7. Conclusion

In this paper, two sets of previously developed models were tested against a new set of data to identify the less biased models. A goodness-of-fit statistic was used to evaluate the performance of each predictive model in this study. As a part of this approach, a new set of data (data set from 2016) was used. The latest data indicates that cluster-based models outperform income-based models in terms

of accuracy. Based on the latest data, it appears that cluster-based models are outperforming income-based models in terms of accuracy. The cluster-based models are performing well on the latest data, suggesting that they may be a useful tool for predicting the outcome of interest in similar contexts. Based on the latest data, cluster-based models appear to be quite effective at predicting similar outcomes. However, when a new set of data is available, further examination of these models is warranted.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

Aggarwal, C., & Reddy, C. K. (Eds.) (2013). *Data Clustering*. Chapman & Hall/CRC. https://doi.org/10.1201/b15410

Dutta, U., Zhong, X. H., & Gsouda, N. (2022). Analysis of Global Road Traffic Death Data Using a Clustering Approach. *Current Urban Studies, 10,* 275-292. https://doi.org/10.4236/cus.2022.102017

Economic Times (2019). Definition of Human Development Index | What Is Human Development Index? Human Development Index Meaning—The Economic Times. *The Economic Times.* https://economictimes.indiatimes.com/definition/human-development-index

Grubesic, T. H., & Murray, A. T. (2001). Detecting Hot Spots Using Cluster Analysis and GIS. In *Proceedings from the 5th Annual International Crime Mapping Research Conference* (Vol. 26). Springer.

Hastie, T. et al. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Vol. 2). Springer.

IGI Global (2018). *What Is Education Index*. https://www.igi-global.com/dictionary/education-index/79488

KOF Swiss Economic Institute (2019). *KOF Globalisation Index*. https://kof.ethz.ch/en/forecasts-and-indicators/indicators/kof-globalisation-index.html

Kumar, S., & Toshniwal, D. (2015). A Data Mining Framework to Analyze Road Accident Data. *Journal of Big Data, 2,* Article No. 26. https://doi.org/10.1186/s40537-015-0035-y

Nouri, G. (2023). *Modeling of Global Traffic Deaths Using a Data-Driven Approach*. Ph.D. Dissertation, University of Detroit Mercy.

Taha, S. et al. (2015). Modeling of California Bearing Ratio Using Basic Engineering Properties. In *8th International Engineering Conference*. https://www.academia.edu/28150821/Modeling_of_California_Bearing_Ratio_using_Basic_Engineering_Properties

Transportation Research Board (2001). *Simple Performance Test for Superpave Mix Design*. NCHRP Report #465.

UNESCO (2015). *Education Index*. United Nations Economic and Social Commission for Western Asia. https://archive.unescwa.org/education-index

United Nations (UN) (2022). Human Development Index. https://hdr.undp.org/content/human-development-report-2021-22

US Census (2024). *International Database*. https://www.census.gov/programs-surveys/international-programs/about/idb.html

WHO (2018). *Global Status Report on Road Safety 2018*. World Health Organization.

WHO (2022). *Road Safety*. World Health Organization. https://www.who.int/health-topics/road-safety#tab=tab_1

World Bank (2022). *Glossary | DataBank*. https://databank.worldbank.org/metadataglossary/all/series