Macroscopic traffic safety data analysis and prediction

Constantinos Antoniou[1*], George Yannis[2] and Eleonora Papadimitriou[2]

National Technical University of Athens

[1] School of Rural and Surveying Engineering, Zografou 15780,

T: +30 210 7722783, F: +30 210 7722629, E: antoniou@central.ntua.gr

[2] Department of Transportation Infrastructure and Engineering, Zografou 15773,
T: +30 210 7721326, F: +30 210 7721454,

E: {geyannis, nopapadi}@central.ntua.gr

*corresponding author

Abstract

Macroscopic modeling of traffic safety can provide insight into this global health problem and help policy-makers in both under-developed and developing countries adjust their policies in reaction to the changing conditions. The objective of this paper is to illustrate how macroscopic road safety data analysis can be useful in explaining road safety trends and patterns and thus supporting road safety policies and initiatives. Statistical techniques for the macroscopic analysis of road safety data are presented, followed by case study results from European countries. Practical issues, such as measures of goodness of fit and model diagnostics are also discussed. A discussion on emerging trends and state-of-the-art in the field concludes the paper.

Keywords road safety; macroscopic analysis; generalized linear models; non-linear models; state-space models

Introduction

Modeling road safety is a complex task, which needs to consider both the quantifiable impact of specific parameters, as well as the underlying trends that cannot always be measured or observed. The sensitivity of users to road safety campaigns, the improved quality of the vehicle fleet, the improvement of the driving skills of the general population, and the overall improvement of the condition of the road network are only some of the aspects that cannot be easily modeled directly. Therefore, modeling should consider both measurable parameters and the dimension of time, which embodies all remaining parameters.

Macroscopic modeling can provide insight into this problem and support policy-makers in developing countries to adjust their policies in reaction to the changing macroscopic conditions. For example, developing insight e.g. regarding the expected breakpoints in road safety fatality trends, as identified from the developed countries time series, can be applied to performing more accurate future predictions for developing countries (which have still not reached the motorization levels of developed countries). Macroscopic, in this context, implies to the analysis of aggregate (monthly or annual) accident and fatality data, rather than the more detailed and disaggregate (in-depth) analysis of individual accidents.

The objective of this paper is to illustrate how macroscopic road safety data analysis can be useful in explaining road safety trends and patterns and thus supporting road safety policies and initiatives. Several types of models are presented in this context, including generalized linear models, non-linear models and state-space models. The paper follows a "why-how-what" approach, motivating the need for this analysis (why), showing which techniques can be used (how) and finally using selected case study results to illustrate typical results of the analysis (what). In this way, the reader is exposed to a thorough coverage of the models that can be useful in the context of macroscopic data analysis, so that it is then possible to follow specific research directions of interest.

The remainder of this paper starts with a literature review, summarizing the use of statistical techniques for the macroscopic analysis of road safety data, indicative examples of which are then presented, followed by a discussion of their key properties. The methods that can be used to analyze macroscopic road safety data are then presented; time series models are first introduced, with an emphasis on their functional form (linear, generalized linear and non-linear models). State-space models are also introduced, along with the concepts of multivariate and multilevel models. These approaches are then demonstrated using case study results from Greece and other European countries. A discussion on emerging trends and state-of-the-art in the field concludes the paper.

Literature review

A macroscopic road-safety model commonly used in the late 60s was proposed by Smeed (1968) linking the number of fatalities with the number of vehicles and the population. Jacobs (1986) repeated this analysis for a number of developed and developing countries using data between 1968 and 1975 while Gharaybeh (1994) applied the same formula to assess the development of road safety in Jordan, relative to that of other middle-eastern and developing countries. Many studies have criticised Smeed's model because it only concentrates on the motorisation level of country and ignores the impact of other variables (cf. Broughton, 1991, Andreassen, 1991, while another useful review is provided by COST329, 2004, where a detailed analysis of the debate surrounding Smeed's formulas and analysis is available).

The comparative analysis of macroscopic trends in road-safety-related issues among countries and regions has attracted the attention of researchers for several decades. A critical review of a number of approaches for modelling road safety trends can be found in Hakim et al. (1991) and Oppe (1989). A review of these concerns, as well as several alternative approaches for the development of road safety models is provided by Al-Haji (2007). Lassarre (2001) presented an analysis of ten European countries' progress in road safety by means of a structural (local linear trend) model, yielding two adjusted trends, one deterministic and one stochastic. Intervention functions related to the major road safety measures were introduced, while an indicator of the rate of progress given risk exposure trends (vehicle-km travelled) was defined.

Page (2001) presented an exponential formula that yields fatalities as the product of all explanatory variables' influence, which could be transformed to a simple algebraic form (first order polynomial with an intercept) by taking the logarithm of both sides. Models with several exogenous variables are developed and attempts to rank countries based on their road mortality level were made. Beenstock and Gafni (2000) show that there is a relationship between the downward trend in the rate of road accidents in Israel and other countries and suggest that this reflects the international propagation of road safety technology as it is embodied in motor vehicles and road design, rather than parochial road safety policy. Van Beeck et al. (2000) examine the association between prosperity and traffic accident mortality in industrialized countries in a long-term perspective (1962-1990) and find that in the long-term the relation between prosperity and traffic accident mortality appears to be non-linear. Kopits and Cropper (2005) use linear and log-linear forms to model region specific trends of traffic fatality risk and per income growth using panel data from 1963 to 1999 for 88 countries. Abbas (2004) compares the road safety of Egypt with that of other Arab nations and G-7 countries, and develops predictive models for road safety.

Other analyses entail a specific road safety related problem, applying international macroscopic comparison techniques to a subset of road network users, such as novice or young drivers. Twisk & Stacey (2007) presented a general study of identified trends in young drivers risk and associated countermeasures in certain European countries. The relationship between general safety levels and young driver risks is stressed: the impact of general safety measures on the subgroup is greater than that of measures specifically targeting young drivers, especially for poorly performing countries.

Another big topic of research relates to the factors that affect road safety and the way that this impact is applied. Several researchers (Hakim et al., 1991; Cameron et al, 1993; Newstead et al., 1995, Lord, 2002), using road accident statistics, have presumed that the explanatory variables have a multiplicative effect on accidents (as opposed to e.g. additive). Henning-Hager (1986) presented a non-linear regression model to express the relationship between traffic fatalities, traffic volumes and the quality of transportation supply and demand in urban areas. Qin et al. (2004) showed that the relationship between crashes and the daily volume (AADT) is non-linear and varies by crash type, and is significantly different from the relationship between crashes and segment length for all crash types. On the other hand, vehicle fleet may affect the number of fatalities, given that an increase in the vehicle number leads to higher average traffic volumes, which in turn may translate to a reduction in average speeds. Moreover, an increase of the vehicle fleet and total mileage in a country increases the need for more and safer road environment, in which the drivers' behaviour tends to be also better (Koornstra, 1992, 1997).

Clearly, the topic of macroscopic road safety modeling and forecasting is an active research area, where active debate is taking place and interesting developments are still being made.

A look at the data – (Why?)

In order to understand why it is useful (and practical) to analyze macroscopic road safety data, as well as some of the caveats involved in doing so, it is useful to look at some indicative data that can be used for demonstration purposes. Figure 1 presents the number of persons killed and seriously injured per month in Greece for the period between 1998 and 2004 (demonstrating the clear road safety improvements obtained between 2000 and 2004 in Greece). Besides a rather clear decreasing trend, perhaps the most striking feature of this graph is the periodicity (annual seasonality) in the data. A failure to identify, capture and analyze this trend can result in loss of understanding of the underlying phenomenon.

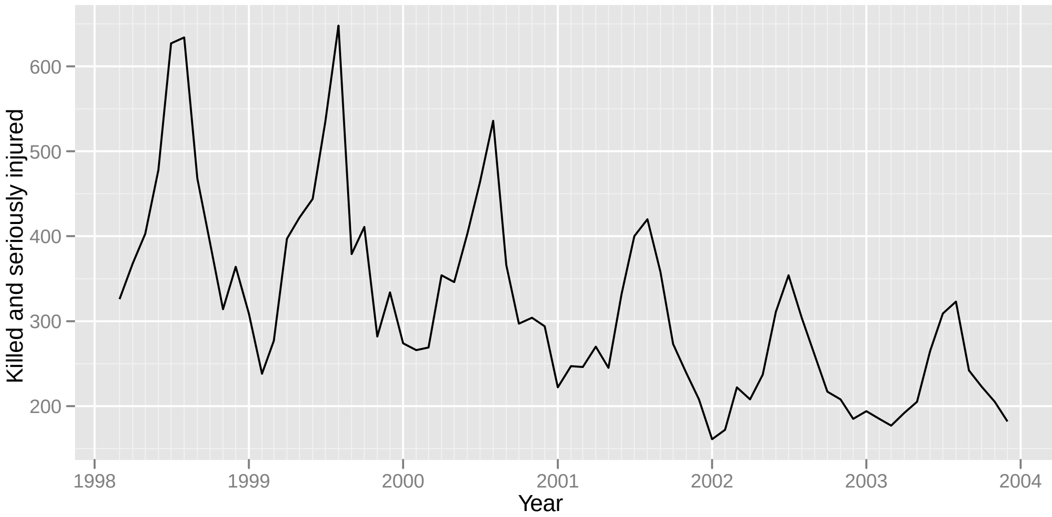


Figure 1. Example of seasonal trends / periodicity in macroscopic road safety data

Many statistical techniques assume independence of observations. Road safety data, however, as shown in Figure 1, are often in the form of time-series of counts observed during successive time periods, e.g. days, months or years. In practice, such observations often tend to be correlated with the respective observations from previous years, months or days, i.e. are usually temporally correlated. The linear regression model -an attractive and simple method- has stringent assumptions that are therefore usually violated when applied to road safety data. Another similar assumption is that of linearity (in the parameters). In later sections, alternative modeling assumptions are presented within the more flexible generalized linear modeling and non-linear regression frameworks.

Figure 2 presents an overview of key road safety indicators (personal risk defined as fatalities per 100.000 population) in several European countries. The main observation that one can make from this data is that –in general- there seems to be an increasing trend in road accident fatalities/personal risk up to a point where this trend is reversed. Understanding why this breakpoint and trend reversal happens is clearly an important element of improving road safety in countries that have yet not experienced this level.

Exposure data are very useful in road safety analysis, as they help illustrate the underlying trends that lead to the road safety situation. Key exposure measures are the vehicle-kilometers or the person-kilometers traveled, or the time traveled. However, collecting or estimating exposure data is a much more difficult endeavor and these data are often unavailable for analysis. One way to overcome this limitation is to seek proxies, i.e. available (or more easily collectable) data that have a high correlation with the actual exposure data. Examples of these are the number of vehicles in circulation or the amount of fuel sold at gas stations.

In this case, as shown in the bottom subfigure of Figure 2, motorization data can be used as a proxy to exposure data. The data suggest that there may be a point, as the motorization level increases, that personal risk stops increasing and starts to decrease. If this point can be located, then this may mean that road safety decision makers in countries that have not yet reached that breakpoint, might be able to foresee it and incorporate it in their strategies and policies. Alternatively, knowing about this breakpoint might protect road safety decision makers in countries that have not yet reached that point, from believing that they have achieved a road safety breakthrough, when they reach it.

Furthermore, it is worth noting that several countries (such as the Czech Republic and Poland in Figure 2) show more than one clear peaks. (In this case, one could argue that they are due to the "opening" of their economies in the late 80s/early 90s, but of course further exploration and verification is required.) Similarly, one can identify smaller peaks in all subfigures and explaining them could lead to the development of interesting insight into the road safety of these countries.

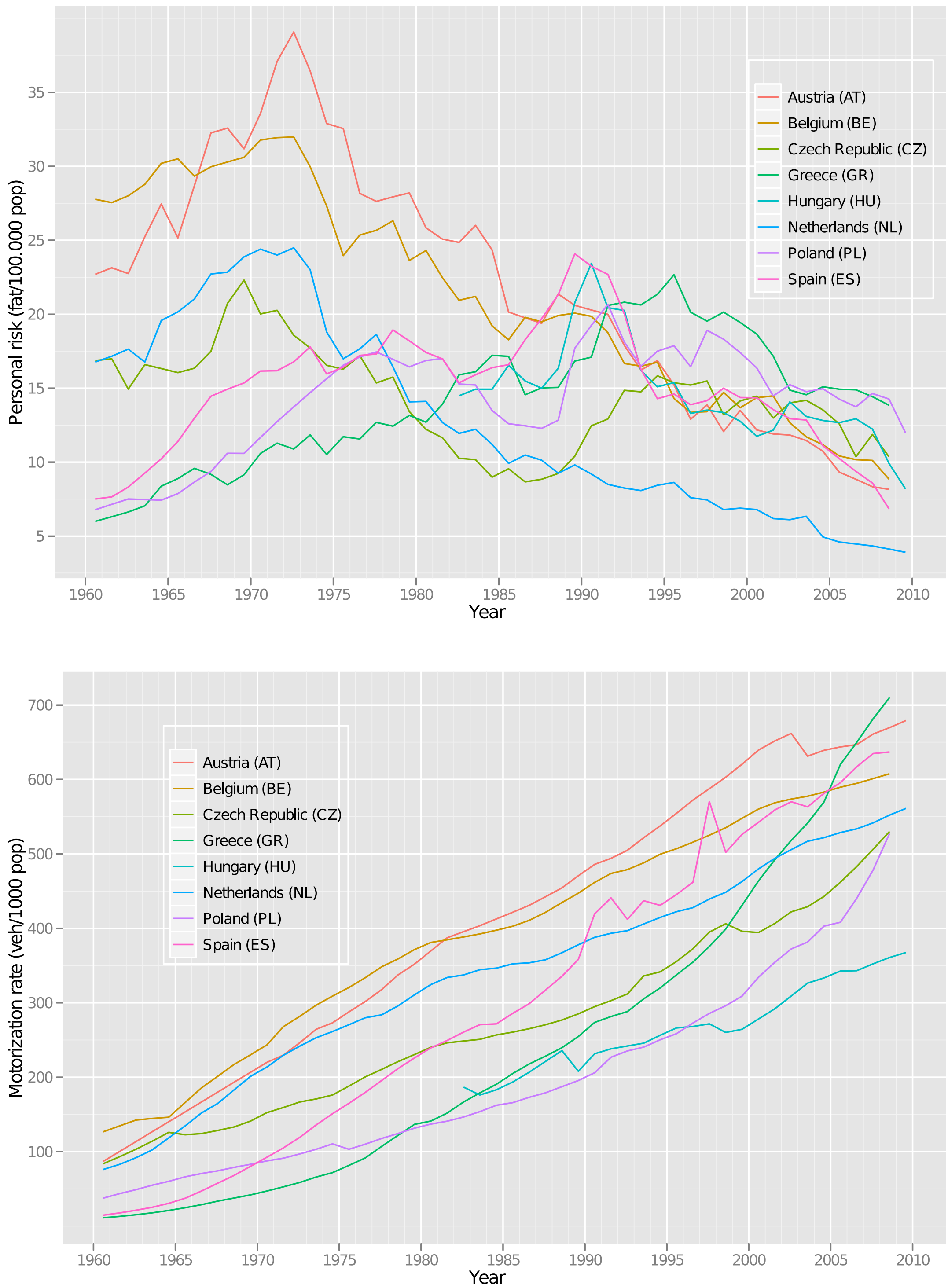


Figure 2. Breakpoints and trend reversals in road safety data (adapted from Yannis et al., 2011)

The increased understanding of trends and the systematic steering of road safety related policies and campaigns could improve their effectiveness. For example, previous research in Greece analyzing data at the county level has shown a significant 2-months time halo effect of enforcement (Agapakis and Mygiaki, 2003), which – if not explicitly considered- may result in misleading observations regarding the effectiveness of the considered measures. Especially as regards the distance halo effect, this would concern the case that the effect of enforcement would ‘‘cross” a county’s borders to the neighbouring counties of the same or neighbouring regions. However, given that the intensification of enforcement was not individually decided by local authorities, but was instead applied on all counties of Greece, it is considered that the cross-county effects of neighbouring counties (and regions) do not significantly affect the within-county (or region) effects.

Of course, it needs to be clearly stated that aggregated, macroscopic data are only one aspect of the overall topic of road safety that can provide only part of the answer. Other types of data sources that can shed light from different angles into the parameters that govern road safety include in-depth accident data (Yannis et al., 2010) and medical information data (Petridou et al., 2009).

Methods - (How?)

Time-series methods have been used to account for and correct temporal correlation, such as that observed typically in macroscopic road-safety data. It is recognized however that traffic fatality risk also depends on other parameters, such as vehicle quality, traffic safety initiatives and regulations, and intensity of police enforcement, which however are not expected to affect the results of such macroscopic analysis. Therefore, this section includes a presentation of methodological instruments that can be used to develop models linking road safety measures and related explanatory data in flexible ways and thus possibly offer insight on the existing or future trends for the same or other environments.

Generalized Linear Models

The linear regression model is simple, elegant and efficient, but it is subject to the fairly stringent Gauss-Markov assumptions (Washington et al., 2003). If these assumptions hold, it can be shown that the solution obtained by minimizing the sum of squared residuals ('least squares') is BLUE, i.e. best linear unbiased estimator. In other words, it is unbiased and has the lowest total variance among all unbiased linear estimators. These assumptions, however, are often violated in practice. Yannis et al. (2007a) illustrate how two of these violations can be explicitly considered, in particular correlated disturbances; and non-normal error structures. The choice of these two violations is not arbitrary; instead it is motivated by the fact that these two violations are more relevant to the nature (time-series count data) of the road safety data. Generalized linear models (GLM), a generalization of the linear regression, can be used to overcome the restriction on the normality of the error structure (McCullagh and Nelder, 1989, Dobson, 1990). Specific treatment of the application of GLM in the presence of serially correlated count data is also presented.

The objective of GLM is to allow for more flexible error structures, besides the Gaussian, which is assumed by –linear and nonlinear– regression. The Poisson distribution has been considered suitable to counts of car crashes for a long time (Nicholson and Wong, 1993). However, the Poisson model -while arguably more appropriate than the Gaussian- is not without weaknesses and technical difficulties. For example, the assumption of a pure Poisson error structure may prove inadequate in the presence of "overdispersed" data (Maycock and Hall, 1984). Overdispersion reflects more variation in the response than what is expected by the Poisson assumption, which assumes that the variance equals the mean. An implication of overdispersion is that the estimates of the standard errors of the parameters will not be correct, and in fact the standard errors will be underestimated.

A straightforward approach to overcome this issue is to use a quasi-Poisson model; i.e. estimate a dispersion parameter for the Poisson model, thus allowing it to take values other than one. Maycock and Hall (1984) showed that the negative binomial model could also be used as an extension to the Poisson.

Miaou (1994) and Wood (2002) have also used the negative binomial model for road safety applications. Maher and Summersgill (1996) mention that, quite often, the two approaches (quasi-Poisson and negative binomial) may provide very similar estimation results. One may then be tempted to think that the two models are equivalent and that it does not really matter which model is selected. Maher and Summersgill further warn that this may not be the case, as the two models may have different prediction properties, as measured, e.g. by the prediction error variance. Lord et al. (2005) present the results of an examination of the applicability of different models, including Poisson, negative binomial (or Poisson-gamma) and zero-inflated Poisson and negative binomial models, to the modeling of accident data.

Furthermore, the generalized linear modeling framework allows the consideration of a limited amount of non-linear structures in the developed models. For example, several researchers have shown that conventional linear regression models lack the distributional property to adequately describe collisions. This inadequacy is due to the random, discrete, non-negative, and typically sporadic nature that characterize the occurrence of vehicle collisions. Several researchers (including Hauer et al.1988, Hakim et al., 1991; Cameron et al., 1993; Newstead et al., 1995), using road accident statistics, have presumed that the explanatory variables have a multiplicative effect on accidents, i.e. $y = ax_1^b x_2^c$ (as opposed to e.g. additive, i.e. $y = a + bx_1 + cx_2$).

Examples of road safety applications involving the use of GLM in temporally correlated data include before/after analysis on the impact of red-light camera presence in crashes (Retting and Kyrychenko, 2001), investigation of relationships between accidents, flows and road or junction geometry, allowing for the presence of a trend over time in accident risk (Maher and Summersgill, 1996), traffic safety comparisons among several counties in France, where the time trend of each index (incidence and severity) is the same across counties and across road types (Amoros et al., 2003), and estimation of expected junction accidents (both in total and disaggregated by severity, road surface condition and lighting condition), which allow for the possibility of accident risk varying over time (Mountain et al., 1998). White and Washington (2001) developed a logistic regression model to gain insight into the relationship between enforcement and the use of safety restraint.

Generalized linear models facilitate the analysis of the effects of explanatory variables in a way that closely resembles the analysis of covariates in a standard linear model, but with less confining assumptions. This is achieved by specifying a *link function*, which links the systematic component of the linear model with a wider class of outcome variables and residual forms.

A key point in the development of GLM was the generalization of the normal distribution (on which the linear regression model relies) to the exponential family of distributions. This idea is not new and was developed by Fisher (1934). Many well-known distributions belong to the exponential family, including –for example– the Poisson, normal, and binomial distributions. On the other hand, examples of well-known and widely used distributions that cannot be expressed in this form are the student's t-distribution and the uniform distribution. The generalized linear model can be defined in terms of a set of independent random variables, each with a distribution from the exponential family.

Non-linear regression

Besides not conforming to the normality and other assumptions, many interesting processes may be more adequately modeled by non-linear models in practice. Linear regression models might have been a practical necessity in the past, but theoretical and computational developments have made the use of more elaborate (appropriate, accurate) methods practical. This can also be seen in road safety research, where while early work used multiple linear regression modeling (assuming normally distributed errors and homoscedasticity), over the past two decades there has been a departure from this model. In the previous section it was shown that generalized linear models allow for some nonlinear relationships to be modeled and relax some restrictions on the distributional assumptions of linear regression. Although many scientific and engineering processes can be described well using linear models, or other relatively simple types of models, there are many processes that are inherently nonlinear. Non-linear models can then be used (see e.g. Bates and Watts, 1988). The biggest advantage of nonlinear regression over many other techniques is the broad range of functions that can be fit.

A non-linear regression model can be written as:

$$Y_m = f\left(x_m, q\right) + Z_m \tag{6}$$

where f is the expectation function, x_m is a vector of associated regressor variables or independent variables for the nth case, Y_m is the dependent variable, θ is a vector of parameters to be estimated and Z_m are random disturbances. This model is of the same general form as the linear model, with the exception that the expected responses are nonlinear functions of the parameters. More formally, for non-linear models, at least one of the derivatives of the expectation function with respect to the parameters depends on at least one of the parameters. Non-linear regression has been widely used in road-safety related research (e.g. Hakim et al., 1991, Qin et al., 2004, among many others).

The Gauss-Markov assumptions from ordinary least square (OLS) procedures still apply in non-linear regression. Therefore, whenever time or distance is involved as a factor in a regression analysis, it is important to check the assumption of independent residuals. When the residuals are not independent, the model for the observations must be altered to account for dependence (e.g. moving average or autoregressive models of variable order).

Of course many other types of models find common use, but not all of them can be covered in detail in this article. Examples of such models include multilevel models, see e.g. Yannis et al., 2007b, multivariate models, see e.g. Yannis et al., 2008, for an example of multivariate multilevel models and state-space time-series analysis, see e.g. Commandeur and Koopman (2007) for an introduction to the topic with practical examples from road safety applications.

Case studies Results – (What?)

Examples of case study results demonstrating the modeling techniques introduced in the previous sections are presented in this section. The following cases are illustrated:

- Modeling of past data in order to obtain insight into past trends and breakpoints
- Modeling of data for prediction of future trends (using multiple techniques)
- Analysis of distributional assumptions through appropriate model diagnostics

All models have been estimated using the R Software for Statistical Computing (RDCT, 2011).

Analyzing road safety data, such as those presented in Figure 2, can provide answers to many interesting questions from multiple perspectives. For example, from a road safety point of view, the following questions are interesting:

- Is the trend "universal"? What causes it?
- Does the trend happen at the same time in all countries?
- Can we use this to make predictions?

But even from a purely statistical point of view there are interesting question relating to the way that these structural changes can be estimated: starting from simple piece-wise regression, to estimation of consistent trends given an exogenous number of breakpoints, to simultaneous estimation of breakpoints and trends.

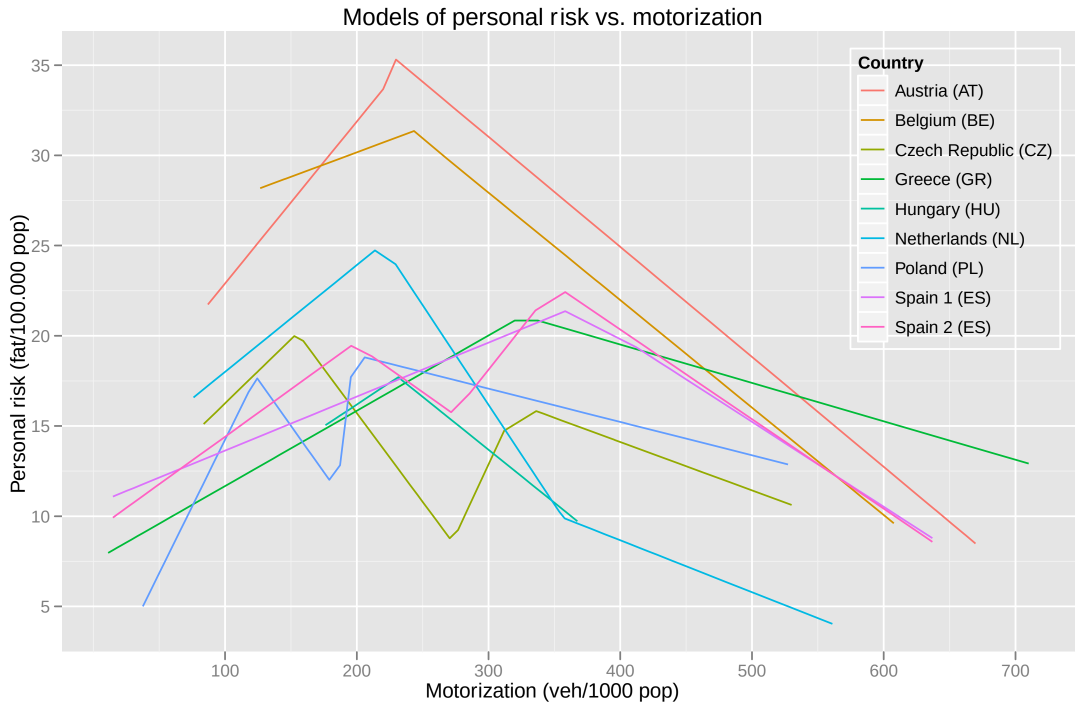
Figure 3 summarizes the estimated models using the data in Figure 2 (estimated using the segmented R package, Muggeo, 2003, 2008), providing a concise overview that can be used to draw conclusions, including the following:

- Different countries reached specific motorization rates at different (and sometimes distant) moments in time (temporal landmarks);

- Some of those countries exhibit a break point within a narrow range of motorization rate values, implying perhaps similar social and economic conditions and/or similar road safety culture;
- This range is different for certain subgroups among the examined countries, providing a hint that some grouping may be of meaning in geographic and socioeconomic context.

It is noted that the estimated models are linear (between breakpoints) and have been estimated using motorization rate as the explanatory variable (top subfigure). In the middle subfigure the same data are plotted against time for visualization purposes.

Before strong conclusions can be drawn based on the interpretation of such results, several considerations must be made to ensure that the models are indeed directly comparable, e.g. the data definition across countries. The numerator of the motorization rate (fatalities), for example, may be regarded more or less well-defined, after many efforts put at pan-European level for a common definition (30-day fatalities). As far as the denominator is concerned, however, available data of vehicle fleet show some slight discrepancies, e.g. the total number of vehicles in Spain reveals some irregular steps for specific years. Furthermore, each vehicle class is ruled by specific particularities, presumably implying a camouflage for systematic errors (Katsochis et al., 2006). The application of common definitions should be further examined, so that there is an as-common-as-possible base for comparison.



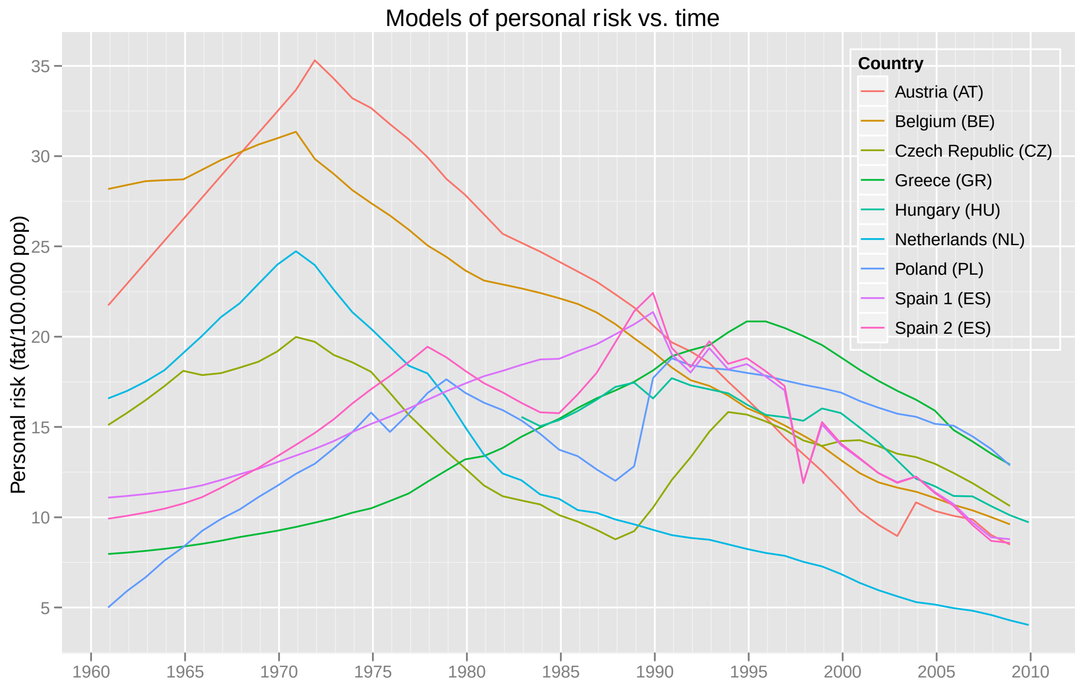


Figure 3. All estimated models. Top → personal risk vs. motorization, Bottom → personal risk vs. time (Adapted from Yannis et al., 2011)

Figure 4 shows the values predicted by the quasi-Poisson model. The dashed line shows the actual observed number of persons killed and seriously injured in Greece (excluding the two major metropolitan areas of Athens and Thessaloniki). The thick solid line represents the model predictions and 95% confidence intervals are also shown with thinner solid lines. The data show a clear seasonality, which is maintained even as the magnitude of the fatalities considerably decreases over time. Interpreting this annual periodicity is an involved process that requires additional data, relating e.g. to weather conditions. During the winter months fatalities decrease. The climate in Greece is mild, meaning that there are limited extreme dangerous conditions during the winter (e.g. frost, ice). On the other hand, daytime is shorter and in general people tend to limit their discretionary trips, which translates into a decrease in the vehicle-kilometers traveled. On the other hand, during the summer the day is longer, people drive more and therefore are exposed more to risk. Furthermore, August is typically the vacation month in Greece. This may have several implications, e.g. more interurban and less urban traffic due to holidays, which may result in higher accident severity. Another possible cause for the increase in the fatalities in August is that drivers spend more time in unknown roads (while on vacation) or perhaps drive more while tired or after having consumed alcohol.

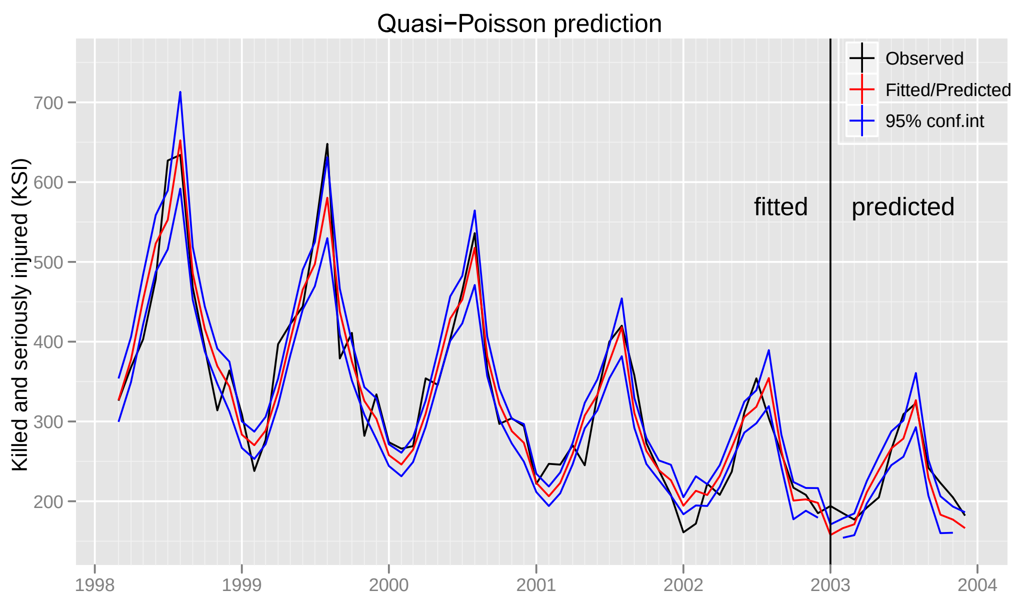


Figure 4. Quasi-Poisson model predictions (adapted from: Yannis et al, 2007a)

State-space models offer another way to analyze macroscopic road safety data. Figure 5 presents the prediction results from a latent risk time-series (LRT) analysis of annual fatality and motorization data in Greece. Data from the period 1960-2008 have been used to make predictions up to 2020, including confidence intervals. The model specification allows for the incorporation of "interventions", i.e. modeling points in time during which significant events occurred that influenced the evolution of the modeled phenomenon, shown by the broken lines in the model results. One question that arises from Figure 5 relates to the expanding margins of the prediction. Considering that road safety is a very complex process, affected by a number of natural causes (such as weather) and man-made effects (such as the economic conditions, the development of new motorways and traffic-related laws and regulations), it is reasonable to expect such a wide range. Disseminating this information to the general public, or even policy-makers and decision-makers, who might not be as comfortable with the underlying statistics, might require a different action plan. For example, instead of showing this one scenario, including prediction and confidence intervals, one might choose to show the prediction from a small number of scenarios (e.g. pessimistic, most likely, optimistic). Certainly, this is not a trivial exercise.

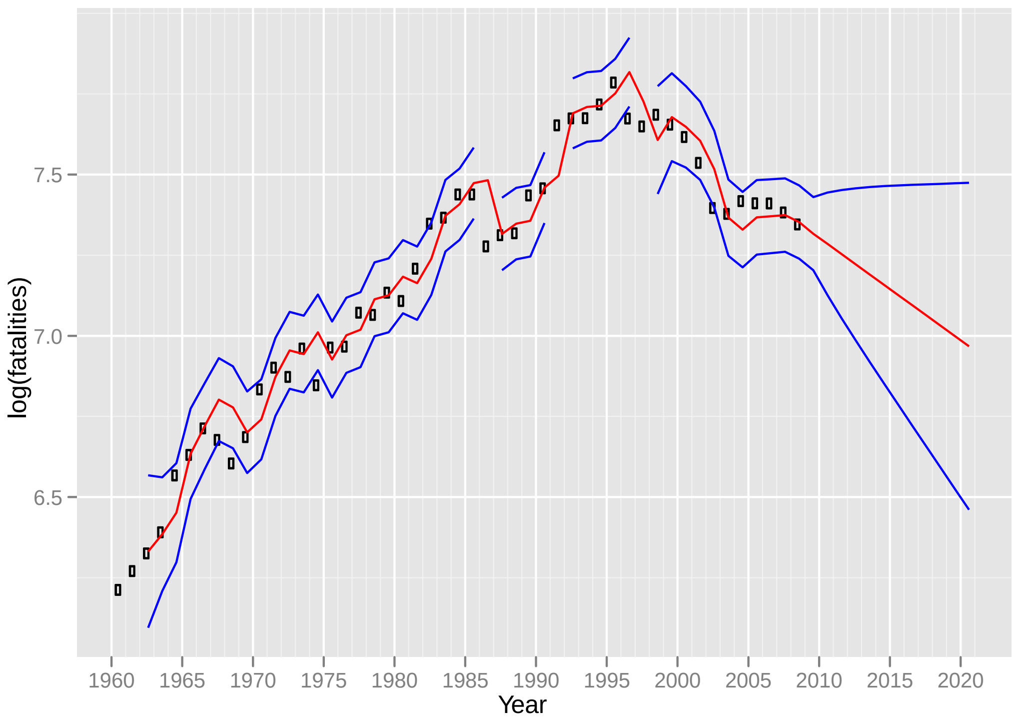


Figure 5. Latent risk time-series with interventions model prediction

Model diagnostics

The advent of powerful computers and sophisticated software has made the specification and estimation of complex model forms possible. However, with power come perils and responsibility, as it is not uncommon for researchers to estimate complex models without being fully aware of the assumptions that these models must comply to and the issues that originate from their violations. Knowing which diagnostics to use and how to apply and interpret them correctly is at least equally important as specifying and estimating a model. While a small number of model diagnostics are mentioned in this section, the appropriate tools for each model application should be sought from the relevant literature.

A large number of aggregate tests have been developed for the assessment of the goodness of fit of alternative models. However, there are several pitfalls that should be avoided when attempting to use such measures. For example, it should be noted that the usual tests for comparing nested models estimated using maximum likelihood estimation, such as the Akaike Information Criterion, AIC, (Akaike, 1973) or the Schwarz/ Bayesian Information Criterion, BIC, (Schwarz, 1978), are not suitable for comparison across these (non-nested) models. For example, AIC or BIC could be used to compare models with different numbers of parameters and the same likelihood function (except for the number of parameters), e.g. two normal or two Poisson models, but not one normal and one Poisson.

Some model diagnostics for the analysis of model residuals are presented in Figure 6 for two models: one in which the dependent variable is assumed to follow a Poisson distribution and one in which it is assumed to follow a quasi-Poisson distribution. Normal scores plot (QQ plot) of standardized deviance residuals are presented in the top subfigures. The x-axis represents the standardized deviance residuals, while the y-axis represents the quantiles of the standard normal. The dotted line in the QQ plot (top) is the expected line if the standardized residuals are normally distributed, i.e. it is the line with intercept 0 and slope 1. If the deviance residuals were normally distributed, all points on the plot would fall on this dotted line. The Poisson model residuals clearly do not follow a normal distribution. The quasi-Poisson model deviance residuals, on the other hand, are practically normally distributed.

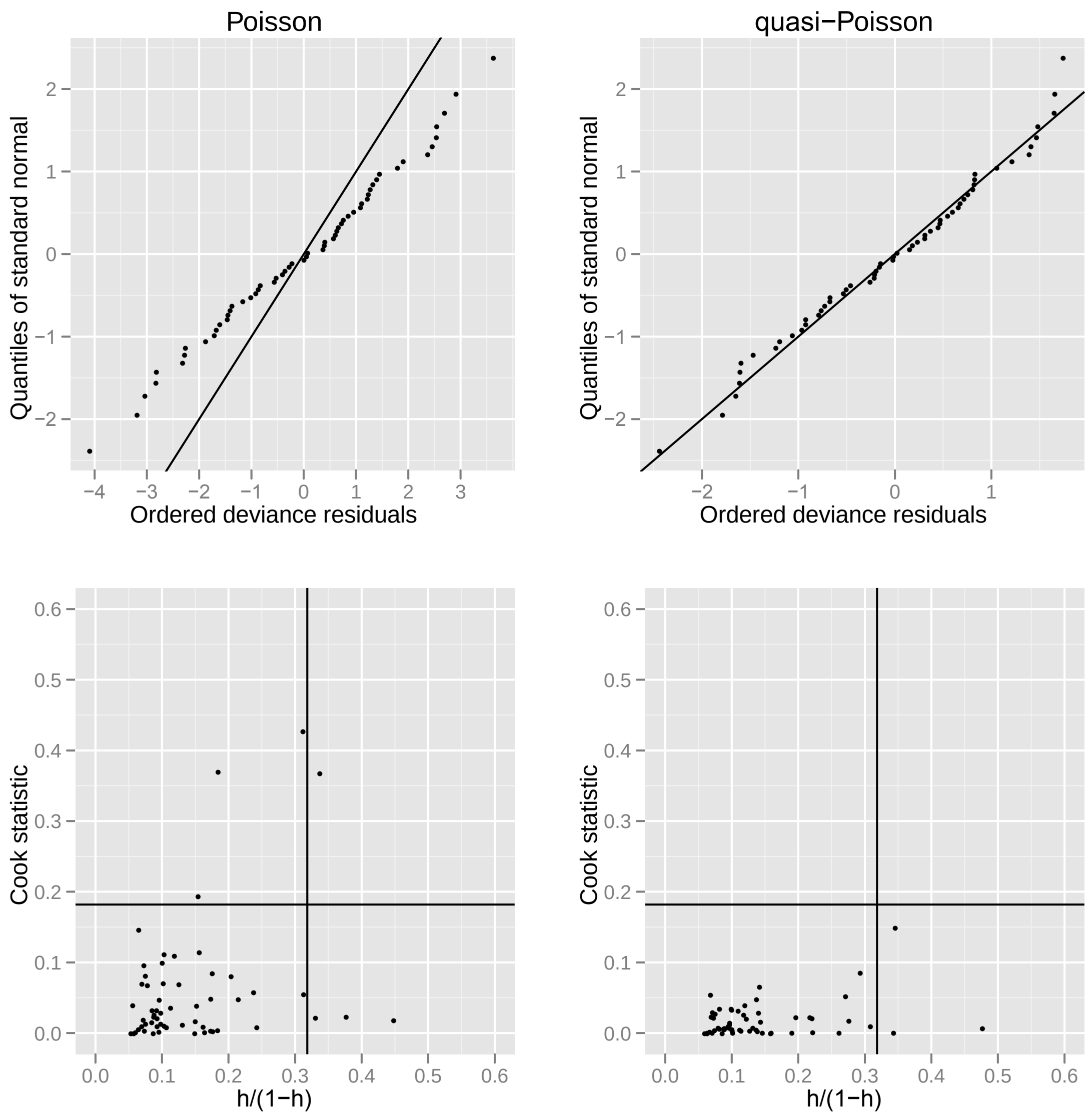


Figure 6. Model fit diagnostic plots (adapted from: Yannis et al., 2007a)

The bottom subfigures feature plots of the Cook statistics against the standardized leverages. The standardized leverage of the i-th observation x_i can be computed as (Belsley et al., 1980):

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x}_i)}{(n-1)s_x^2} \quad (7)$$

where n is the number of observations, the overbar indicates the predicted value, and s_x is the standard error. There are two dotted lines on each plot. The horizontal line is at 8/(n-2p) where n is the number of observations and p is the number of parameters estimated. Points above this line may be points with high influence on the model. The vertical line is at 2p/(n-2p) and points to the right of this line have high leverage compared to the variance of the raw residual at that point. If all points are below the horizontal line or to the left of the vertical line then the line is not shown. For example, in the quasi-Poisson plots, the horizontal line is not present, since no point lies above it.

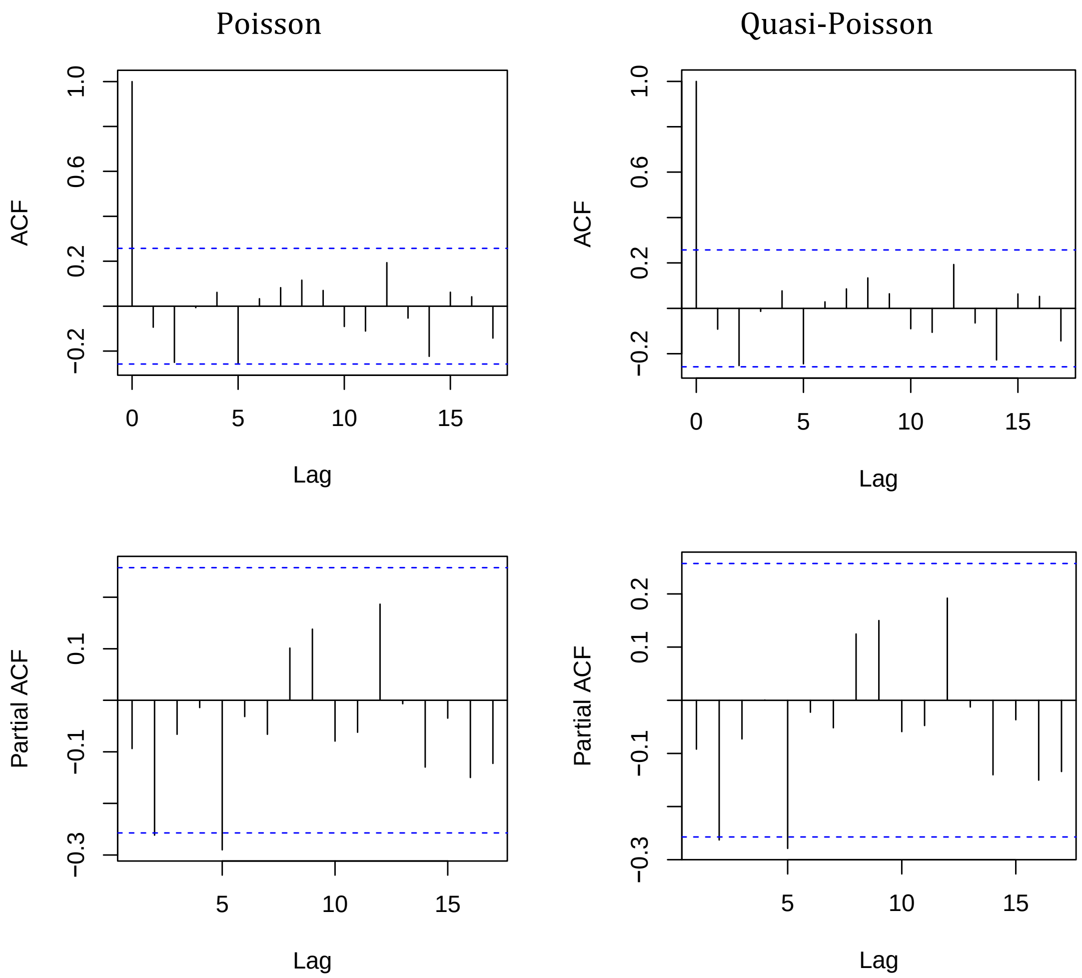


Figure 7. Residual autocorrelation plots

Finally, an important consideration when dealing with serially correlated data is the autocorrelation of the residuals. Residual plots should be analyzed to check for autocorrelation, while autocorrelation (ACF) and partial autocorrelation function (PACF) plots are also very helpful. Figure 7 shows the ACF and PACF plots for the Poisson and quasi-Poisson models discussed above, indicating that there are no serious autocorrelation issues in the residuals of either model (as the only value that exceeds the threshold is in the partial ACF for a lag of five).

Tests for other properties and assumption may also be used as needed (e.g. normality, heteroscedasticity, skewness). Tests for serial correlation are of particular interest in time-series contexts (e.g. "portmanteau" and Ljung-Box tests, Ljung and Box, 1978). A large number of measures have also been developed for the assessment of the predictive performance of these models, such as RMSPE, MPE, ME, MEN (Pindyck and Rubinfeld, 1997).

Discussion

Many other techniques can and have been used for the analysis of macroscopic road safety data, including classification of data. For example, Wegman and Oppe (2010) used Singular Value Decomposition and Multiple Correspondence Analysis of a number of observed characteristics to group European countries into more homogeneous classes in terms of road safety, while Gitelman et al. (2010) used Principal Components Analyses and Factor Analyses on European countries' data in an attempt to design a composite indicator for road safety.

Another way that the analyses presented in this paper could be further enhanced is through stratification involving specific vehicle types and population subsets (e.g. age groups or gender) (Stipdonk et al., 2010). It will then be much easier to distinguish cases and consider the presence of true impact due to GDP, vehicle fleet or other growth-related parameters; so, it is not advised to neglect the study of such elementary indicators, especially when difficulties are encountered in the reliability of more exposure-oriented analyses (e.g. using vehicle-kilometres travelled). Further research directions include the enrichment of the model with additional macroscopic parameters, as well as the investigation of other functional forms and model specifications. Additional parameters (such as the Gross Domestic Product, GDP) may help separate exogenous effects and isolate road safety trends and can be used to construct appropriate indicators. Hollo et al. (2010) use road safety performance indicators to analyze the trends in casualties in several Central European countries.

Other functional forms may also provide valuable insight into the road-safety problem. One relevant question is whether road safety trends are similar for best and worst performing countries and subsequently to find the inflection points defining the thresholds between the changing trends. This question may be proved very beneficial mainly for the less developed countries from a road safety point of view. An alternative modeling approach would have been the use of structural time-series models, such as those proposed by Harvey and Shephard (1993), Harvey (1994), which belong to the family of unobserved component models.

Acknowledement

This paper is based on the work carried out in the SAFETYNET and DACOTA research projects, partly financed by the European Commission.

References

Abbas, K. A., 2004. Traffic safety assessment and development of predictive models for accidents on rural roads in Egypt. Accident Analysis and Prevention, 36 (2), pp. 149-163.

Agapakis, J., Mygiaki, E., 2003. Macroscopic investigation of the effect of enforcement on the improvement of road safety on Greece. Diploma Thesis, NTUA, School of Civil Engineering, Department of Transportation Planning and Engineering, Athens, Greece (in Greek).

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Second International Symposium on Information Theory (B. Petrox and F. Caski, eds.), 267–281. Akademia Kiado, Budapest.

(Reprinted in Breakthroughs in Statistics, eds Kotz, S. & Johnson, N. L. (1992), volume I, pp. 599–624. Springer, New York.

Al-Haji, G., 2007. Road Safety Development Index (RSDI). Theory, Philosophy and Practice. Linkoeping Studies in Science and Technology, Dissertation No. 1100, Norrkoeping, Sweden.

Amoros, E, Martin, J., and B. Laumon (2003). Comparison of road crashes incidence and severity between some French counties. Accident Analysis and Prevention, Vol 35, pp. 537-547.

Andreassen, D., 1991. Population and registered vehicle data vs. road deaths. Accident Analysis and Prevention, 23(5), October, pp 343-351.

Bates, D. M., Watts, D. G., 1988. Nonlinear regression analysis and its applications, John Wiley & Sons, Inc., New York.

Beenstock, M. and D. Gafni, 2000. Globalization in road safety: explaining the downward trend in road accident rates in a single country (Israel). Accident Analysis and Prevention, 32, pp. 71-84.

Belsley, D., Kuh, E., and Welsch, R. (1980). Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. John Wiley and Sons, New York.

Broughton, J., 1991. Forecasting road accident casualties in Great Britain. Accident Analysis and Prevention, 23(5), October, pp 353-362.

Cameron, MH, Haworth, N, Oxley, J, Newstead, S, & Le, T., 1993. Evaluation of Transport Accident Commission road safety television advertising. Report No.52, Monash University Accident Research Centre.

Commandeur, J.J.F., and S.J. Koopman (2007). An Introduction to State Space Time Series Analysis. Oxford University Press.

COST329, 2004, Models for traffic and safety development and interventions. Final Report of the Action. European Commission, Luxembourg.

Dobson, A.J., 1990. *An Introduction to Generalized Linear Models*. Second edition, Chapman and Hall, London.

Dupont, E., H.Martensen, H.Stipdonk, F.Bijleveld, J.Commandeur, E.Papadimitriou, G.Yannis, C.Antoniou, R.Bergel, C.Brandstaetter (2009). "Analyses using the European Road Safety Observatory", Proceedings of the 4th IRTAD Conference, Seoul, September 2009.

Fisher, R. A. (1934), Two new properties of mathematical likelihood. Proceedings of the Royal Society A, 144, pp 285-307.

Gharaybeh, F. A., 1994. Application of Smeed's formula to assess development of traffic safety in Jordan, Accident Analysis and Prevention, Volume 26, Issue 1, pp. 113-120.

Gitelman, V., E. Doveh and S. Hakkert (2010). Designing a composite indicator for road safety. Safety Science, 48, pp. 1212-1224.

Harvey, A. C. and N. Shephard, 1993. Structural Time Series Models. In G. S. Maddala, C. R. Rao and H. D. Vinod, eds., Handbook of Statistics, Vol. 11, pp. 261-302, Elsevier Science Publishers, B. V.

Harvey A.C., 1994. Forecasting, Structural Time Series Models and the Kalman Filter, Cambridge University Press.

Henning-Hager, U., 1986. Urban development and road safety. Accident Analysis and Prevention. 18(2):135-45.

Hakim, S; Shefer, D; Hakkert, A S & Hocherman, I., 1991. A critical review of macro models for road accidents. Accident Analysis and Prevention, 23 (5), 379-400.

Hauer, E., Ng, J. C. N., and Lovell, J. (1988). "Estimation of Safety at Signalized Intersections", Transportation Research Record, 1185, Transportation Research Board, National Research Council, Washington, D.C., pp. 48-61.

Hollo, P., V. Eksler and J. Zukowska (2010). Road safety performance indicators and their explanatory value: A critical view based on the experience of Central European countries. Safety Science, 48, 1142-1150.

Jacobs, G. D., 1986. Road accident fatality rates in developing countries-a reappraisal. In: PTRC. Summer Annual Meeting, University of Sussex, 14-17 July 1986., Proc of Seminar H. London: PTRC Education and Research Services, pp. 107-119.

Koornstra, M. J. (1992) The evolution of road safety and mobility, IATSS Research, 16: 129-148.

Koornstra, M. J. (1997) Trends and forecasts in motor vehicle Kilometrage, road safety, and environmental quality, pp: 21-32 in Roller, D., (ed.) The motor vehicle and the environment – Entering a new century. Proceedings of the 30th International Symposium on Automotive Technology & Automation, Automotive Automation Limited, Croydon.

Kopits, E., and M. Cropper, 2005. Traffic fatalities and economic growth. Accident Analysis and Prevention, 37, pp. 169-178.

Lassarre, S., 2001. Analysis of progress in road safety in ten European countries. Accident Analysis and Prevention 33, pp. 743 – 751.

Ljung, G. M., and G. E. P. Box, 1978. On a measure of lack of fit in time series models. Biometrika, 65, 297-303.

Lord, D., 2002. Application of Accident Prediction Models for Computation of Accident Risk on Transportation Networks. Transportation Research Record: Journal of the Transportation Research Board 1784, pp. 17-26.

Lord, D., S. P. Washington, and J. N. Ivan (2005). Poisson, Poisson-gamma and zero-inflated regression models of motor vehicle crashes: balancing statistical fit and theory. Accident Analysis and Prevention 37, pp. 35-46.

Maher M.J. and I. Summersgill (1996). A comprehensive methodology for the fitting of predictive accident models. Accident Analysis and Prevention 28(3), pp. 281-296.

Maycock, G., and Hall, R. D. (1984). "Accidents at 4-Arm Roundabouts." TRRL Laboratory Report 1120, Transport and Road Research Laboratory, Crowthorne, Berkshire, UK.

McCullagh, P. and Nelder, J. A., 1989. *Generalized Linear Models*. Second edition. Chapman Hall, New York.

Miaou, S. (1994). The relationship between truck accidents and geometric design of road sections: Poisson versus negative binomial regressions. Proceedings of the 73rd Annual Meeting of the Transportation Research Board, Washington, D.C.

Mountain, L., Maher, M., and B. Fawaz (1998). The influence of trend on estimates of accidents at junctions. Accident Analysis and Prevention, Vol. 30, No. 5, pp. 641-649.

Muggeo, V. (2003) Estimating regression models with unknown break-points. Statistics in Medicine, 22: p.p. 3055-3071.

Muggeo, V. M. R. (2008). segmented: An R package to fit regression models with broken-line relationships. R News, 8(1):20-25, May. Available online: http://CRAN.R-project.org/doc/Rnews/Rnews_2008-1.pdf (Accessed: November 15, 2010)

Newstead, S; Cameron, M H; Gantzer, S & Vulcan, P., 1995. Modeling of some major factors influencing road trauma trends in Victoria 1989 - 93. Report No. 74, Monash University Accident Research Centre.

Nicholson, A., and Y-D. Wong (1993). Are accidents Poisson distributed? A statistical test. Accident Analysis & Prevention, Volume 25, Issue 1, February 1993, Pages 91-97

Oppe, S., 1989. Macroscopic Models for Traffic and Traffic Safety. Accident Analysis and Prevention, 21 (3), pp. 225 232.

Page, Y., 2001. A statistical model to compare road mortality in OECD countries. Accident Analysis and Prevention, 33, pp. 371-385.

Petridou, E., G.Yannis, A. Terzidis, N. Dessypris, E. Germeni, P.Evgenikos, N.Tselenti, A.Chaziris and I. Skalkidis (2009). "Linking Emergency Medical Department and Road Traffic Police casualty data: a tool in assessing the burden of injuries in less resourced countries", Traffic Injury Prevention, Vol. 10, No. 1, 2009, pp. 37-43.

Pindyck, R. S. and D. L. Rubinfeld (1997). Econometric models and economic forecasts. Fourth edition. Irwin/McGraw-Hill.

Qin, X., J. N. Ivan and N. Ravishanker, 2004. Selecting exposure measures in crash rate prediction for two-lane highway segments, Accident Analysis and Prevention, 36 (2), pp.183-191.

R Development Core Team, 2011. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org (accessed July 20, 2011).

Retting, R.A. and Kyrychenko, S.Y. (2001). Crash Reductions Associated with Red Light Camera Enforcement in Oxnard, California. Insurance Institute for Highway Safety, Arlington, VA.

Schwarz, G. (1978). Estimating the Dimension of a Model. Annals of Statistics. 6, pp. 461–464.

Smeed, R J., 1968. Variations in the pattern of accident rates in different countries and their causes. Traffic Engineering Control, 10(7), 364-371.

Stipdonk, H., P. Wesemann and B. Ale (2010). The expected number of road traffic casualties using stratified data. Safety Science, 48, pp. 1123-1133.

Twisk, D.A.M. and Stacey, C. (2007). Trends in young driver risk and countermeasures in European countries. Journal of Safety Research, Vol. 38 (2007), p.p.245–257

van Beeck, E. F, G. J. J. Borsboom, and J. P. Mackenbach, 2000. Economic development and traffic accident mortality in the industrialized world, 1962-1990. International Journal of Epidemiology, 29, 503-509.

Washington, S. P., Karlaftis, M. G., and F. L. Mannering (2003). Statistical and Econometric Models for Transportation Data Analysis. Chapman & Hall/CRC.

Wegman, F. and S. Oppe (2010). Benchmarking road safety performances of countries. Safety Science, 48, pp. 1203-1211.

White, D. J. and S. P. Washington (2001). Safety restraint use as a function of law enforcement and other factors, Transportation Research Record, Journal of the Transportation research Board, 1779: pp. 109-115.

Wood, G.R. (2002). Generalized Linear Accident Models and Goodness of Fit Testing. Accident Analysis & Prevention, Vol. 34, pp. 417-427.

Yannis, G., C. Antoniou and E. Papadimitriou (2007a). Road casualties and enforcement: Distributional assumptions of serially correlated count data. Traffic Injury Prevention, 8:3, pp. 300 - 308.

Yannis, G., E. Papadimitriou, and C. Antoniou (2007b). Multilevel modeling for the regional effect of enforcement on road accidents. Accident Analysis and Prevention, 39, pp. 818-825.

Yannis, G., E. Papadimitriou, and C. Antoniou (2008). Impact of enforcement on traffic accidents and fatalities: A multivariate multilevel analysis. Safety Science, 46, pp. 738-750.

Yannis, G., P.Evgenikos and A.Chaziris (2009). "CADaS - A common road accident data framework in Europe", Proceedings of the 4th IRTAD Conference, Seoul, September 2009.

Yannis, G., E.Papadimitriou, E.Dupont and H.Martensen (2010). "Estimation of fatality and injury risk by means of in-depth fatal accident investigation data", Traffic Injury Prevention, Vol. 11, No. 5, 2010, pp. 492-502.

Yannis, G., C. Antoniou, E. Papadimitriou and D. Katsochis (2011). When may road fatalities start to decrease? Journal of Safety Research, Vol. 42, Iss. 1, February 2011, pp. 17-25.