Accident Analysis and Prevention 130 (2019) 38–53

Contents lists available at ScienceDirect

Accident Analysis and Prevention

journal homepage: www.elsevier.com/locate/aap

Exploring the temporal stability of global road safety statistics

Loukas Dimitriou[a,*], Paraskevas Nikolaou[a], Constantinos Antoniou[b]

[a] Laboratory for Transport Engineering, Department of Civil and Environmental Engineering, University of Cyprus, 75 Kallipoleos Str., P.O. Box 20537, 1678 Nicosia, Cyprus

[b] Chair of Transportation Systems Engineering, Technical University of Munich, Arcisstrasse 21, 80333 Munich, Germany

ARTICLE INFO

Keywords:
Road traffic fatalities
Data inconsistencies
Data visualization
Principal component analysis
Hierarchical clustering
Structural equation modeling

ABSTRACT

Given the importance of rigorous quantitative reasoning in supporting national, regional or global road safety policies, data quality, reliability, and stability are of the upmost importance. This study focuses on macroscopic properties of road safety statistics and the temporal stability of these statistics at a global level. A thorough investigation of two years of measurements was conducted to identify any unexpected gaps that could highlight the existence of inconsistent measurements. The database used in this research includes 121 member countries of the United Nation (UN-121) with a population of at least one million (smaller country data shows higher instability) and includes road safety and socioeconomic variables collected from a number of international databases (e.g. WHO and World Bank) for the years 2010 and 2013. For the fulfillment of the earlier stated goal, a number of data visualization and exploratory analyses (Hierarchical Clustering and Principal Component Analysis) were conducted. Furthermore, in order to provide a richer analysis of the data, we developed and compared the specification of a number of Structural Equation Models for the years 2010 and 2013. Different scenarios have been developed, with different endogenous variables (indicators of mortality rate and fatality risk) and structural forms. The findings of the current research indicate inconsistency phenomena in global statistics of different instances/years. Finally, the results of this research provide evidence on the importance of careful and systematic data collection for developing advanced statistical and econometric techniques and furthermore for developing road safety policies.

1. Introduction

Road safety research, like many research fields, relies on painstakingly collected data from various sources. Many studies have dealt with various deficiencies and limitations of road safety data, such as underreporting of road accidents. Thus, in order to study road safety phenomena such as road traffic fatalities, collecting data from only one source for a wide range of countries is often impossible. However, collecting data from several sources can lead to errors (Hellerstein, 2008). In many cases, these anomalies are difficult to identify, especially when the data are used for extrapolation. This effects the creation of robust models in terms of Goodness-of-Fit (GoF), and in terms of meaning, and can be hazardous, especially when used for policy/decision making.

The objective of this research is to provide and demonstrate an efficient approach on investigating (not measuring) the magnitude of data inconsistencies, using an extensive historical global database by implementing data-model analysis, such as data visualization, Principal Component Analysis (PCA), Hierarchical Clustering (HC) and Structural Equation Modeling (SEM). Individual models are developed for two time-points: 2010 and 2013. In this way, any possible gaps between the 2010 and 2013 HCs, PCAs and SEM models might indicate the existence of inconsistent information, a fact that should be taken under consideration in studies supporting policy making. In particular, for the purposes of this research, 25 socio-economic variables regarding UN-121 countries for two time-points (2010 and 2013) were collected. The 25 socio-economic variables were further subdivided into four sectors (Economy, Network, Demographic/Geographic and Enforcement). These variables were used for studying the effects they might have on road traffic fatalities and in particular on mortality rate and fatality risk indices. Following this arduous data collection effort, the data have been thoroughly analyzed through the use of multiple suitable statistical techniques.

On one hand, data exploration procedure did not provide any evidence for possible existence of data inconsistencies, except from the case of the enforcement variables which in the correlation graphs showed some differences. On the other hand, multivariate exploratory techniques (PCA and HC) showed differences between the two instance

* Corresponding author.
E-mail addresses: lucdimit@ucy.ac.cy (L. Dimitriou), nikolaou.paraskevas@ucy.ac.cy (P. Nikolaou), c.antoniou@tum.de (C. Antoniou).

https://doi.org/10.1016/j.aap.2017.12.015
Received 5 July 2017; Received in revised form 4 December 2017; Accepted 17 December 2017
Available online 21 February 2018
0001-4575/

years' measurements. Thus, the next approach was to conduct a thorough methodological investigation on the effects that these data might have on mortality rate (Scenario 1) and on fatality risk (Scenario 2). The most popular models used in analyzing traffic fatalities, namely, linear regression models, cannot handle possible latent information involved in analyzing such complex phenomena. In order to consider such data and information relationships SEM models were adopted here. Therefore, for achieving the scope of the current paper SEM models for two different instances/years (2010 and 2013) were compared and unacceptable differences between the models were identified, a fact that indicates possible inconsistencies in global statistics. The only SEM model that showed consistency (referring to the coefficients of the variables) is the SEM Model 1 – Scenario 1. Overall, the findings of the current research exhibited that the gaps of the PCAs, HC graphs, and SEM coefficients indicate and highlight the existence of inconsistencies in the different in instance years' global data. Finally, the idea offered here for future work is the use of statistics (if they are directly available) from national authorities and checked again for consistency.

The rest of this paper is organized as follows. Section 2 outlines relevant research, while Section 3 presents the data collection procedure, followed by the data visualization and the multivariate exploratory analysis of the data. Section 4 describes the methodology for the SEM development and the comparison of the models. Section 5 provides a discussion based on the model estimation results. Finally relevant conclusions, remarks and highlights on possible further research directions are offered in the last section.

2. Background and review of the state-of-the-art

Road safety analyses are often based on macroscopic, aggregate data (Dupont et al., 2014; Yannis et al., 2014; Yu, 2015; Antoniou et al., 2016; Wegman et al., 2017). However, it is well-known that the use of such macro-level data, such as socio-economic and network-level, is accompanied by inherent risks, especially when their scope is towards policy/decision making.

Many researchers have implemented various methods for detecting and addressing data inconsistencies. Ma et al. (2009) developed a method for information inconsistencies detection for real-time information in dynamic decision-making. Fomina et al. (2014) presented some methods and approaches to deal with inconsistent and noisy databases used for the inductive notion formation. Deb and Liew (2016) developed a methodology for imputing missing data of numerical or categorical values in a traffic accident historical database.

Visual analysis (data visualization) is a common tool in identifying data issues in various disciplines. For instance, Baur et al. (2015) collected biomarkers of ageing analytical, anthropometric and demographic data from about 3000 volunteers in the MARK-AGE database and applied a data visualization method, among other methods, for dealing with errors in the database. However, data visualization cannot guaranty or quantitatively capture the identification of data inconsistencies, even if the phenomenon seems to be stationary over the years.

In order to identify patterns among the 2010 and 2013 samples and the relation that these samples might have, two appropriate multivariate exploratory techniques were considered, PCA and HC. Regarding the literature, several studies were conducted using these two techniques. For instance, Depaire et al. (2008) examined the effectiveness of cluster analysis as a technique for identifying homogenous traffic accident types and evaluated if it allows subsequent traffic accident analysis to reveal new information. Additionally, Saha et al. (2016) investigated the statistical plots of PCA for detecting anomalous sample data. PCA was considered to be a very promising technique in making a perceptible contribution to the detection of outliers from the observed data set and was accordingly applied on the proposed level-of service measures.

After identifying the patterns that these two samples have, and the differences between them, the next step was the estimation of the effects that the 2010 and 2013 macro-level socio-economic and demographic data might have on road traffic fatalities (mortality rate and fatality risk). However, these variables are considered to have a latent information structure that might have a consequent effect on road traffic fatalities estimation. Thus, a suitable method for considering this information is SEM. Moreover, SEM is widely used in several research fields, including economics (Tahmasebi and Rocca, 2015), health (Lai et al., 2015), and road safety (Hassan et al., 2013). "By segregating measurement errors from the true score of attributes, SEM provides a methodology to model the latent variables directly" (Yuan and Bentler, 2006).

Other road safety studies using SEM include Zhou et al. (2015), who analyzed pedestrians' self-reported violating crossing behavior intensions, and Hassan and Abdel-Aty (2011), who examined the responses of drivers under low visibility conditions and quantified the impacts and values of various factors found to be related to drivers' compliance and drivers' satisfaction with variable speed limit and changeable message signs instructions in different visibility, traffic conditions, and on two types of roadway.

Despite the differences of the different years SEM coefficients, a fact that indicates the existence of data inconsistencies, current research provides information on selecting SEM models with good fit. However, the complexity of SEM makes model selection harder than with simpler modeling approaches. Therefore, a lot of attention is given on the model Goodness-of-Fit (GoF) criteria and overall selection procedure. For example, Hooper et al. (2008) introduce a variety of fit indices and can be used as a guideline for prospective structural equation modelers. Preacher and Merkle (2012) discussed problems stemming from sampling variability in selection indices and show that selection decisions using information criteria (specifically the Bayesian information criterion) can be highly unstable over repeated sampling, even in large samples.

The current paper takes advantage of this experience and background and develops an elaborate data collection and modeling framework that allows the analysis of cross-sectional and temporal data for the identification and isolation of data issues. Ultimately, this study aims to contribute to the literature evidence of data inconsistencies in global statistics, using samples from two different instances/years and using straight forward procedures and several analyses (data visualization, PCA, and HC), followed by a model development approach, namely SEM. The findings of this research can be possibly used for selecting consistent macro-level variables/statistics (e.g. GNI, diesel price) for policy making purposes.

3. Data collection and exploratory analysis

3.1. Data coverage

Collecting global data poses significant challenges. Choices of variables can have considerable impact on the quality, completeness, and coverage of the final data. The 25 variables shown in Table 1 were selected for this analysis because they are reasonably reliable, available globally, and cover socio-economic, demographic, road network and road safety policy aspects. A number of sources were considered for the

Table 1
The collected (25) socio-economic variables.

Sector	Var. No	Abbreviation	Variable	Type
Economy	1	Income	Income level	Categorical[a]
	2	GNI	GNI per capita (US $)	Continuous
	3	GDP	GDP per capita (US $)	Continuous
	4	Food_prod	Food production index	Continuous
	5	Tax	Total tax rate (% of commercial profits)	Continuous
	6	Unemp	% Unemployment of total labor force	Continuous
	7	Diesel_price	Pump price for diesel fuel (US$ per liter)	Continuous
	8	Gasol_price	Pump price for gasoline (US$ per liter)	Continuous
	9	Int_users	Internet users (per 100 people)	Continuous
	10	Num_reg_veh	Number of registered vehicles (per 100,000 vehicles)	Continuous
Network	11	Tot_nodes_net	Total Nodes	Continuous
	12	Tot_length_net	Total Network's Length (Km)	Continuous
Demographic/Geographic	13	Con	Continents	Categorical[b]
	14	Popul	Population (per 1000,000 people)	Continuous
	15	Area	Land area (Km^2)	Continuous
	16	Popul_growth	Annual % population growth	Continuous
	17	Birth_rate	Birth rate, crude per 1000 people	Continuous
	18	Death_rate	Death rate, crude per 1000 people	Continuous
	19	Popul_15_64	Population aged 15-64	Continuous
Enforcement	20	Nat_speed_lim_law	National speed limit law	Categorical[c]
	21	Nat_drink_law	National drink–driving law	Categorical[c]
	22	Nat_helmet_law	National motorcycle helmet law	Categorical[c]
	23	Nat_seat_belt_law	National seat-belt law	Categorical[c]
	24	Nat_child_rest_law	National child restraint law	Categorical[c]
	25	Nat_mob_use_law	National law on mobile phone use while driving	Categorical[c]

Notes:
[a] 0: Low, 1: Middle, 2: High.
[b] 0: Asia, 1: Europe, 2: Africa, 3: North America, 4: South America, 5: Oceania.
[c] 0: Yes, 1: No.

data collection, and finally, data from the following sources were aggregated: World Health Organization (WHO), World Bank, Natural Earth and World Atlas. All data is from two different years: 2010 and 2013. These years were chosen because they represent the most recent year good global data coverage was available and a previous year which is close enough to be comparable and far enough to allow for changes in variables to become evident. The initial data collection effort considered all 193 UN countries, but gradually reduced this number to 121 by omitting countries with less than one million inhabitants or countries with missing information on several variables.

3.2. Data exploration

Scientific data visualization analyses data and calculation results in depth in order to obtain the understanding and insight on data, represented with images, graphs, maps etc. (Chaolong et al., 2016). Within this section, we focus the discussion on the 2010 and 2013 collected data. From a macroscopic point of view, the differences between these two samples were not directly evident from the data visualization process. The finer analyses in the next section highlight these aspects. Fig. 1, presents the main variables of population and reported road traffic fatalities of the UN-121 countries, regarding both years. Fig. 2, provides a similar view, this time based on registered vehicles and reported traffic fatalities in 2010 and 2013, respectively.

The visual comparisons of the socio-economic, demographic and traffic fatality measures of 2010 and 2013, showed some minor differences, which cannot be used for indicating any possible data inconsistency. Thus, in order to observe the data more precisely, correlation matrices were created for both samples and depicted in Fig. 3. In the analysis all variables are entered in natural units (as presented in Table 1). It is noted that due to the extended size of variables set only direct effects are considered in this research. From the correlation matrix of the 2010 and 2013 samples, it can be observed that road traffic fatalities are highly correlated (above 0.7 or below −0.7), with the variables: number of registered vehicles and population. This fact indicates that over the years road traffic fatalities are highly correlated with the same socio-economic and demographic factors. Additionally, the differences between the 2010 and 2013 samples, as they been identified from the correlation graphs, are the correlation that enforcement variables have between them.

3.3. Multivariate exploratory techniques

In order to identify the patterns that exist in the two samples, two multivariate exploratory techniques were used, PCA and HC. These techniques have the ability to view the relative contribution of variables in both datasets. Additionally, the use of these techniques offers preliminary indications of possible similarities/differences between the two datasets, identifying which justify further investigation. HC was first implemented by the use of the dissimilarity matrix, which was calculated according to Gower's distance. The dissimilarities between the countries are the weighted mean of the contribution of each data. The hierarchical clustering results are presented as a dendrogram (Fig. 4), which provides a succinct and eloquent visualization tool. The dendrograms were divided into four clusters by cutting the tree at about the 0.6 level; the clusters are depicted with the (red) borders. The height of the dendrograms refers to the dissimilarity values of the pairs.

The first HC of the UN-121 countries (Fig. 4, a) divided into 4 clusters, which included 22, 62, 3 and 34 countries, respectively. The dissimilarity value of the first cluster seems to be the smaller one than

the other three clusters. This might be due the fact that the majority of

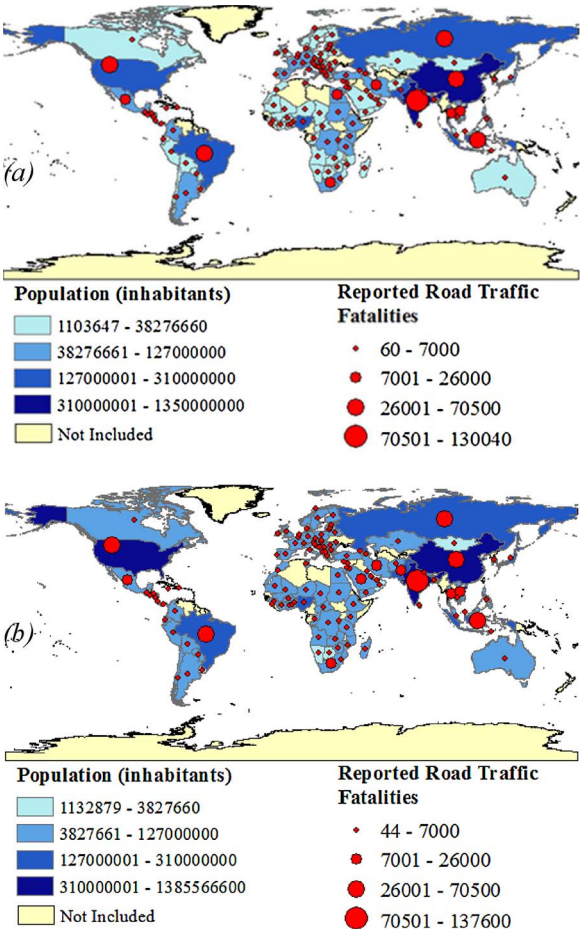

Fig. 1. Comparison between the road traffic fatalities and the population (a) in 2010 and; (b) in 2013.

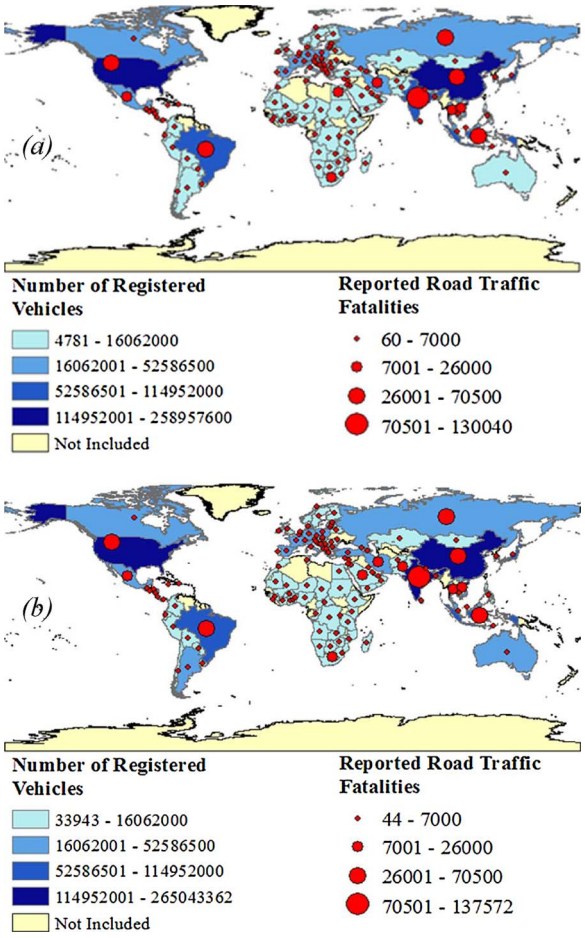

Fig. 2. Comparison between the road traffic fatalities and the number of registered vehicles (a) in 2010 and; (b) in 2013.

the countries included in the cluster are African (except for Afghanistan), which have almost the same socio-economic conditions. The third cluster includes only three countries (United States, Australia, and Canada), who seem that they approximately had the same socio-economic context in 2010. The second HC of the UN-121 countries regarding the 2013 data is depicted in Fig. 4, b. From this HC analysis four clusters emerged, including 24, 30, 30 and 37 countries, respectively. Looking at the assignment of the countries within the clusters, one can assess that the right most cluster includes the more advanced countries (from a road safety point of view), including most of the Western European countries (as well as several Arab countries). Moving towards the left of the figure, the road safety level seems to decrease. Naturally, this is not a straightforward and simple process and some observations are not clearly aligned. For example, the US is clustered in the second cluster (albeit at the border with the first one), along with Russia, China, and India. This cluster also includes Eastern European, as well as Central and South American countries. The third cluster includes primarily countries from Asia and North Africa, while the fourth one comprises countries from the rest of Africa and some less developed parts of Asia.

Principal Component Analysis has also been performed on the data, in order to extract a small number of vectors that can be used to summarize the data. PCA aims to explain the variance-covariance structure using a few linear combinations of the originally measured variables. Through this process, a more parsimonious description of the data is provided; reducing or explaining the variance of many variables with fewer well-chosen combinations of variables (Washington et al., 2011).

In particular, the current PCAs were developed by analyzing the correlation matrix, which takes the standardized form of the matrix. In the current study the two datasets (2010 and 2013) do not have the measurement scale, and thus analyzing the correlation matrix ensures that differences in measurement scales are accounted for (Field, 2009).

Fig. 5, visualizes the two PCAs for the UN-121 countries, regarding the 2010 and 2013 dataset. The eight first principal components, shown in Fig. 5, cumulatively explain more than 80% of the variance in the data. While the two first components seem to be very consistent across the two considered time-periods, the following components are increasingly more variable. This suggests that the main components might be stable across time, and thus more reliable for long term policy support. This can also be explained by looking at the main variables that contribute to these first components, which are more slowly-drifting variables, relating to the socioeconomic characteristics and infrastructure network measures.

4. Multiple year consistency analysis using SEM

4.1. Overview and variable transformation

Following up on the single-year analysis of the road safety data, we embark on a temporal analysis of the data stability. While a simple

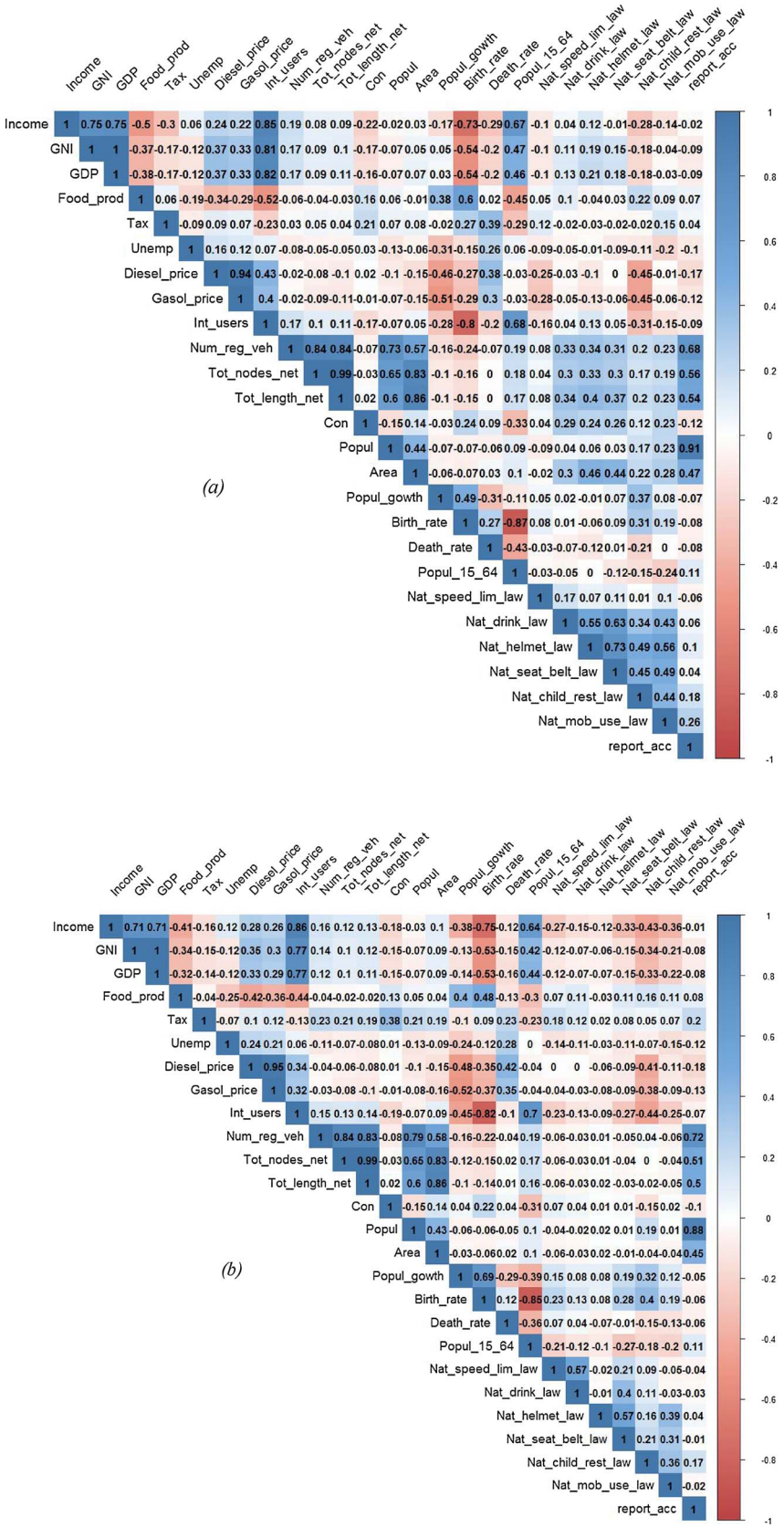

Fig. 3. Correlation graphs for the (a) 2010 and; (b) 2013 datasets.

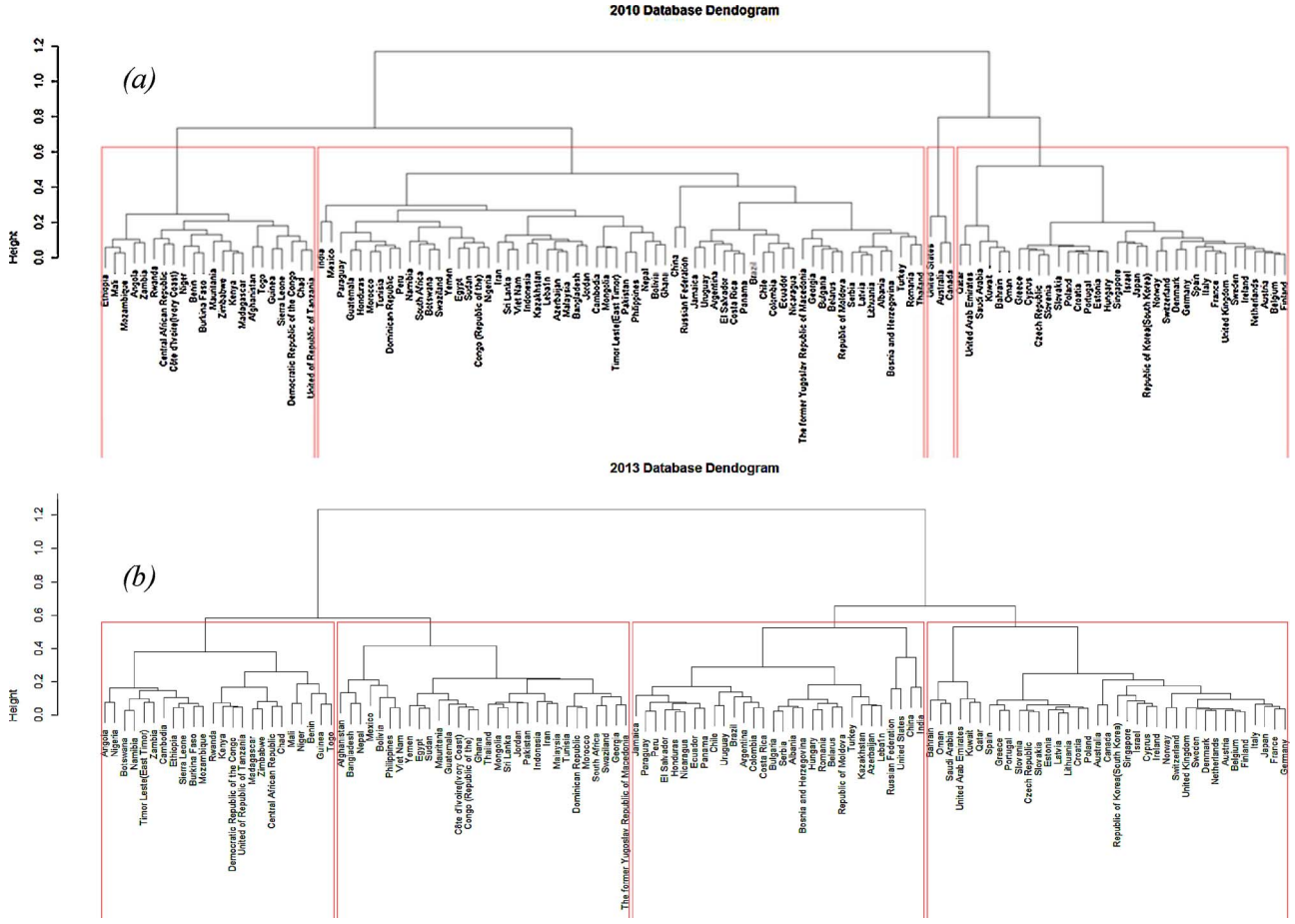

Fig. 4. The Hierarchical Clustering of the UN-121 countries according their socio-economic characteristics (a) in 2010 and; (b) in 2013. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

statistical analysis could provide some insights, we instead attempt to uncover the underlying data variability through the estimation of structural equation models for different time-points. Besides 2013, which – as discussed before – is the latest, complete year, we also consider 2010. The choice of the former year is indeed a bit arbitrary, and other years could also be chosen. The motivation for choosing this is that, on the one hand, it is recent enough, and therefore reasonable data can be collected, while, on the other hand, the difference of three years allows for some differences in the collected data to emerge. Table 2, presents the variable indices that were also considered for this analysis. Some have been obtained by a simple transformation of the original data variables; for example, mortality rate (road traffic fatalities per million inhabitants) and fatality risk (road traffic fatalities per 100,000 registered vehicles).

4.2. Experimental design

A number of models have been specified and estimated, in order to explore the stability of the data across the considered time-points (2010 and 2013). The dimensions of the experimental design are:

Year (two levels): (i) 2010 and (ii) 2013

Endogenous variable (two levels): (i) mortality rate or (ii) fatality risk, and

Considered explanatory sectors (four levels): the four considered sectors were considered in an additive fashion, i.e. the first type of models only included variables related to the sector “Economy”, the second to the sectors “Economy” and “Network” and so on. Thus, four combinations were considered (all possible combinations were not considered, as that would lead to a very tedious experimental setup, which could make interpretation of the results impractical).

A total of sixteen SEM models were thus developed (two years times two endogenous variables times four sets of explanatory variables to start the specification from). The models were then compared against each other via different metrics, in order to assess their degree of (dis) similarity. This comparative analysis includes aggregate model statistics and GoF measures, as well as path diagrams of the various models.

Prior to the development of the SEM models, an investigation for information inflation (e.g., multicollinearity) was performed through correlation matrices. After omitting the collinear variables from the models, the procedure continued with the SEM implementation. The following subsections outline this process.

4.3. Information inflation (multicollinearity) treatment

One of the common pitfalls of model development lurks in the existence of collinear variables. Socioeconomic variables are often susceptible to this issue, and therefore a step of explicitly considering this possibility preceded the modeling effort. In particular, correlation matrices were developed for all variables (at each time-point) and if any of the sample correlation values were highly positive (greater than 0.7) or

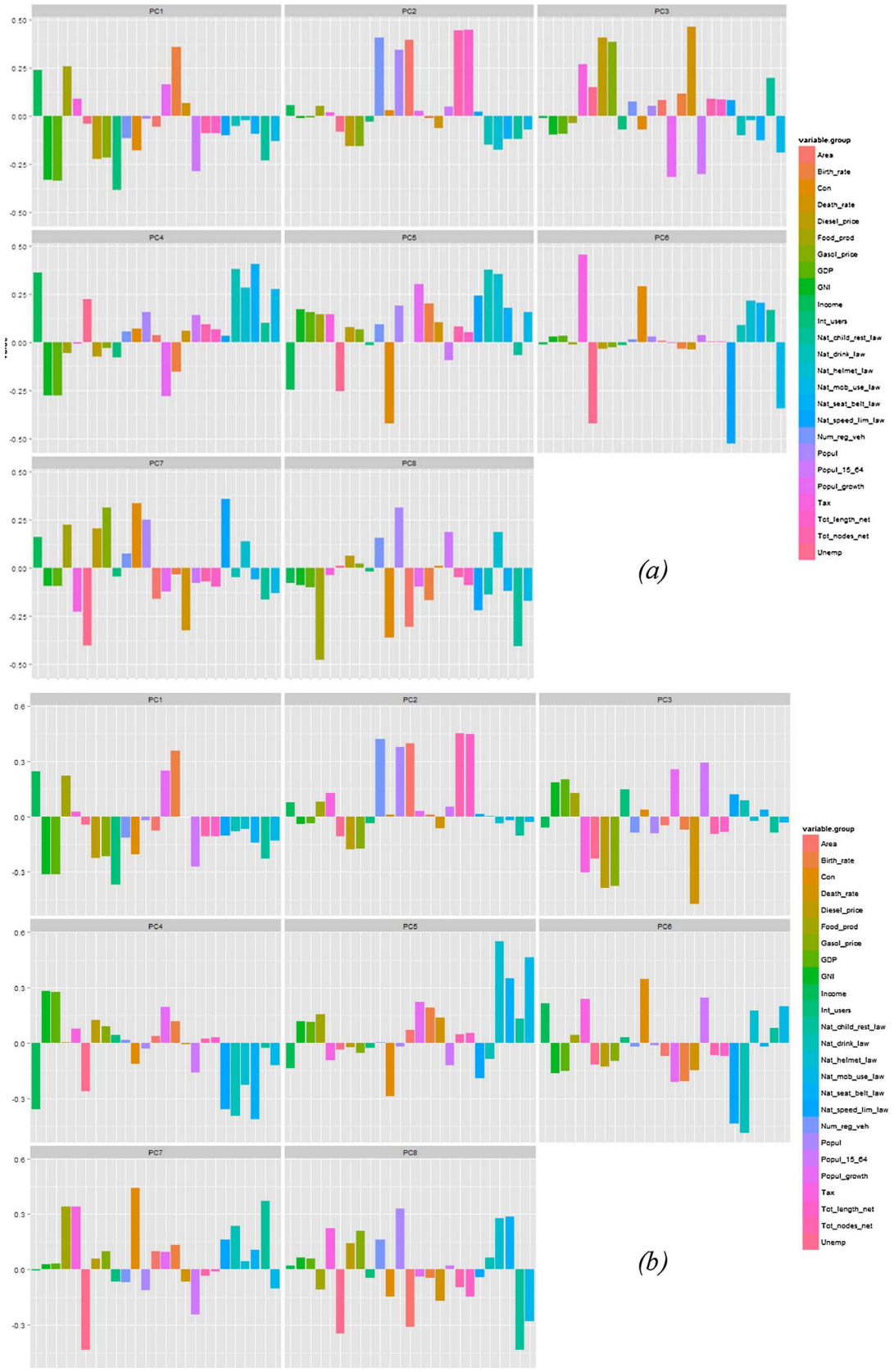

Fig. 5. Principal components for the socio-economic data (a) in 2010 and; (b) in 2013.

Table 2
The modified endogenous and exogenous variables.

Sector	Var. No	Abbreviation	Variable	Type
Network	1	Tot_nodes_net_veh	Total Nodes per 100 000 registered vehicles	Continuous
	2	Tot_length_net_veh	Total Network's Length (Km) per 100 000 registered vehicles	Continuous
Accidents	3	Report_acc_rates_popul	Reported road traffic fatalities per 1 000 000 people	Continuous
	4	Report_acc_rates_reg_veh	Reported road traffic fatalities per 100 000 registered vehicles	Continuous
Demo- graphic/ Geo- graphic	5	Popul_veh	Population per 100 000 registered vehicles	Continuous
	6	Area_pop	Land area (Km^2) per 1 000 000 people	Continuous
	7	Area_veh	Land area (Km^2) per 100 000 registered vehicles	Continuous
Eco- nomy	8	Num_reg_veh_pop	Registered vehicles per 1 000 000 people	Continuous

highly negative (smaller than -0.7) then there is probably evidence of multicollinearity. The multicollinearity problem can usually be eliminated by removing one of the exogenous variables in question from the set of exogenous variables (Bertsimas and Freund, 2004). The criterion for removing one variable from the high correlated pairs was the correlation of both variables with the endogenous variable (road traffic fatalities), i.e., the variable with the smallest correlation value with the endogenous variable was discarded. Binary variables were excluded from this procedure.

4.4. Structural equation modeling (SEM)

SEMs, similar to other statistical models, are used to evaluate theories or hypotheses using empirical data. All variables in the model, whether observed or latent, are classified as either exogenous or endogenous (Washington et al., 2011).

The SEM equation is:

$$\eta = \beta\eta + \gamma\xi + \varepsilon \quad (1)$$

where:

η: vector of endogenous variables.

ξ: vector of exogenous variables.

β and γ: coefficient vectors to be estimated (coefficients that contain regression coefficients for the endogenous and exogenous variables, respectively).

ε: vector of regression errors.

A backward stepwise regression analysis was implemented according to the regression weight tables that SEM software provided. Backward stepwise regression starts by comparing models with large numbers of exogenous variables and sequentially removing one exogenous variable at each step. The removed variable is the one that contributes least to the GoF criterion. The procedure iterates until a regression model is obtained in the final step (Washington et al. 2011).

Tables 3–6 are presented in Appendix B and show a summary view of the resulting SEM models for all scenarios concerning the 2010 and 2013 variables. In order to maintain an overview, the comparison of the models is not done at the level of individual variables, but at a more aggregate level, driven primarily by the GoF indicators. In particular, the following GoF indices [Akaike's Information Criterion (AIC), Schwarz Bayesian Information Criterion (BIC), Goodness-of-Fit Index (GFI) and Root Mean Square Error of Approximation (RMSEA)] are considered. The path diagrams of all models are presented in Appendix A (saturated model specification, followed by the final selected model for each year).

Concerning the Models' 2 final form, from Scenario 2 (Fig. 7 in Appendix A), it seems that the 2010 model includes some different variables compared to the respective 2013 model. The same applies for Models 1 (Fig. 6 in Appendix A) and 3 (Fig. 8 in Appendix A), in Scenario 1. As for Model's 4 final form, Scenario 1 (Fig. 9 in Appendix A), it can be observed that the 2013 model does not include the variables from the sector "Enforcement" and this is due either for collinearity reasons or for negative variance or for statistically insignificancy. However, the 2010 model appears to include almost all the variables of the sector "Enforcement". This particular difference is not expected, due the short time gap between the two years and thus can be justified due to information inconsistency. Overall, from these models (Scenario 1), the exogenous variables GNI, food production and diesel price appeared to be consistent and might be considered in future studies.

Regarding Models' 1 final form, from Scenario 2 (Fig. 10 in Appendix A), it seems that in the 2013 model the sector "Economy" is increasing the fatality risk indicator, in contrast with the 2010 model, where the particular sector is decreasing the fatality risk indicator. Once again, the particular differential between the concluded models can be interpreted as possible existence of data inconsistency. Continuing with Models' 4 final form, from Scenario 2 (Fig. 13 in Appendix A), it appears that in the 2010 model the sector 'Demographic/Geographic' tends to increase fatality risk. Moreover, in the respective 2013 model, only one variable for the factor 'Enforcements' was remained in the model ('Nat_dring_law'). However, observing the positive effect of this variable on fatality risk increment strengthens the speculations for data inconsistency existence inside the database. Furthermore, the sector 'Economy' in the 2010 model, seems to decrease the fatality risk. Concerning, the respective 2013 model, it appears that the sector 'Economy' is increasing the fatality risk indicator. Additionally, this model does not include the sector 'Demographic/Geographic' but only two variables of the sector which have a direct relationship with the endogenous variable.

From this particular implementation, it can be observed that when building separate models using different combinations of socio-economic factors, certain socio-economic factors are not statistically significant (e.g. "Network") regarding road traffic fatalities. The sector "Economy" appeared to be efficient and was included in all sixteen models. Moreover, it seems that the "Economy" sector is the most robust sector, taking into consideration the correlation that it has with mortality rate and fatality risk indicators. This revelation in consequent when related to European countries (e.g. Yannis et al., 2014; Antoniou et al., 2016), and it appears that these results are on global reference. Notwithstanding, the fact that the sector "Network" was added to all models except Models' 1 in every scenario, it appeared to be omitted in every model's final form due to collinearity reasons or negative variance reasons or due statistical insignificancy. This evidence leads to the conclusion that more comprehensive variables concerning the networks' infrastructure of the countries could be used on future works.

Notwithstanding, the fact that some of the data might be inconsistent and thus not reliable for investigating phenomena such as road traffic fatalities it does not nullify all the models and their potential for providing robust estimations for the particular indicators. For selecting the models which provided robust estimations, the GoF measures are taken under consideration.

5. Discussion

The GoF of a statistical model describes how well it fits into a set of observations. Two popular GoF indices are AIC and BIC. The AIC and BIC are not used to test the model in the sense of hypothesis testing, but for model selection. Of the two, the BIC penalizes by adding parameters to the model more strongly than the AIC (Maydeu-Olivares and Garcia-Forero, 2010).

The most popular alternative measures of fit for SEM analysis, however, are the GFI, the Normed Fit Index (NFI), the Comparative Fit Index (CFI), and the RMSEA. The GFI, NFI, and CFI all have values ranging from 0 to 1; a good fit is indicated by values greater than 0.90 for GFI and NFI and 0.95 and greater for CFI. For RMSEA, a value of 0 is interpreted as an exact fit; values less than 0.05 are a close fit, values between 0.05 and 0.08 are a fair fit, values between 0.08 and 0.10 are a mediocre fit, and values more than 0.10 are a poor fit (Chan et al., 2007).

In the current research, the following measures were used for model selection: GFI, RMSEA, AIC, and BIC. In the current experimental setup, as it concerns the 2010 models from the Scenario 1 (Table 3 in Appendix B), Model 1 has a better fit according to the AIC, GFI, and BIC, while Model 4 has a better fit than the other three models according to RMSEA. Thus, Model 1 is identified as the most significant in estimating the road traffic fatalities with reference to the population. Using the same procedure, the following models outperformed the others:

3 Models-Scenario 1: Model 1 (Fig. 6-a in Appendix A)
4 Models-Scenario 2: Model 1 (Fig. 10-a in Appendix A)
5 Models-Scenario 1: Model 1 (Fig. 6-b in Appendix A)
6 Models-Scenario 2: Model 1 (Fig. 10-b in Appendix A)

The information about the models is complete, especially the Figs. 6–13 in Appendix A which can be examined to try understand the "logics" of the data and their effects on road traffic fatalities. It appears the models with mortality rates (Scenario 1) perform poorly. One general observation is that while a single model did not dominate all scenarios, simpler models tended to perform better. This is not a surprising finding and is consistent with the expectation of parsimony. In this case, this is an encouraging finding, which suggests that in order to obtain a reasonable model specification for road safety, a reasonable dataset could suffice. Naturally, there are many ways to interpret this finding, and there needs to be a lot of caution in how this is phrased and interpreted. In fact, instead of arguing that more road safety data are not needed, the more rational approach is to use this finding to support the notion that *better* data are needed. Then, richer models would be able to be specified and estimated with much better results. Such models would arguably lead to better road safety forecasts and therefore better support for developing evidence-based road safety policy.

6. Conclusions

Global databases have been considered to be sources of information for several factors (e.g. economy and demography). However, these databases have been carried out by humans, a fact that makes the information imperfect to some cases. Thus, mistaken conclusions might occur when these data are used. The scope of the current paper is the investigation of two different instant years' measurements in order to identify any unexpected gaps that might be due the existence of inconsistent information.

For the purposes of this research UN-121 countries and information (from global organizations; e.g. WHO and World Bank) about their socio-economic, demographic/geographic, network infrastructure and enforcement context were considered for the years 2010 and 2013. This information was related with the road traffic fatalities of the countries. The investigation of the collected data was initially approached by the illustration of the data on both years (data visualization) and with the implementation of multivariate exploratory techniques (e.g. PCA and HC). Data visualization approach showed that their no significant inconsistency. However, HC and PCA techniques show some differences between the different years. In detail, the PCAs difference was detected from the third component.

In order to further explore and validate this gap, the 2010 and 2013 datasets were used for developing models that will quantitatively capture the effect that such data have on two different risk exposure indicators, namely mortality rate (Scenario 1) and fatality risk (Scenario 2). SEM is used here in order to capture primarily the latent information that these data might have. On the other hand, the measurement equation could be better handled by a count regression model. This would result in a treatment by a hybrid model. The estimated SEM models, regarding Scenario 1, showed that there is a strong difference between them (see Table 3–4 in Appendix B and Figs. 6–9 in Appendix A), which indicates a possible existence of data inconsistency. Additionally, it could be said that the socio-economic variables, GNI, food production and diesel price are included in both 2010 and 2013 models, a fact that suggests the use of these variables in related works. As for the SEM models, regarding Scenario 2 (see Tables 5–6 in Appendix B and Figs. 10–13 in Appendix A), it appears that the assumption for data inconsistency is valid. In particular, in all the 2010 models the effect of "Economy" on fatality risk is strongly negative, in contrast with the 2013 models where "Economy" appears with a strongly positive effect. This outcome might be justified due to the economic changes (e.g. economic crisis) that occurred in the period 2010–2013. However, the differences of the 2010 and 2013 models are 'strong' enough when the models become additively larger (see Model 4). Overall, the correlations between the latent variables are rather small and consequently, the differences between the two periods have to be minimized. Additionally, the differences between the two periods are more evident with the fatality risk. Also, the effects of some exogenous variables do not have the expected sign (e.g. "Nat_drink_law" in Model 4 – Scenario 2), a fact that might be justified as inconsistent information or it can be justified in the failure of one country to pass a law on drinking. As it was exposed in the literature review section, although these macro-level data are commonly used in several studies of road traffic fatalities, collecting data from several global organizations may introduce information inconsistency and thus biases to models. In order to avoid this phenomenon, in cases of realistic analysis it is recommended to cross-check the information from trustworthy national measure centers (e.g. Police). However, the majority of countries included in this research, do not have an accessible system for providing such information, and thus a global investigation would be a tedious task to be carried out.

The next research work is to develop a joint model with the two periods to test the differences. Furthermore, the focus will be given on the investigation of data inconsistencies inside national databases, likely Cyprus. Finally, the addition of stratification or decomposition according to road users would be interesting in future studies.

Appendix A

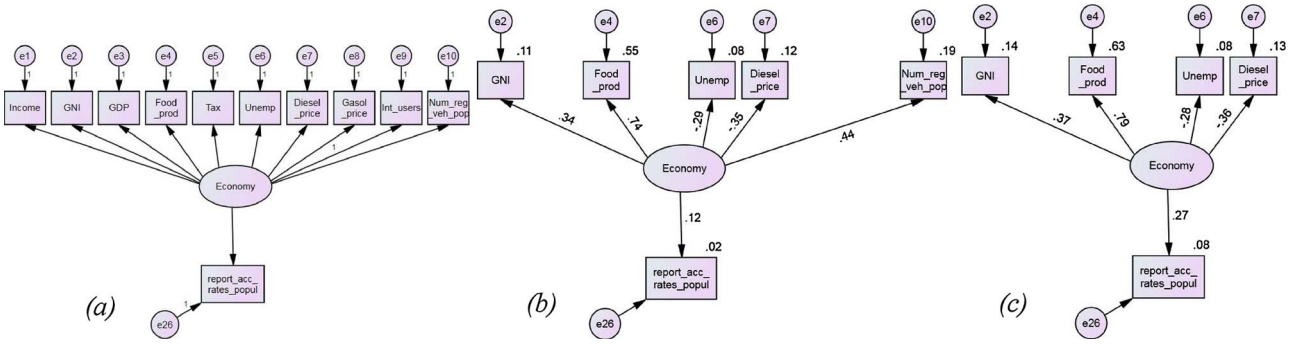

Fig. 6. Model 1-Scenario 1 (a) initial form for 2010 and 2013; (b) final form for 2010; (c) final form for 2013.

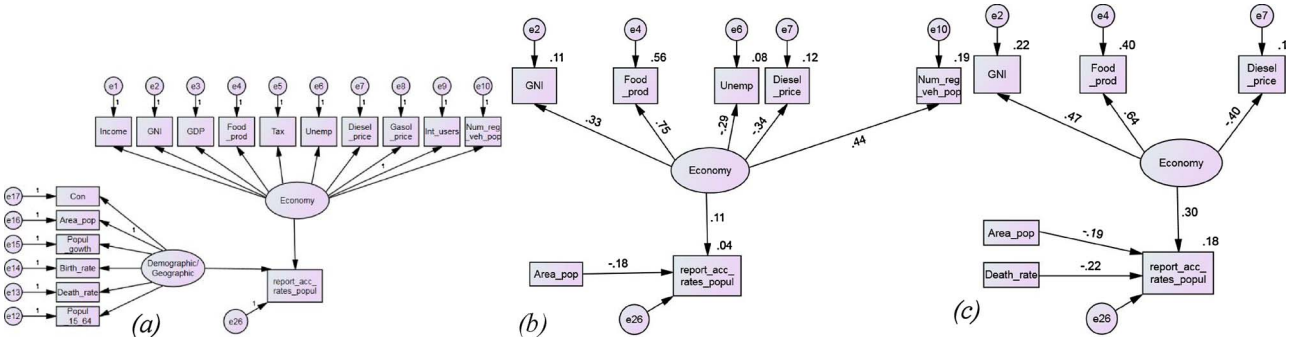

Fig. 7. Model 2-Scenario 1 (a) initial form for 2010 and 2013; (b) final form for 2010; (c) final form for 2013.

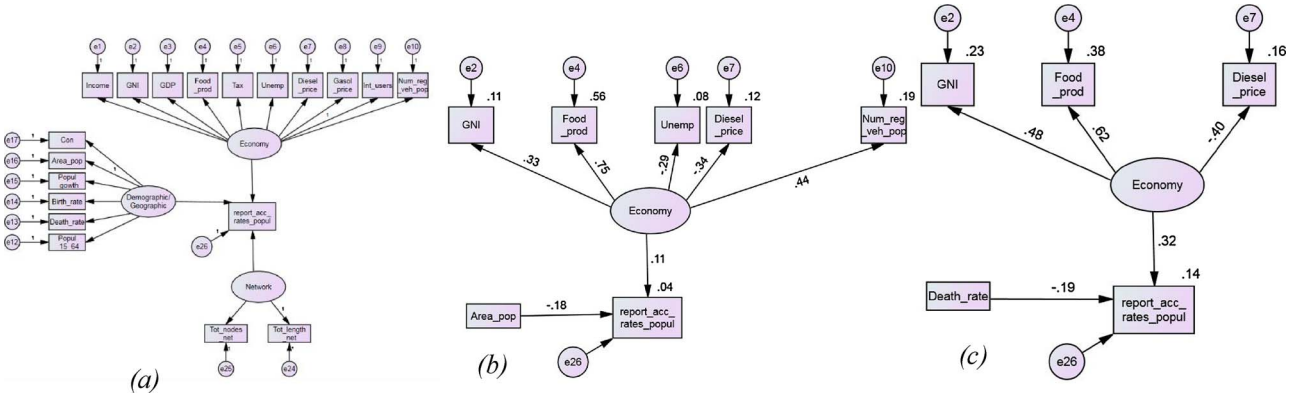

Fig. 8. Model 3-Scenario 1 (a) initial form for 2010 and 2013; (b) final form for 2010; (c) final form for 2013.

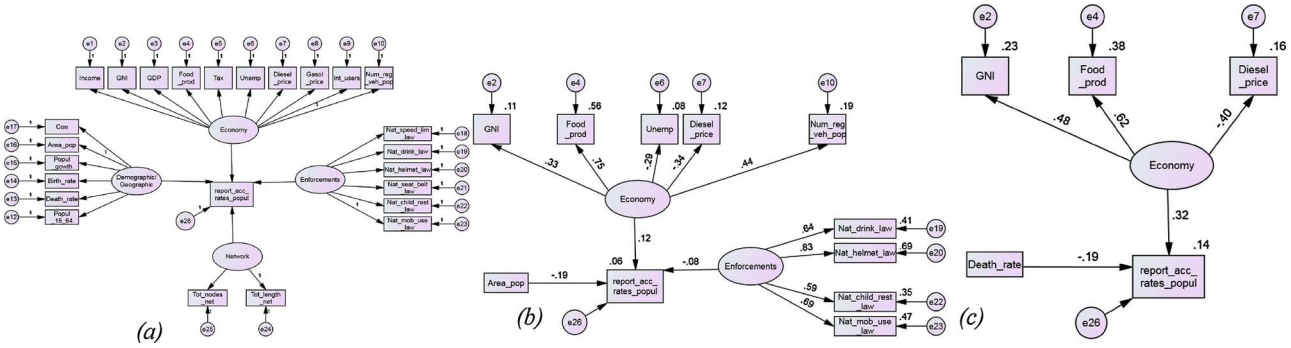

Fig. 9. Model 4-Scenario 1 (a) initial form for 2010 and 2013; (b) final form for 2010; (c) final form for 2013.

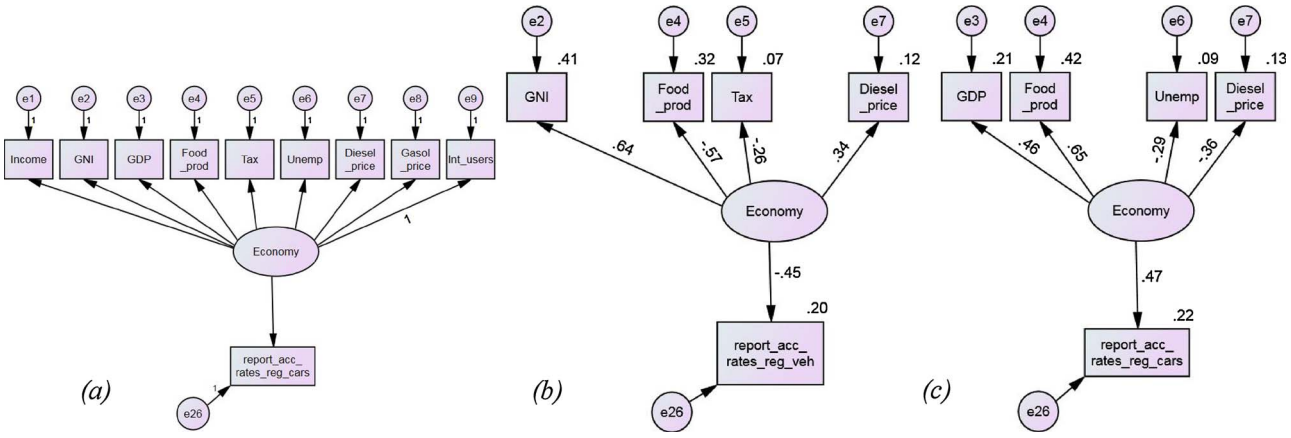

Fig. 10. Model 1-Scenario 2 (a) initial form for 2010 and 2013; (b) final form for 2010; (c) final form for 2013.

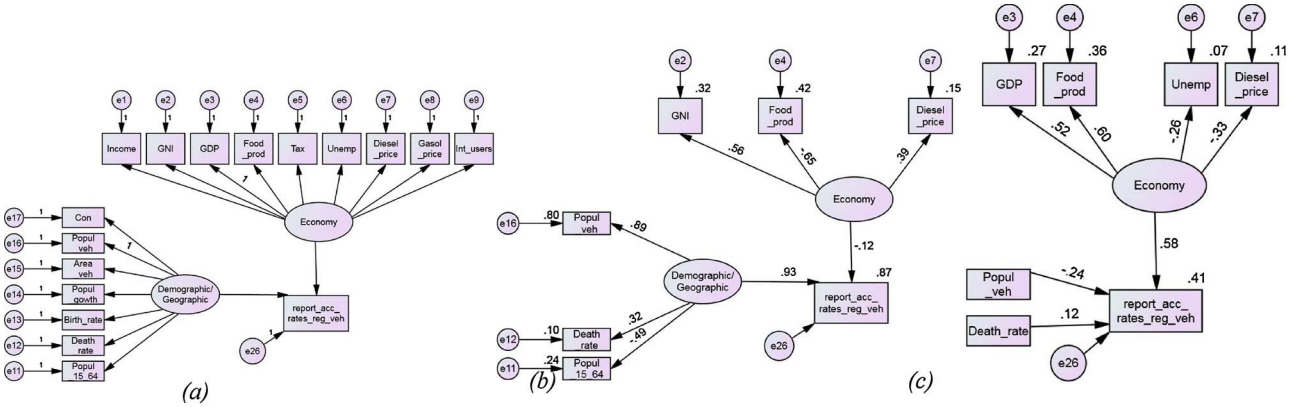

Fig. 11. Model 2-Scenario 2 (a) initial form for 2010 and 2013; (b) final form for 2010; (c) final form for 2013.

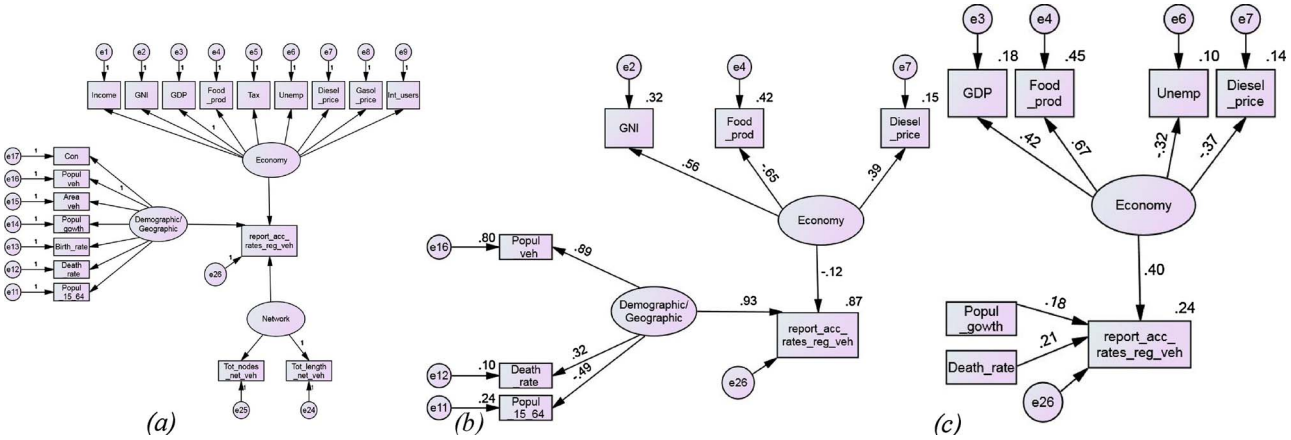

Fig. 12. Model 3-Scenario 2 (a) initial form for 2010 and 2013; (b) final form for 2010; (c) final form for 2013.

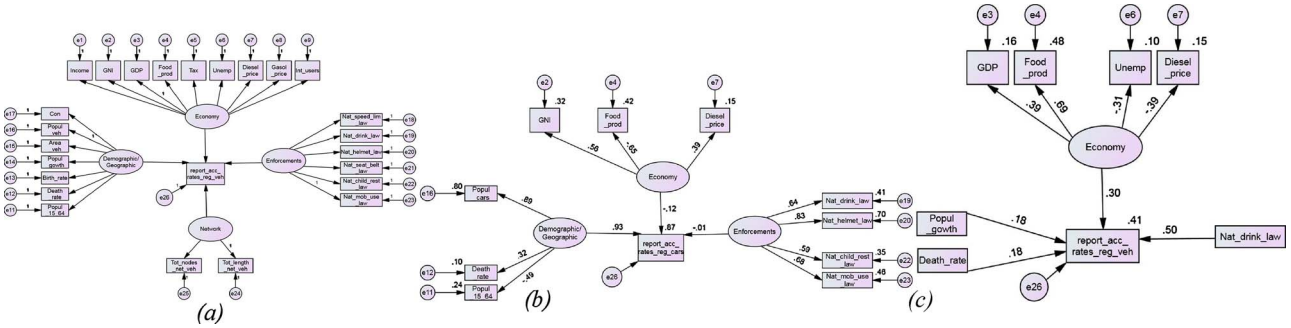

Fig. 13. Model 4-Scenario 2 (a) initial form for 2010 and 2013; (b) final form for 2010; (c) final form for 2013.

Appendix B

Table 3
Models' statistical information according 2010 data, Scenario 1.

	Model 1			Model 2			Model 3			Model 4		
	Standardized weight	Regression weight		Standardized weight	Regression weight		Standardized weight	Regression weight		Standardized weight	Regression weight	
		Estimate	SE		Estimate	SE		Estimate	SE		Estimate	SE
Income ⟵ Economy	[a]	[a]		[a]	[a]		[a]	[a]		[a]	[a]	
GNI ⟵ Economy	0.336	0.072 (**)	0.03	0.331	0.07 (*)	0.03	0.331	0.07 (*)	0.03	0.332	0.07 (*)	0.03
GDP ⟵ Economy	[a]	[a]		[a]	[a]		[a]	[a]		[a]	[a]	
Food_prod ⟵ Economy	0.741	1		0.746	1		0.746	1		0.748	1	
Tax ⟵ Economy	[c]	[c]		[c]	[c]		[c]	[c]		[c]	[c]	
Unemp ⟵ Economy	−0.289	−0.137 (*)	0.063	−0.291	−0.137 (*)	0.063	−0.291	−0.137 (*)	0.063	−0.29	−0.136 (*)	0.063
Diesel_price ⟵ Economy	−0.347	−0.02 (*)	0.008	−0.343	−0.019 (*)	0.008	−0.343	−0.019 (*)	0.008	−0.342	−0.019 (*)	0.008
Gasol_price ⟵ Economy	[a]	[a]		[a]	[a]		[a]	[a]		[a]	[a]	
Int_users ⟵ Economy	[a]	[a]		[a]	[a]		[a]	[a]		[a]	[a]	
Num_reg_veh_pop ⟵ Economy	0.438	0.093 (**)	0.034	0.438	0.092 (**)	0.034	0.438	0.092 (**)	0.034	0.436	0.091 (**)	0.034
Con ⟵ Demographic/ Geographic				[c]	[c]		[c]	[c]		[c]	[c]	
Populd												
Popul_growth ⟵ Demographic/ Geographic				[c]	[c]		[c]	[c]		[c]	[c]	
Birth_rate ⟵ Demographic/ Geographic				[a]	[a]		[a]	[a]		[a]	[a]	
Death_rate ⟵ Demographic/ Geographic				[b]	[b]		[b]	[b]		[b]	[b]	
Popul_15_64 ⟵ Demographic/ Geographic				[b]	[b]		[b]	[b]		[b]	[b]	
Tot_nodes_net ⟵ Network										[a]	[a]	
Tot_length_net ⟵ Network										[a]	[a]	
Nat_speed_lim_law ⟵ Enforcements										[c]	[c]	
Nat_drink_law ⟵ Enforcements										0.639	0.548 (***)	0.094
Nat_helmet_law ⟵ Enforcements										0.833	0.806 (***)	0.123
Nat_seat_belt_law ⟵ Enforcements										[a]	[a]	
Nat_child_rest_law ⟵ Enforcements										0.592	0.877 (***)	0.16
Nat_mob_use_law ⟵ Enforcements										0.685	1	
Report_acc_rates_popul ⟵ Economy	0.123	6.735 (0.292)	6.394	0.11	0.555 (0.335)	0.576	0.11	0.555 (0.335)	0.576	0.119	0.596 (0.299)	0.574
Report_acc_rates_popul ⟵ Area_pop				−0.178	−4.026 (*)	2.019	−0.178	−4.026 (*)	2.019	−0.19	−4.307 (*)	2.013
Report_acc_rates_popul ⟵ Enforcements										−0.083	−13.633 (0.399)	16.166
d.f.	9			14			14			53		
AIC	45.773			64.205			64.205			149.297		
BIC	79.323			103.346			103.346			213.6		
GFI	0.937			0.919			0.919			0.863		
RMSEA	0.109			0.115			0.115			0.108		

[a] Collinear Omission.
[b] Negative Variance Omission.
[c] Non-Statistical Significant Omission.
[d] The variable was omitted from the models' data sets and were used for creating the endogenous variables in the respective scenarios.
*Significant at the 0.05 level.
**Significant at the 0.01 level.
***Significant at the 0.001 level.

Table 4
Models' statistical information according 2013 data, Scenario 1.

	Model 1			Model 2			Model 3			Model 4		
	Standardized weight	Regression weight		Standardized weight	Regression weight		Standardized weight	Regression weight		Standardized weight	Regression weight	
		Estimate	SE		Estimate	SE		Estimate	SE		Estimate	SE
Income ⟵ Economy	[a]	[a]		[a]	[a]		[a]	[a]		[a]	[a]	
GNI ⟵ Economy	0.369	0.053 (*)	0.023	0.466	0.084 (*)	0.035	0.482	0.089 (*)	0.037	0.482	0.089 (*)	0.037
GDP ⟵ Economy	[a]	[a]		[a]	[a]		[a]	[a]		[a]	[a]	
Food_prod ⟵ Economy	0.793	1		0.635	1		0.619	1		0.619	1	
Tax ⟵ Economy	[c]	[c]		[c]	[c]		[c]	[c]		[c]	[c]	
Unemp ⟵ Economy	−0.281	−0.094 (*)	0.047	[c]	[c]		[c]	[c]		[c]	[c]	
Diesel_price ⟵ Economy	−0.358	−0.013 (*)	0.006	−0.397	−0.018 (*)	0.008	−0.395	−0.019 (*)	0.008	−0.395	−0.019 (*)	0.008
Gasol_price ⟵ Economy	[a]	[a]		[a]	[a]		[a]	[a]		[a]	[a]	
Int_users ⟵ Economy	[a]	[a]		[a]	[a]		[a]	[a]		[a]	[a]	
Num_reg_veh_pop ⟵ Economy	[c]	[c]		[c]	[c]		[c]	[c]		[c]	[c]	
Con ⟵ Demographic/ Geographic Populd				[c]	[c]		[c]	[c]		[c]	[c]	
Popul_growth ⟵ Demographic/ Geographic				[c]	[c]		[c]	[c]		[c]	[c]	
Birth_rate ⟵ Demographic/ Geographic				[a]	[a]		[a]	[a]		[a]	[a]	
Popul_15_64 ⟵ Demographic/ Geographic				[b]	[b]		[b]	[b]		[b]	[b]	
Area_pop ⟵ Demographic/ Geographic							[a]	[a]		[a]	[a]	
Tot_nodes_net ⟵ Network							[a]	[a]		[a]	[a]	
Tot_length_net ⟵ Network							[a]	[a]		[a]	[a]	
Nat_speed_lim_law ⟵ Enforcements										[c]	[c]	
Nat_drink_law ⟵ Enforcements										[c]	[c]	
Nat_helmet_law ⟵ Enforcements										[c]	[c]	
Nat_seat_belt_law ⟵ Enforcements										[b]	[b]	
Nat_child_rest_law ⟵ Enforcements										[c]	[c]	
Nat_mob_use_law ⟵ Enforcements										[c]	[c]	
Report_acc_rates_popul ⟵ Economy	0.275	0.841 (*)	0.427	0.303	1.157 (*)	0.561	0.322	1.255 (*)	0.591	0.322	1.255 (*)	0.591
Report_acc_rates_popul ⟵ Area_pop				−0.188	−0.016 (*)	0.007						
Report_acc_rates_popul ⟵ Death_rate				−0.224	−4.176 (**)	1.582	−0.191	−3.543 (*)	1.612	−0.191	−3.543 (*)	1.612
Report_acc_rates_popul ⟵ Network												
Report_acc_rates_popul ⟵ Enforcements												
d.f.	5			9			5			5		
AIC	25.049			35.077			25.858			25.858		
BIC	53.007			68.627			53.816			53.816		
GFI	0.983			0.969			0.98			0.98		
RMSEA	0.009			0.044			0.038			0.038		

[a] Collinear Omission.
[b] Negative Variance Omission.
[c] Non-Statistical Significant Omission.
[d] The variable was omitted from the models' data sets and were used for creating the endogenous variables in the respective scenarios.
*Significant at the 0.05 level.
**Significant at the 0.01 level.
***Significant at the 0.001 level.

Table 5
Models' statistical information according 2010 data, Scenario 2

	Model 1			Model 2			Model 3			Model 4		
	Standardized weight	Regression weight		Standardized weight	Regression weight		Standardized weight	Regression weight		Standardized weight	Regression weight	
		Estimate	SE		Estimate	SE		Estimate	SE		Estimate	SE
Income ⟵ Economy	[a]	[a]		[a]	[a]		[a]	[a]		[a]	[a]	
GNI ⟵ Economy	0.643	1		0.563	1		0.563	1		0.562	1	
GDP ⟵ Economy	[a]	[a]		[a]	[a]		[a]	[a]		[a]	[a]	
Food_prod ⟵ Economy	−0.566	−0.001 (**)	0	−0.647	−0.001 (*)	0	−0.647	−0.001 (*)	0	−0.648	−0.001 (*)	0
Tax ⟵ Economy	−0.256	−0.001 (*)	0	[c]	[c]		[c]	[c]		[c]	[c]	
Unemp ⟵ Economy	[c]	[c]		[c]	[c]		[c]	[c]		[c]	[c]	
Diesel_price ⟵ Economy	0.341	0 (*)	0	0.387	0 (**)	0	0.387	0 (**)	0	0.387	0 (**)	0
Gasol_price ⟵ Economy	[a]	[a]		[a]	[a]		[a]	[a]		[a]	[a]	
Int_users ⟵ Economy	[a]	[a]		[a]	[a]		[a]	[a]		[a]	[a]	
Num_reg_veh_popd												
Con ⟵ Demographic/ Geographic				[c]	[c]		[c]	[c]		[c]	[c]	
Popul_veh				0.895	1		0.895	1		0.895	1	
Popul_growth ⟵ Demographic/ Geographic				[c]	[c]		[c]	[c]		[c]	[c]	
Birth_rate ⟵ Demographic/ Geographic				[a]	[a]		[a]	[a]		[a]	[a]	
Death_rate ⟵ Demographic/ Geographic				0.318	0 (***)	0	0.318	0 (***)	0	0.318	0 (***)	0
Popul_15_64 ⟵ Demographic/ Geographic				−0.486	0 (***)	0	−0.486	0 (***)	0	−0.486	0 (***)	0
Area_veh ⟵ Demographic/ Geographic				[a]	[a]		[a]	[a]		[a]	[a]	
Tot_nodes_net_veh ⟵ Network							[a]	[a]		[a]	[a]	
Tot_length_net_veh ⟵ Network							[a]	[a]		[a]	[a]	
Nat_speed_lim_law ⟵ Enforcements										[c]	[c]	
Nat_drink_law ⟵ Enforcements										0.641	0.552 (***)	0.094
Nat_helmet_law ⟵ Enforcements										0.834	0.812 (***)	0.124
Nat_seat_belt_law ⟵ Enforcements										a	a	
Nat_child_rest_law ⟵ Enforcements										0.592	0.882 (***)	0.161
Nat_mob_use_law ⟵ Enforcements										0.681	1	
Report_acc_rates_reg_veh ⟵ Economy	−0.451	−0.021 (**)	0.007	−0.121	−0.006 (0.086)	00.004	−0.121	−0.006 (0.086)	00.004	−0.121	−.006 (0.86)	.004
Report_acc_rates_reg_veh ⟵ Demographic/ Geographic				0.926	0 (***)	0	0.926	0 (***)	0	0.927	0 (***)	0
Report_acc_rates_reg_veh ⟵ Enforcements										−0.008	−10.945 (0.886)	76.551
d.f.	5			13			13			42		
AIC	26.502			100.234			100.234			178.481		
BIC	54.460			142.171			142.171			183.815		
GFI	0.979			0.877			0.877			0.846		
RMSEA	0.05			0.192			0.192			.132		

[a]Collinear Omission.
[b]Negative Variance Omission.
[c]Non-Statistical Significant Omission.
[d]The variable was omitted from the models' data sets and were used for creating the endogenous variables in the respective scenarios.
*Significant at the 0.05 level.
**Significant at the 0.01 level.
***Significant at the 0.001 level.

Table 6
Models' statistical information according 2013 data, Scenario 2.

	Model 1			Model 2			Model 3			Model 4		
	Standardized weight	Regression weight		Standardized weight	Regression weight		Standardized weight	Regression weight		Standardized weight	Regression weight	
		Estimate	SE		Estimate	SE		Estimate	SE		Estimate	SE
Income ⟵ Economy	[a]	[a]		[a]	[a]		[a]	[a]		[a]	[a]	
GNI ⟵ Economy	[a]	[a]		[a]	[a]		[a]	[a]		[a]	[a]	
GDP ⟵ Economy	0.456	0.083 ([b])	0.029	0.517	0.102 ([c])	0.03	0.42	0.074 ([b])	0.028	0.395	0.067 ([a])	0.027
Food_prod ⟵ Economy	0.652	1		0.599	1		0.668	1		0.692	1	
Tax ⟵ Economy	c	c		c	c		c	c		c	c	
Unemp ⟵ Economy	−0.292	−0.12 ([a])	0.054	−0.259	−0.115 ([a])	0.055	−0.318	−0.127 ([a])	0.055	−0.311	−0.12 ([a])	0.054
Diesel_price ⟵ Economy	−0.357	−0.016 ([a])	0.006	−0.331	−0.016 ([a])	0.006	−0.372	−0.016 ([a])	0.006	−0.387	−0.120 ([a])	0.006
Gasol_price ⟵ Economy	[a]	[a]		[a]	[a]		[a]	[a]		[a]	[a]	
Int_users ⟵ Economy	[a]	[a]		[a]	[a]		[a]	[a]		[a]	[a]	
Num_reg_veh_popd												
Con ⟵ Demographic/ Geographic				c	c		c	c				
Popul_veh							[a]	[a]		[a]	[a]	
Birth_rate ⟵ Demographic/ Geographic				[a]	[a]		[a]	[a]		[a]	[a]	
Popul_15_64 ⟵ Demographic/ Geographic				[b]	[b]		[b]	[b]		[b]	[b]	
Area_veh ⟵ Demographic/ Geographic				[c]	[c]		[a]	[a]		[a]	[a]	
Tot_nodes_net_veh ⟵ Network							[c]	[c]		[c]	[c]	
Tot_length_net_veh ⟵ Network							[a]	[a]		[a]	[a]	
Nat_speed_lim_law ⟵ Enforcements										[a]	[a]	
Nat_helmet_law ⟵ Enforcements										[c]	[c]	
Nat_seat_belt_law ⟵ Enforcements										[b]	[b]	
Nat_child_rest_law ⟵ Enforcements										[c]	[c]	
Nat_mob_use_law ⟵ Enforcements										[c]	[c]	
Report_acc_rates_reg_veh ⟵ Economy	0.469	8.807 ([b])	3.004	0.582	12.162 ([c])	3.477	0.402	7.34 ([b])	2.726	0.301	5.191 ([a])	2.133
Report_acc_rates_reg_veh ⟵ Popul_growth							0.182	33.49 ([a])	15.301	0.181	32.456 ([a])	12.988
Report_acc_rates_reg_veh ⟵ Death_rate				0.125	12.016 (0.113)	7.591	0.206	19.249 ([a])	7.801	0.181	16.582 ([a])	6.622
Report_acc_rates_reg_veh ⟵ Popul_veh				−0.243	−32.422 ([b])	10.534						
Report_acc_rates_reg_veh ⟵ Nat_drink_law										0.504	1608.774 ([c])	231.098
d.f.	5			14			14			20		
AIC	30.905			75.952			79.205			86.481		
BIC	58.863			115.093			118.346			131.214		
GFI	0.966			0.901			0.891			0.899		
RMSEA	0.099			0.142			0.149			0.12		

Notes:
[a]Collinear Omission.
[b]Negative Variance Omission.
[c]Non-Statistical Significant Omission.
[d]The variable was omitted from the models' data sets and were used for creating the endogenous variables in the respective scenarios.
[*]Significant at the 0.05 level.
[**]Significant at the 0.01 level.
[***]Significant at the 0.001 level.

References

Antoniou, C., Yannis, G., Papadimitriou, E., Lassarre, S., 2016. Relating traffic fatalities to GDP in Europe on the long term. Accid. Anal. Prev. 92, 89–96.

Baur, J., Moreno-Villanueva, M., Kotter, T., Sindlinger, T., Burkle, A., Berthold, M.R., Junk, M., 2015. MARK-AGE data management: cleaning, exploration and visualization of data. Mech. Ageing Dev. 151, 38–44.

Bertsimas, D., Freund, R., 2004. Data, Models, and Decisions: The Fundamentals of Management Science. South-Western College Publishing.

Chan, F., Lee, G.K., Lee, E.-J., Kubota, C., Allen, C.A., 2007. Structural equation modeling in rehabilitation counseling research. Rehabil. Couns. Bull. 51 (1), 53–66.

Chaolong, J., Hanning, W., Lili, W., 2016. Research on visualization of multi-dimensional real-time traffic data stream based on cloud computing. Procedia Eng. 137, 709–718.

Deb, R., Liew, A.W.-C., 2016. Missing value imputation for the analysis of incomplete traffic accident data. Inf. Sci. 339, 274–289.

Depaire, B., Wets, G., Vanhoof, K., 2008. Traffic accident segmentation by means of latent class clustering. Accid. Anal. Prev. 40, 1257–1266.

Dupont, E., Commandeur, J.J., Lassarre, S., Bijleveld, F., Martensen, H., Antoniou, C., Papadimitriou, E., Yannis, G., Hermans, E., Perez, K., Santamarina-Rubio, E., Usami, D.S., Giustiniani, G., 2014. Latent risk and trend models for the evolution of annual fatality numbers in 30 European countries. Accid. Anal. Prev. 71, 327–336.

Field, A., 2009. Discover Statistics Using SPSS, third edition. SAGE Publications.

Fomina, M., Morosin, O., Vagin, V., 2014. Argumentation approach and learning methods in intelligent decision support systems in the presence of inconsistent data. Procedia Comput. Sci. 29, 1569–1579.

Hassan, H.M., Abdel-Aty, M.A., 2011. Analysis of drivers' behavior under reduced visibility conditions using a structural equation modeling approach. Transp. Res. 14 (Part F), 614–625.

Hassan, H.M., Dimitriou, L., Abdel-Aty, M.A., Al-Ghamdi, A.S., 2013. Analysis of Risk Factors Affecting the Size and Severity of Traffic Crashes in Riyadh. Transportation Research Board Compendium.

Hellerstein, J.M., 2008. Quantitative data cleaning for large databases. Survey for the United Nations Economic Commission for Europe.

Hooper, D., Coughlan, J., Mullen, M.R., 2008. Structural equation modelling: guidelines for determining model fit. Electron. J. Bus. Res. Methods 6 (1), 53–60.

Lai, C.M., Mak, K.K., Watanabe, H., Jeong, J., Kim, D., Bahar, N., Ramos, M., Chen, S.H., Cheng, C., 2015. The mediating role of internet addiction in depression, social anxiety, and psychosocial well-being among adolescents in six Asian countries: a structural equation modeling approach. Public Health 129, 1224–1236.

Ma, J., Lu, J., Zhang, G., 2009. Information inconsistencies detection using a rule-map technique. Expert Syst. Appl. 36, 12510–12519.

Maydeu-Olivares, A., Garcia-Forero, C., 2010. Goodness-of-fit testing. In: third edition. In: Peterson, P., Baker, E., McGaw, B. (Eds.), International Encyclopedia of Education Vol. 1. pp. 190–196.

Preacher, K.J., Merkle, E.C., 2012. The problem of model selection uncertainty in structural equation modeling. Psychol. Methods 17 (1), 1–14.

Saha, P., Roy, N., Mukherjee, D., Sarkar, A.K., 2016. Application of principal component analysis for outlier detection in heterogeneous traffic data. Procedia Comput. Sci. 83, 107–114.

Tahmasebi, M., Rocca, M., 2015. A fuzzy model to estimate the size of the underground economy applying structural equation modeling. J. Appl. Econ. 2, 347–368.

Washington, S.P., Karlaftis, M.G., Mannering, F.L., 2011. Statistical and Econometric Methods for Transportation Data Analysis, second edition. CRC Press.

Wegman, F., Allsop, R., Antoniou, C., Bergel-Hayat, R., Elvik, R., Lassarre, S., Lloyd, D., Wijnen, W., 2017. How did the economic recession (2008–2010) influence traffic fatalities in OECD-countries? Accid. Anal. Prev. 102 (May (1)), 51–59.

Yannis, G., Papadimitriou, E., Folla, K., 2014. Effect of GDP changes on road traffic fatalities. Saf. Sci. 63, 42–49.

Yu, C.-Y., 2015. Disparity in Traffic Safety Across Neighbor-Hoods With Different Economic Statuses and Ethnic Com-Positions. Transportation Research Board Compendium.

Yuan, K.-H., Bentler, P.M., 2006. Structural equation modeling. Handbook of Statistics Vol. 26. pp. 297–358.

Zhou, H., Romero, S.B., Qin, X., 2016. An extension of the theory of planned behavior to predict pedestrians' violating crossing behavior using structural equation modeling. Accid. Anal. Prev. 95 (Part B), 417–424.