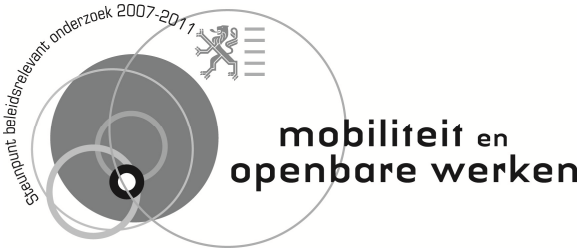
Time Series Models for Road Safety Accident Prediction

RA-MOW-2011-020

D. Karlis, E. Hermans

Onderzoekslijn Risicobepaling



DIEPENBEEK, 2012
STEUNPUNT MOBILITEIT & OPENBARE WERKEN
SPOOR VERKEERSVEILIGHEID

Documentbeschrijving

Rapportnummer:	RA-MOW-2011-020
Titel:	Time Series Models for Road Safety Accident Prediction
Auteur(s):	D. Karlis, E. Hermans
Promotor:	Prof. dr. Geert Wets
Onderzoekslijn:	Risicobepaling
Partner:	Universiteit Hasselt
Aantal pagina's:	27
Projectnummer Steunpunt:	6.1
Projectinhoud:	In dit project worden prognoses op vlak van verkeersveiligheid in Vlaanderen gemaakt. Een eerste stap hierin is de bepaling van een geschikte methodologie.

Uitgave: Steunpunt Mobiliteit & Openbare Werken – Spoor Verkeersveiligheid, februari 2012.

Steunpunt Mobiliteit & Openbare Werken
Spoor Verkeersveiligheid
Wetenschapspark 5
B 3590 Diepenbeek

T 011 26 91 12
F 011 26 91 99
E info@steunpuntmowverkeersveiligheid.be
I www.steunpuntmowverkeersveiligheid.be

Samenvatting

Titel: Tijdreeksmodellen voor verkeersongevallen predictie

Korte samenvatting:

In dit overzichtsrapport wordt een kritische blik geworpen op de variëteit aan prognosemodellen die er bestaan in de literatuur en welke bruikbaar kunnen zijn voor het maken van voorspellingen omtrent de evolutie in de tijd van verkeersongevallen. Er wordt hierbij de nadruk gelegd op gedisaggregeerde analyse, oftewel analyse gericht op verschillende subgroepen uit de verkeersveiligheidpopulatie. Naast de klassieke tijdreeksmodellen en modellen voor discrete waardes komen de moderne en krachtige state-space modellen aan bod. Deze blijken bijzonder geschikt om verder te gebruiken bij toekomstige modelberekeningen. Ook aspecten die te maken hebben met de beschikbaarheid van benodigde gegevens worden verder besproken waarna een algemeen besluit wordt gevormd met betrekking tot de selectie van geschikte prognosemodellen die toelaten om analyses omtrent verkeersveiligheid in de toekomst voor verschillende subgroepen in Vlaanderen uit te voeren.

English summary

Title: Time series models for road safety accident prediction

Abstract:

In the present report we aim at providing a critical view to some existing time-series models for road safety accident prediction in order to allow for a better picture of the wide scope of the problem and also to be able to criticize approaches. A special focus is given to models that allow for disaggregate analysis, i.e. analysis that focus on different subgroups of the road safety population. Apart from simple time series models and discrete valued models, the currently fashionable and powerful state space models are investigated in more detail as they prove to be a promising technique for further analysis. Some issues related to data availability are also discussed. Conclusions and general comments with respect to the selection of appropriate models to forecast accident rates for different groups in Flanders for the forthcoming years are given.

Inhoudsopgave

1. INTRODUCTION

Road traffic crashes are one of the world's largest public health and injury prevention problems. The problem has important consequences because the victims are overwhelmingly healthy prior to their crashes. A report published by the WHO (2009) estimated that approximately 1.3 million people die each year on the world's roads, between 20 and 50 million sustain non-fatal injuries and traffic accidents were the leading cause of death among children of 10 – 19 years of age.

Undoubtedly there is awareness in most societies about this issue and reducing the fatalities from road accidents is always on every political agenda. Also, the issue of traffic safety is high in the academic agenda and a lot of research is undertaken in order to examine and improve traffic safety issues. Lately, there was a downward trend in the number of fatalities in most countries in Western Europe, North America and Oceania (see Elvik, 2010, see also Lassarre, 2001), reflecting the awareness of the problems as well as all the measures undertaken to decrease it. However, apart from fatalities there is also a great concern for the public with respect to other types of non-fatal accidents as they also produce significant losses and thereby contribute to the economic costs. In Flanders, almost 40.000 casualties were registered in 2009 (FOD Economie, 2011).

All the above issues make the need for successful accident analysis tools obvious. Organizations at national and regional level can benefit from advanced models that can be used for various scopes of the accident analysis agenda like:

- Prediction of the future amount of unsafety and identification of groups at enlarged risk.
- Examination of the important factors that may lead to an increase/decrease of accidents.

Also, the evaluation of any safety measure taken by examining the pre- and post-period of a new measure is important for policymakers. However the focus in the present report is on prediction.

However accident prediction modeling is not easy and depends on various aspects, like the scope of the analysis, the available data and the level of application; several competing models are available and could be considered for an application. A variety of reasons illustrating the difficulty with this kind of analyses is given below. We would like to refer to some of them in order to show the level of the problem, later on we are going to discuss some of the issues in more detail and explain how problems can be overcome to a certain extent. Some of the reasons are the following:

- Inhomogeneous definition: for certain series the definitions of the quantities measured are not homogeneous, especially if data from different sources or countries are used.
- Data availability: in certain levels the data are far from being considered as complete. Elvik and Mysen (1999) reported that while data on fatalities can be trusted data on other categories like damage only accidents are by far underreported and

hence their completeness is questionable for any reasonable analysis if this factor is not accounted for.

- Various covariate information: while the literature contains many models that try to fit and forecast accident fatalities in the future, it also has a huge variety on the covariate information used and the level of availability of such information. On the one hand this limits the comparability of different models not allowing to clearly select one as a general purpose model (if such a model exists) but also complicates the usage of particular variables as drivers of measuring fatalities.

- Aggregation level: most of the available data are complete in an aggregated level but when we try to disaggregate them in different types of accidents or different areas this is difficult or even impossible. This implies that models developed for aggregated data are not necessarily valid for other categories/areas and hence special effort needs to be taken while interpretations should be cautious with respect to the disaggregated level.

- Small area estimation problem: if prediction of fatalities is to be made at the local level typically the data are sparse and not detailed enough to achieve reliable estimators. On the other hand, such data would be very helpful for evaluating local measures of safety. In such circumstances special models need to be developed which allow for small area estimation and prediction. Such models, typically, borrow information from adjacent areas and use more advanced statistical methodologies.

- Data discontinuity: for some data series it is evident that due to some policy measure there is a change in the time series data (i.e. a decrease (hopefully) in the number of fatalities). Hence the data show some kind of discontinuity that needs to be taken into account. Such an example, illustrating this point, is the downward trend in the fatalities that occurred after the energy crisis in 1973. This trend should be taken into account when data are analyzed and appropriate methods need to be used.

- Missing information: while a considerable improvement has been made towards the completeness of the data and in general the covariate information used, it is still possible that full data are not available and in some cases extrapolations or proxies need to be used. A typical example relates to the exposure which, while recognized as of particular importance, is still very difficult to have a clear number of the kilometers travelled or the time spent in traffic (at a detailed level). The number of vehicles or the consumption of fuels is then used as a proxy. Such proxies cannot be very detailed (for example consumption for small areas) and hence fail to provide sufficient information for example at a disaggregated level.

- Purpose of the derived model: in certain cases the model has merely monitoring purposes, it tries to identify the important factors that may explain the current situation or evaluate a safety decision taken at the past by looking in the data for possible changes in the pattern. Such a model while interesting itself does not have necessarily a forecasting potential if the model development has mainly

descriptive purposes. Hence various models developed, based on different scopes are not comparable.

- Different modeling approaches have been used to model accident fatalities across different countries. This implies that comparisons are not easy. For example, in different countries it is possible that the models used for prediction have different variables. Perhaps this has some theoretical reason (some variable while significant in some countries are not significant in other countries) but still problems of comparability across countries arise.

- Statistical methods used: finally, the literature has a huge number of different methodologies applied to such data. While it is not the purpose of this section to criticize or not these methods, we put emphasis on the fact that from a statistical point of view it is hard to assume that one method fits in all cases and hence special effort is needed for any particular data set. Also note that the level of information makes some models applicable or not, e.g. when examining a small area where the fatalities are small counts, time series models based on normal random variables are meaningless. Note that if we treat data at a large aggregation level, e.g. the country level, then reasonably enough one may ignore the discreteness of the data and fit normal based models.

Hakim et al (1991) provided a wide review on macro models for accident prediction. They also provided a long list (up to that date) of empirical studies including a comprehensive table about the variables that were considered for creating the models. The variables concerned demographics, economics variables, driving behavior variables and also dummies to allow for intervention effects. Of course after that time (1991) more studies have appeared increasing the level of "candidate" covariates to be used in an accident prediction model. To conclude, there is a wide range of covariate information in which the number of variables increases. In recent days, much more variables have been considered mainly accounting for different effects like weather conditions, geometrical details of the road and other.

Hakim et al (1991) distinguished between 3 main avenues to model accidents and create accident prediction models. The first type of model is time series models where data from one area are considered across time perhaps with the available covariate information. The second category refers to panel data where more areas are considered. Such an approach allows for examining covariates related to the area and hence creates more detailed models. Availability of panel data of course implies much more detailed data and also implies more refined modeling because effects may disappear if they are not present to all the areas. The third avenue is the combination of time series and panel data. In this report we focus on time series models since the interest lies on models for one country (more specifically one region) only.

In general, three important aspects of any accident forecasting approach can be stated; they relate to accident exposure, accident risk and accident severity. Exposure relates to the level of participation in traffic. That way, people are exposed to the risk of

having an accident. As mentioned above, exposure is considered as a dominant variable in order to examine accident data. A natural measure of exposure would be the mileages, i.e. the total kilometers travelled (or driven by cars on a given road segment of interest), but this is not available in most of the cases, especially if smaller subgroups are examined. Hence, several other proxies are used in literature. The SafetyNet (2008) project funded by the EU tried to list and qualify exposure candidates. Modeling the exposure variable is an important task for accident forecasting models. Secondly, it is important to define and work with appropriate risk measures in order to quantify the importance of the accident analysis. Typical risk measures are the number of accidents per kilometres travelled or the number of accidents per population. Finally, the third important component is the severity of the accident. Given the fact that an accident occurred, the severity of the casualties can be taken into account. The number of severely (or fatally) injured persons per accident can then be considered.

In the present report we aim at providing a critical view to some of the existing models so as to allow for a clear picture of the wide scope of the problem. This report serves mostly as a guide to the selection of models for predicting accidents in Flanders for the period up to 2020, so some selection on the description of methods is made based on the data available for this study. It is not easy to include all kind of models, we have tried to include the most important ones. We focus on dynamic models like state-space models as we found them as interesting candidates for the application at hand. State–space models are very flexible as they account for trend, covariate information while at the same time they allow the effects to change dynamically.

Throughout the report, we will try to keep the mathematical details in a low level since such details can be found in the papers cited. More detailed reviews can be found in COST 329 (2004) and OECD (1997).

The next section will introduce time series models. We provide a general framework, later on we describe discrete valued models that are recently popular. Next we discuss problems related to disaggregate analyses, i.e. analyses that focus on subgroups, and we show the kind of problems such approaches have. Then we present the currently fashionable and powerful state space models in section 4. Some issues related to data availability are discussed in Section 5. Conclusions and general comments with respect to the selection of appropriate models to forecast accident rates for different groups in Flanders for the forthcoming years can be found in Section 6.

2. TIME SERIES MODELS

Most accident data are in a time series form, i.e. we have available observations in consecutive time intervals, like every day, every month or even on an annual basis. Time series modelling is a well understood and studied area in most of the disciplines and a wide range of time series models have been developed and applied to real data from several disciplines.

A typical time series model assumes certain different effects like:

- Trend, the general trend of the data (upwards or downwards for example),
- Seasonal effects (for example when examining monthly fatal accident counts certain months have a larger frequency),
- Cycles which are periodic effects not captured by the seasonal effects, typical for example in an economic model where economic cycles are present,
- Covariate effects, i.e. some other variables that affect the variables we examine,
- Random error term that creates the uncertainty around a hypothesized model.

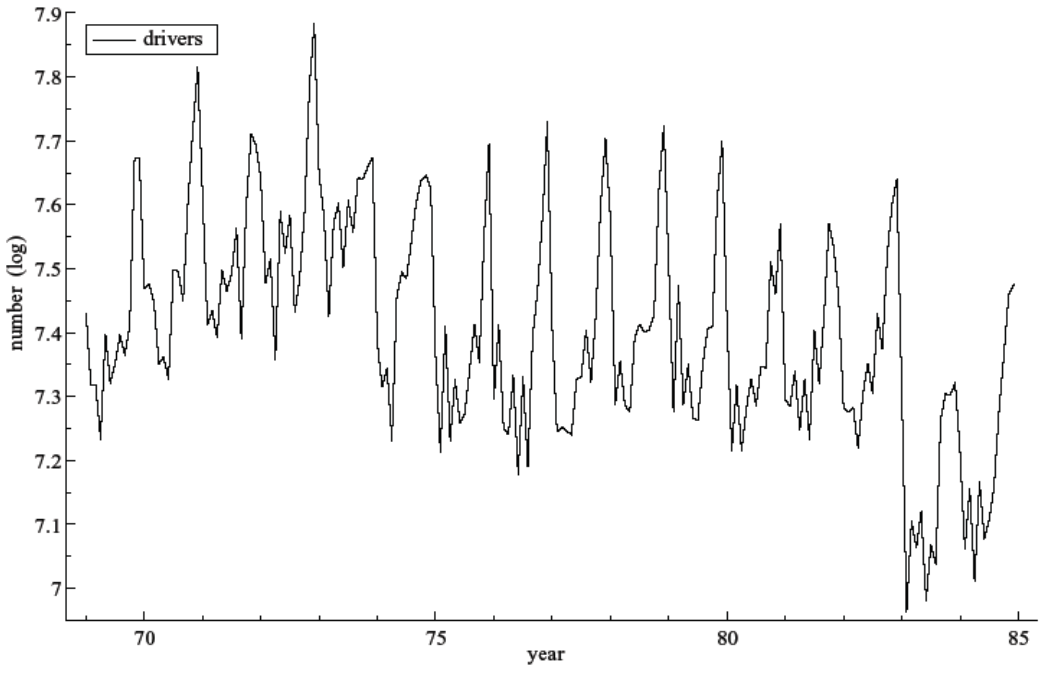


Figure 1. Monthly numbers (logged) of car drivers who were killed or seriously injured in road accidents in Great Britain. (Reproduced from Fig 1.1 from Harvey et al, 2004)

Figure 1 is a typical time series for accidents. One can see that examining the whole period from 1970 to 1985 there is a downward trend. Within each year we can see the repeated (more or less) pattern which is the seasonal effect. At the beginning of 1983 we can see that the series jumped to a lower level. This is due to the effects of the seat belt law enforced in January 1983. Such effects can be taken into account by appropriate usage of covariates (a dummy in this case). The remaining fluctuation can be considered as the random error term. Note that in this example there is no cyclical effect since it is not common in accidents.

Existing time series models try to separate such effects and allow for prediction of forthcoming time periods. Classical textbooks on time series models can be used to check for the type of models available, we do not repeat this here. The interested reader can consult, for example, the books from Chatfield (2003) and Brillinger (1975)

A natural question that arises is why accident counts should be correlated across time. Based on a Poisson process, accidents should occur randomly in time. However this is a rather simplistic assumption. Drivers share the same environment every day, road conditions are the same day by day and also driving behaviours do not change in short time intervals. Thus we expect that accident generating mechanisms are the same in successive time periods leading to correlated accident counts.

2.1 Continuous data

Various models were considered and applied to accident data: Spline models or any other general trend models, Box-Jenkins (ARIMA) models, and DRAG-type structural explanatory models. Spline models can be used to detect the overall trend of casualty series without considering a specific functional form as for example a linear trend model does. The short-term predictive capability of the ARIMA model is shown to be rather good. The DRAG-type models can investigate the influence of social and mobility factors on the development of safety (see, e.g. Van den Bossche and Wets, 2003, Van den Bossche, 2006). It has been found once again that the choice of the most appropriate type of model will depend upon the policy context (in the sense of the kind of problems that are treated) and data availability. There are great differences between countries with respect to the time series models that have been applied. This results from the variety of national exposure data, scientific backgrounds and the particular aspects of road safety that have been of interest. Thus, comparisons are difficult with these types of models.

One of the first models developed is the one proposed by Smeed (1949) where the number of deaths in car accidents was modelled as:

$$D_t = AN_t^{\ a} P_t^{\ b}$$

where D_t is the number of dead persons at time *t*, N_t the number of registered vehicles at time *t* and P_t the population size at time *t*. The model actually assumes what is still

valid, i.e. that exposure is the most important factor. Parameters A, α and b were estimated. This law was then used in certain countries with great success and still now this kind of relationship which can be also put in a log scale as

$$\log D_t = A' + a \log N_t + b \log P_t$$

is used as a basis for more complicated models. This model is not explicitly a time series model as current observations are not related to previous ones but since the covariates (population and vehicle numbers) are time series, autocorrelation is present. Figure 2 presents the model fitted to data from the UK from the original paper of Smeed (1949)

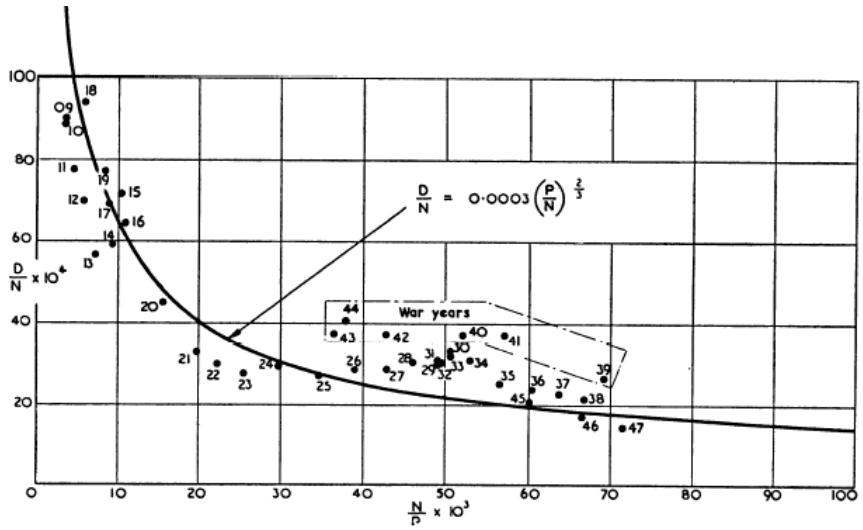


FIG. 2A

Relation between Number of Deaths per 10,000 Registered Vehicles and Number of Vehicles per 1,000 Population for Great Britain, Period 1909-1947. (The curve shown is the same as that shown in Fig. 1a. It is not a curve of best fit for the points of this diagram)

Figure 2: Smeed's law. Figure taken from the original paper.

After these basic models a lot of development of models took place. Most of the models assumed a normal random error term ignoring the discrete nature of the data. This can be attributed to the fact that continuous data time series are much better developed than integer valued time series and since most of the applications treated aggregated data, a normal approximation was reasonable.

Following COST 239 (2004), the DRAG1 model (Gaudry 1984) did an important step forward by considering, (i) a substantially extended set of explanatory factors, (ii) a multi-equation modelling approach, in which the severity and the underlying amount of exposure were treated as endogenous variables to be explained, and (iii) an estimation technique allowing for estimably flexible (non-linear) functional forms for several dependent and independent variables. This technique, due to Liem et al (1993), is known under the acronym BC-GAUHESEQ (Box-Cox Generalised AUtoregressive HEteroskedastic Single EQuation estimation). Later modelling efforts within the DRAG tradition include a German model (Gaudry and Blum 1993), a French model (Jaeger and Lassarre 1997), a

local Swedish model (Tegnér and Loncar-Lucassi 1997), and a Norwegian model (Fridstrøm 1999).

A detailed description of all the time series models applied to accident data is beyond the scope of the present report. However note that typically accident counts do not show complicated autocorrelation structures and hence rather simple (from the time series point of view) models are sufficient, like autoregressive and moving average models. State-space models are reviewed in a later section.

2.2 Discrete models

The discrete nature of the data has been recognized early in the literature. Especially, when particular types of accidents are considered or the areas or time of interest is small, then the data at hand are small integer numbers and it is not suitable (and for many reasons not reasonable) to apply time series models for normal/continuous data in that case.

For example, if the observed data are small counts with an excess of zero values, normal models may fail considerably as they, for example, give positive probability to negative counts or since the data are typically skewed. This can be seen in Figure 3, which shows the frequencies from daily data on the number of fatal accidents (Brijs et al, 2008). For the upper left graph, the number of accidents is large and hence some continuous approximation would make sense. However note that the data are skewed and the normality assumption is questionable. For the other two series with small counts the deviation from normality is apparent.

Note that this depends on the nature of the data. When considering subgroups it is reasonable that we may have small counts and hence this family of models is more appropriate to use. It also depends on the time scale which is used. For daily accidents it is reasonable that the counts are small, aggregated yearly data may have a different behaviour. Thus in this report we describe different models with potential usage rather than already limiting ourselves to a particular class of model (as both aggregated and disaggregated data will be used).

Next we describe the two most important classes of models in this framework:

- Poisson regression models which assume that the number of accidents follow a Poisson distribution with a mean which is related to some covariate information, typically in the log scale, since the parameter of the Poisson distribution is positive. There are several applications of such models (see e.g. Fridstrøm and Ingebrigtsen 1991, Brännäs and Johansson 1992). A potential limitation of the model is the fact that the Poisson distribution assumes equality between the variance and the mean which is a rather restrictive assumption for real data. Note, however, that this cannot be seen easily from raw data since the introduction of covariates in fact assumes that the equality of the variance and the mean is conditional on the same covariate information, which is not easy to be

checked before the model is fitted to the data and a residual analysis is done. This is a common fallacy in certain applications. The choice of covariate information is a delicate task. Most of the existing models aim at examining rather certain potential covariates rather than creating a global model.

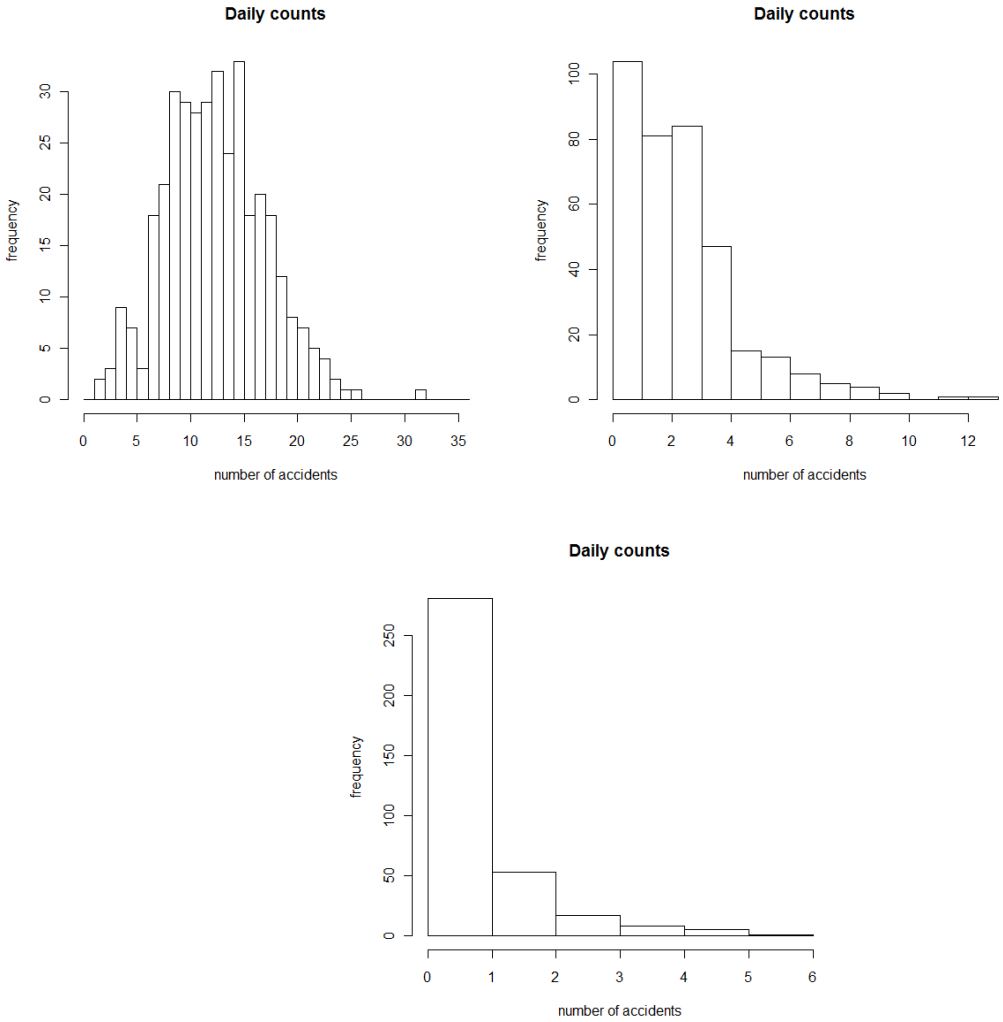


Figure 3: Frequencies of daily counts from three different areas. Data taken from Brijs et al, (2008)

- A next step, if the Poisson regression model fails, is a negative binomial regression model. First of all note that different models are used under the same name (see Cameron and Trivedi, 1998). The negative binomial model has an extra overdispersion parameter which allows the variance to exceed the mean, which is the limitation mentioned above for the Poisson regression model. Fitting negative binomial regression models is easy as several packages offer this capability. Again, covariates are assumed to the log scale of the mean, while one can select to add covariates to model the variance of the model as well. The negative binomial has a larger variance than the Poisson model with the same mean and hence typically has larger tails and a larger probability of a zero count.

Applications of a negative binomial regression model for accidents are given by Washington et al. (2005, see also the references therein). It is imperative to acknowledge that if a Poisson model is assumed while ignoring the overdispersion this may lead to erroneous conclusions. The reason is that since the assumed variability is smaller than the true one the standard errors derived underestimate the real ones and hence covariates may be found significant (since we divide by a smaller quantity when using a Z-score test). Hence one has to be cautious in order to select appropriately the underlying model in order to avoid erroneous results.

These two models, the Poisson regression and the negative binomial model, are perhaps the most widely used ones. Several alternatives/improvements have been proposed however.

Zero inflated Poisson and negative binomial models have been fitted in order to account for the excess of zeros found in several data sets (see Washington et al, 2005). Zero inflated models assume the existence of some additional component which has always zero accidents, in practice we assume that some sites have zero accidents for some reason. Such models have been criticized as not being reasonable to model the generating mechanism of accidents (see Washington et al, 2007) however they can have several other interpretations that make such model plausible with real data, especially for small areas or time frames, where an excess of zero accidents is observed.

In addition, several other models have been considered. Like the generalized Poisson regression models which allow for both over- and underdispersion with respect to the Poisson model. Note that underdispersed data sets are rather rare in practice and typically they may reflect some underreporting mechanism. To this direction, recently Lord et al (2008) proposed a Conway-Maxwell Poisson regression model which can also fit both over- and underdispersion situations while Lord and Geedipally (2011) propose a Negative Binomial-Lindley Distribution model. It is evident that several other distributions could be fitted for accident data.

In all the above, the time series framework can be implicitly introduced in the model by considering lagged observations. Discrete valued time series models are of increased interest especially the last years. There are two broad categories of such models: the observation driven model and the parameter driven ones (see Cox, 1981). The key difference is that in the parameter driven model the time correlation is introduced by a latent unobserved time series process in the parameter space, in fact they are a special case of state-space models (see later). Such models while they borrow strength from the corresponding time series models for continuous data are hard to work with without additional relaxing assumptions. The second category is the observation driven model where the current observation is directly related to the previous one(s). For some applications of such models we refer to Brijs et al (2008) and Quddus (2009).

2.3 Other models

The range of potential models for accident prediction is wide and it would be difficult to mention all these models in this report. For example, there are applications where more than one type of accidents is considered together and they have been jointly modelled (see Park et al, 2007 or the work of Brijs et al, 2007). Hierarchical Bayesian models have also been proposed for modelling motor vehicle collisions (Schluter et al., 1997, Tunaru, 2002). Also other approaches in the boundary between statistics and artificial intelligence have been applied like support vector machines (Li et al, 2008), Bayesian networks (Xie et al, 2007) or other artificial intelligence models (see Mussone et al, 1999 and Abdelwahab and Aty, 2002).

3. DISAGGREGATE MODELS

In most situations, data are aggregated in some extent, while existence of disaggregated data, i.e. data at a lower, more detailed level, are not available, or if they are they cover small time periods and hence they are not easy to use.

It is acceptable that while aggregate models can be used to describe general trends for some country (i.e. increasing or decreasing rate of accident casualties), they are unable to represent changes in specific parts of the transport system or in the safety of subgroups of road users. So, while the general picture can be captured it is difficult to have a more detailed view focusing on specific groups of road user modes, or type of roads, or even age categories. From the policy point of view such approaches would be more interesting as many of the safety measures undertaken focus on particular subgroups and not necessarily on the entire range of road users (or at least we expect to see larger effects in some subgroups which may disappear or cancel out if the whole range of users is considered). It is therefore desirable to be able to analyse the current and future states of mobility and safety for each subgroup. Such approaches can be based on disaggregate modelling.

For example, disaggregation of vehicle population data (by age and type), fuel consumption (by fuel sales by type) and population (by age and sex) are also commonly available annually, but details of traffic volume with respect to road type, vehicle type and distance travelled by mode are generally difficult to obtain.

Accident data are now available in great detail in most countries, and over a long time period (dating back to 1970s in most cases). Their reliability, however, decreases with the level of disaggregation, and there are also problems of underreporting of non-fatal accidents. Changes over time in the rate of underreporting cause significant issues for the analyses.

Examples of disaggregated data analysis refer to examining the evolution in the accident or risk pattern of different transport modes, different age categories (i.e. young drivers versus elderly drivers), different genders, different road types (motorways, regional roads and local roads) or even examining particular areas of interest (e.g. crossroads, specific patches of a highway). As one can see, detailed accident data for some of the above categories exist but, data with respect to the covariates are not available at this detailed level. As an example, even if data are available for patches of roads, it is not easy to have the exposure data available also, and perhaps some more refined type of modelling is needed where the exposures in a disaggregated level are also estimated using some auxiliary model for example.

Disaggregate models are the tools for assessing different policy options, setting goals for safety programmes and predicting future safety developments at the disaggregated level. The explanation of accidents, in terms of accident causation, however, should take place with more precise and detailed models.

Disaggregate modelling suffers from some problems like:

- Scarcity of available data: typically, the available data are aggregated in some extent and hence it is hard to find reliable data so detailed in order to be able to look for small categories. This aggregation has to do with issues related to the way the data are collected (typically through official statistics resources) but also due to lack of appropriate definitions that are needed in order to classify appropriately according to subgroups. Moreover, data related to exposure are hard to find. As we have mentioned, typically exposure is measured through proxies and it is hard to disaggregate such proxies to the level that one would need in order to fit reasonable models.

- Need for specific methodologies including special design in the collection of the data: typically, data in subgroups have a lot of zero or small counts, they suffer from some kind of underreporting and hence one may need special methodologies in order to handle these cases. This also implies that general methodologies applicable to the aggregated data due to their size are not available (e.g. typically with aggregated data the counts are large and hence assuming a continuous model is not a bad idea, but for small counts this can lead to large problems). Moreover, working with small data sets implies higher standard errors and hence prediction becomes much more difficult. Finally, disaggregated analysis aims at examining the effect of certain policies to specific subgroups and hence in most cases special designs (e.g. need of controls) must be taken into account making the analysis even more difficult.

- Behavioural considerations generally play an important role in these modelling techniques since the level of analysis is focused on subgroups and hence particular information may be useful.

In case of lack of disaggregated data one may work in the aggregated level and then based on particular (and typically strong) assumptions disaggregate the forecasts to lower levels. For example, Broughton (1988) derived forecasts based on an aggregated level. Then this total number was split in subgroups based on forecasts for the trends of different subgroups (types of road users).

Forecasting in the case when disaggregated data are available is straightforward. However, even in this case the detailed data set may not be complete and hence extrapolation or smoothing approaches may be needed.

4. STATE SPACE MODELS

When modelling time series fitting dynamic models, i.e. models where the parameters may change over time, is commonly used. There are two main classes of univariate dynamic models: ARIMA models studied by Box and Jenkins and unobserved component models which are called structural models, by Harvey and Shephard (2003). In a structural model each component or equation is intended to represent a specific feature or relationship in the system under study. State space methods described in this section, belong to the latter group of models.

State space time series analysis began with the path breaking paper of Kalman (1960) and early developments in the subject took place in the field of engineering.

The state space model in its simple form can be expressed as

$$y_t = Z_t a_t + \varepsilon_t, \quad \varepsilon_t \sim N(0, H_t),$$
$$a_{t+1} = T_t a_t + R_t \eta_t, \quad \eta_t \sim N(0, Q_t),$$

with initial value

$$a_1 \sim N(\alpha_1, P_1),$$

where matrices Z_t, H_t, T_t, R_t and Q_t are assumed known (however this assumption can be relaxed). The key idea of state space models is that a certain parameter a_t relates to the parameter at the previous time point, inducing a dynamic linear model. The first equation is called the observation equation and the second equation is called the state equation.

In fact state space models generalise regression models so that the parameters can vary over time. The first equation is simply a linear regression equation. The state equation introduces a structure and time series component at the same time since parameters change dynamically across time. State space models are also called dynamic linear models.

A typical time series may be decomposed in a trend, a seasonal and an irregular part. An important characteristic is that the components are stochastic. Moreover, explanatory variables can be added and intervention analysis carried out. The principal structural time series models are therefore nothing more than regression models in which the explanatory variables are functions of time and the parameters are time-varying. The key to handle structural time series models is the state space form, with the state of the system representing the various unobserved components. Once in state space form, the Kalman filter may be applied and this in turn leads to estimation, analysis and forecasting.

The state space formulation for time series models is quite general and encompasses most of the classical time series models like MA and ARIMA models for example. Hence state space model is quite general. Also since the state equation(s) can

capture in a very flexible way the behaviour of the underlying (and unobservable) variables it offers great flexibility with real data.

The advantages of state space modelling can be summarized as:

- The basic advantage of state space models is that it is based on a structural analysis of the problem at hand. The different components that may comprise a time series model, like for example, the trend, the seasonal effect and the cycle, can themselves be modelled separately, offering great flexibility but also allowing to see more in detail their effect in the final time series. Hence the researcher can identify the components that (s)he really needs in the model and create a model that better fits the case under investigation.

- Moreover, the state space model offers greater generality. In fact, several other models can be seen as special case of the state space models. Covariates can be easily added to the model in a clearly interpretable way while they can vary stochastically over time.

- In fact, state space models satisfy the Markovian property and hence the necessary calculation can be put in a typical recursive manner. This also implies that the added computational difficulty is not large and it is in any case handable. Most well-known statistical packages offer state space models. Also the univariate theory can be extended to cover the multivariate case in a neat way. Even complicated multivariate state space models can be fitted rather easily.

- Forecasting with state space models is relatively easy and simple. State space models in fact apply some smoothing in the data and hence forecasts are also smooth. Also diagnostic checking is simple as the Kalman filter employed provides such a framework.

- State space models are adaptive and the benefits of this are usually realised by implementing them in real time since only minor calculations are needed. Hence, they have found tremendous applications to forecasting in real time situations. They also can be adapted by minor changes to create more complicated structures like nonlinear models for example.

- Finally, they offer great flexibility as they can be used in certain circumstances, allowing for refined modelling in several problems. Other models are special cases while certain components of a state space model can be modelled separately adding a wide range of possible models that can be fitted to the data and also allowing testing several research hypotheses for the underlying structure of the data. State space models offer the ability to easily handle systems with multiple inputs and outputs in a reasonable and simple way and hence they allow for modelling rather complicated problems.

At the same time, some disadvantages should be mentioned. The models are usually more complicated and less interpretable than standard time series models, especially for non-treated researchers making their acceptance in some problems not easy. In

addition, some added computational effort is needed with respect to much simpler models and hence the lack of their implementation in some widely used statistical packages (like SPSS) make a lot of researchers less reluctant to use the models. Finally, note that while for certain models state space modelling is well established and easy to use, there are models where it is not so easy, like for example discrete valued time series models. The model developed by Zeger (1988) is in fact a state-space model for modelling discrete time series. However, assuming a Poisson distribution leads to rather complicated recursion for the state equation and makes estimation difficult.

State space models are currently popular models for accident prediction mainly due to their generality and flexibility (see e.g. Gould et al 2004, Hermans et al 2006a, 2006b, Bijleveld 2008). Several software packages (like R, EVIEWS, MATLAB, SAS just to name a few) are available for fitting such models removing computational difficulties. State space models provide a convenient and powerful framework for analyzing time series data. More details can be found in several textbooks devoted to such models, see e.g. Durbin and Koopman (2001) and Commandeur and Koopman (2007).

5. DATA AVAILABILITY

When modelling accidents, one of the very difficult problems relates to data availability. For example, while it is considered that in general data related to fatal accidents are reliable in most countries, for less severe accidents it is known that a certain amount of underreporting is present.

This lack of complete data is only one issue related to data availability. A second one relates to the availability of certain covariates that are considered of particular interest for accident forecasting models, like the exposure. Exposure defined as the number of kilometres travelled by the population of interest is typically not available or at least not in the full form that a researcher would like to have. Such difficulties are well described in literature (see e.g. Joly et al 1991, Qin et al 2004, among others).

Even though exposure is recognized as playing a very important role to the development of accident models, its measurement is neither easy nor cheap. Data concerning proxies like the number of vehicles, fuel sales, road length and population size are available in most cases so in many cases in literature, a proxy is used. Thus all models considered are vulnerable on this "assumption" that the proxy measures well the variable that we aim to measure.

Another issue on the data availability refers to the level at which they are available. We have discussed issues about disaggregation of the data. We also emphasize as mentioned above that the level of detail in the data is important for the selection of an appropriate methodology to work with.

6. CONCLUSIONS

The literature contains a huge variety of different models that can be used as accident prediction models. Their usage depends on the scope of the analysis as well as the availability of the data. For the latter, it is important to separate models that aim at examining the effect of some covariates on the accident counts and models created in order to predict future accidents/accident rates. Models that succeed in the former are not necessarily good for the latter. Also covariates that are significant are perhaps not so helpful for prediction as well. Note also that in order to predict future accidents one may need to have the future values of the predictor covariates available, which is an elaborated task and typically not an easy one. Thus the scope of the modelling is important for selecting the appropriate model to work with.

Some more issues with respect to accident prediction modelling are important for developing a successful model although often ignored in practice.

Goodness of fit of the derived model is important. In many published works it has been ignored to report how well the fitted model fits the data although the model is used for arguments and predictions. In a similar fashion, when developing a prediction model the prediction ability of the model needs to be checked carefully typically with out-of-sample predictions.

The existing literature has a large number of developed models which also contain a large number of potential covariates. As time passes more and more data are available and hence more candidate covariate information is available. This creates some new challenges related to variable selection problems. Model parsimony arises as an important aspect of the accident prediction modelling, where covariates measuring similar things need to be removed from the final model. The selected models then are not comparable across different surveys something that is typically ignored. We also mention the model uncertainty issue described in Draper (1995), which emphasizes that using only the selected model for inference ignores the variability present due to the model selection procedure.

Another point that needs to be kept in mind is that nowadays more and more sophisticated models are available that take into account the nature of the data as well as the mechanisms that may have generated them. This invalidates to a great extent models used in the past that could not capture all the features of the data. As an example, the usage of continuous time series models when the accident counts are too small can be dangerous. In recent years, there are several models (and perhaps more models will be developed in the forthcoming years) to better tackle such data.

In this report, we tried to report on a wide range of models existing in the literature in order to capture the general framework. Among them, state space models are promising as they generalize most of the existing models in the literature while they are becoming more and more available in standard packages, including also that their interpretation is now easier and well understood. In the next step, the state space methodology will be applied to Flemish time series data consisting of accident and

casualty data, exposure data and possibly other covariate data, both at the aggregate and disaggregate level in order to identify the groups (i.e. transport modes, age classes and road types) that are expected to have a relatively high risk in the future and therefore need special attention.

7. REFERENCES

1. Abdelwahab, H.T. and Abdel-Aty, M.A. (2002). Artificial neural networks and logit models for traffic safety analysis of toll plazas. *Transportation Research Records*, 1784, 115–125.
2. Bijleveld, F. (2008). Time series analysis in road safety research using state space methods. PhD Thesis, SWOV.
3. Brännäs, K. and Johansson, P. (1992). Time series count data regression. Umeå Economic Studies no 289, University of Umeå.
4. Brijs, T., Karlis, D, Van der Bosche, P and Wets, G. (2007). Bayesian Ranking of accident sites. *Journal of the Royal statistical Society, Series A*, 170, 1001-1017
5. Brijs, T., Karlis, D. and Weets, G. (2008). An Integer Autoregressive Model to Study the Effect of Weather Conditions on Daily Car Accident Counts. *Accident analysis and Prevention*, 40, 1180-1190.
6. Brillinger, D. R. (1975). Time series: Data analysis and theory. New York: Holt, Rinehart. & Winston.
7. Broughton, J. (1988). Predictive models of road accident fatalities. *Traffic Engineering and Control*, 29, 296-300.
8. Cameron, C. and Trivedi, P. (1998). *Regression Analysis of Count Data.* Econometric Society Monograph No.30, Cambridge University Press.
9. Chatfield, C. (2003). *The Analysis of Time-series: An Introduction.* 6th edition, CRC Press.
10. Commandeur, J.J.F. and Koopman S.J. (2007). *An introduction to state space time series analysis*. Oxford: Oxford University Press.
11. COST 329. (2004). Models for traffic and safety development and interventions (Final report of the action EUR 20913). Brussels, Belgium: Directorate General for Transport, European Commission.
12. Cox, D. R. (1981). Statistical issues in the analysis of time series: some recent developments. *Scandinavian Journal of Statistics*, 8, 93–115.
13. Draper, D. (1995) Assessment and Propagation of model uncertainty. *Journal of the Royal Statistical Society, Series B*, 57, 45-97.
14. Durbin, J. and Koopman, S.J. (2001). *Time series analysis by state space methods*. Oxford: Oxford University Press.
15. Elvik, R. (2010). The stability of long-term trends in the number of traffic fatalities in a sample of highly motorized countries. *Accident Analysis and Prevention*, 42, 245-260.
16. Elvik, R. and Mysen, A.B. (1999) Incomplete Accident reporting: meta-analysis of studies made in 13 countries. *Transportation Research Records*, 1665, 133-140.
17. FOD Economie (2011). Verkeersongevallen- en slachtofferdata in Vlaanderen.
18. Fridstrøm, L. and Ingebrigtsen, S. (1991). An aggregate accident model based on pooled, regional time-series data. *Accident Analysis and Prevention* 23, 363-378.
19. Fridstrøm, L. (1999). Econometric models of road use, accidents and road investment decisions. PhD dissertation, Institute of economics, University of Oslo, April 1999.
20. Gaudry, M. and Blum, U. (1993). Une présentation brève du modèle SNUS-1 (Straßenverkehrs- Nachfrage, Unfälle und ihre Schwere). Modélisation de l'insécurité routière. *Collection Transport et Communication*, 47:37-44, Paradigme, Caen.
21. Gaudry, M. (1984). DRAG, un modèle de la Demande Routière, des Accidents et de leur Gravité, appliqué au Québec de 1956 à 1982. *Publication 359, Centre de Recherche sur les Transports (CRT), Université de Montréal*.
22. Gould, P.G., Bijleveld, F.D and Commandeur, J.J.F. (2004). Forecasting road crashes: a comparison of state space models. *Paper presented at the 24th International Symposium on Forecasting, 4-7 July 2004, Sydney, Australia*.

23. Hakim, S., Shefer, D., Hakkert, A. S., & Hocherman, I. (1991). A critical review of macro models for road accidents. *Accident Analysis and Prevention,* 23, 379-400.
24. Harvey, A. C. and Shephard, N. (1993). Structural Time Series Models. *Handbook of Statistics 11, 261-302*.
25. Harvey, A.C., Koopman, D. and Shephard, N. (2004). *State space and unobserved component models: theory and applications*. Oxford University Press.
26. Hermans, E., Wets, G. and Van den Bossche, F. (2006a). Describing the Evolution in the Number of Highway Deaths by Decomposition in Exposure. Accident Risk and Fatality Risk, *Transportation Research Record,* 1950, 1-8.
27. Hermans, E., Wets, G. and Van den Bossche, F. (2006b), The Frequency and Severity of Belgian Road Traffic Accidents studied by State Space Methods. *Journal of Transportation and Statistics*, 9,.63-76.
28. Jaeger, L. and Lassarre, S. (1997) Pour une modélisation de l'évolution de l'insécurité routière. Estimation du kilométrage mensuel en France de 1957 à 1993: méthodologie et résultats. *Rapport DERA no 9709, Convention DRAST/INRETS, Strasbourg/Paris*.
29. Joly, M.F., Joly, P. , Bergeron, J., Desjardins, D., Ekoe, J.M., Ghadirian, P., Gravel, S., Hamet, P., Laberge-Nadeau, C. (1991). Exposure to the risk of traffic accidents, a fundamental epidemiological parameter, and one difficult to measure. *Rev Epidemiol Sante Publique*, 39(3), 307-13.
30. Kalman, R.E. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering,* 82, 35–45.
31. Lassarre, S. (2001). Analysis of progress in road safety in ten European countries. *Accident Analysis and Prevention*, 33, 743-751.
32. Li, X., Lord, D., Zhang, Y. and Xie, Y. (2008). Predicting Motor Vehicle Crashes using Support Vector Machine Models. *Accident Analysis and Prevention*, 40, 1611-1618.
33. Liem, T., Dagenais, M. & Gaudry, M . (1993). LEVEL: the L-1.4 program for BCGAUHESEQ regression - Box-Cox Generalized AUtoregressive HEteroskedastic Single EQuation models. *Publication 510, Centre de recherche sur les transports, Université de Montréal.*
34. Lord, D., and Geedipally, S.R. (2011). The Negative Binomial-Lindley Distribution as a Tool for Analyzing Crash Data Characterized by a Large Amount of Zeros. *Accident Analysis and Prevention*, 43, 1738-1742.
35. Lord, D., Guikema, S.D. and Geedipally, S. (2008). Application of the Conway-Maxwell-Poisson Generalized Linear Model for Analyzing Motor Vehicle Crashes. *Accident Analysis and Prevention*, 40, 1123-1134.
36. Mussone, L. , A. Ferrari and M. Oneta, (1999). An analysis of urban collisions using an artificial intelligence model. *Accident Analysis and Prevention*, 31, 705–718.
37. OECD (1997). Road safety principles and models: review of descriptive, predictive, risk and accident consequence models (OECD/GD(97)153). Paris, France: OCDE-OECD.
38. Park, B.-J., Lord, D. and Hart, J. (2010). Bias Properties of Bayesian Statistics in Finite Mixture of Negative Regression Models for Crash Data Analysis. *Accident Analysis and Prevention*, 42, 741-749.
39. Park, E.S., and Lord, D. (2007). Multivariate Poisson-Lognormal Models for Jointly Modeling Crash Frequency by Severity. *Transportation Research Record* 2019, 1-6.
40. Qin, X., Ivan, J.N and Ravishanker, N. (2004). Selecting exposure measures in crash rate prediction for two-lane highway segments. *Accident Analysis and Prevention*, 36, 183–191.
41. Quddus MA. (2009). Time series count data models: an empirical application to traffic accidents. *Accident Analysis and Prevention*, 40, 1732-41.
42. Schluter, P.J., J.J. Deely and Nicholson A.J.,(1997). Ranking and selecting motor vehicle accident sites by using a hierarchical Bayesian model. *Statistician*, 46, 293–316.

43. Smeed, R. J. (1949). Some Statistical Aspects of Road Safety Research. *Journal of the Royal Statistical Society, Series A*, 1-34.
44. Tegnér, G & Loncar-Lucassi, V. (1997). Tidsseriemodeller över trafik- och olycksutvecklingen. Transek AB, Stockholm.
45. Tunaru, R.(2002). Hierarchical Bayesian models for multiple count data. *Austrian Journal of Statistcs*, 31, 221–229.
46. Van den Bossche, F. (2006). Road Safety, Risk and Exposure in Belgium: an Econometric Approach. PhD Thesis, Hasselt University.
47. Van den Bossche, F. and Wets, G. (2003) Macro Models in Traffic Safety and the DRAG Family: Literature Review. RA-2003-08.
48. Washington, S., Ivan, J.N. and Lord, D. (2007). Further notes on the application of zero-inflated models in highway safety. *Accident Analysis and Prevention*, 39, 53–57.
49. Washington, S., Ivan, J.N. and Lord, D. (2005). Poisson, Poisson-gamma and zero-inflated regression models of motor vehicle crashes: Balancing statistical fit and theory. *Accident Analysis and Prevention,* 37, 35–46.
50. WHO (2009) World report on road traffic injury prevention. World Health Organization.
51. Xie, Y., D. Lord, and Zhang, Y. (2007). Predicting Motor Vehicle Collisions using Bayesian Neural Networks: An Empirical Analysis. *Accident Analysis and Prevention*, 39, 922-933.
52. Zeger, S.L. (1988). A regression model for time series of counts. *Biometrika,* 75, 621–629.