Journal of Safety Research 49 (2014) 1

Contents lists available at ScienceDirect

Journal of Safety Research

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

Editorial

# Letter from the Editors

The *Journal of Safety Research* is pleased to publish in this special issue the proceedings of several papers presented at the 4th International Conference on Road Safety and Simulation convened at Roma Tre University in Rome, Italy, October 2013. This conference serves as an interdisciplinary forum for the exchange of ideas, methodologies, research, and applications aimed at improving road safety globally.

Conference proceedings provide the opportunity for research in its formative stages to be shared, allowing our readers to gain early insights in the type of work currently being conducted and for the researchers to receive valuable feedback to help inform ongoing activities. This conference in particular offers an array of research topics not often covered by this journal from researchers practicing in over 11 countries. As is common with publishing conference proceedings, the papers published in this issue did not go through the normal *JSR* review process. Each paper included in this issue did meet the Road Safety and Simulation conference review requirements. They reflect varying degrees of scientific rigor, methodological design, and groundbreaking application.

The proceedings published in this special issue of *JSR* draw from the following road safety research sectors represented at the conference: driving simulation, crash causality, naturalistic driving, and new research methods.

It is our hope that the publication of these important proceedings will stimulate vigorous dialogue, rigorous research, and continuing innovative initiatives and applications, leading, ultimately, to fewer traffic fatalities, injuries, and crashes.

Thomas W. Planek
Editor-in-Chief

Sergey Sinelnikov
Associate Editor

Jonathan Thomas
Associate Editor

Kenneth Kolosh
Associate Editor

Kathleen Porretta
Managing Editor

18 February 2014

http://dx.doi.org/10.1016/j.jsr.2014.02.001

Journal of Safety Research 49 (2014) 121–127

Contents lists available at ScienceDirect

Journal of Safety Research

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

# Analyzing road design risk factors for run-off-road crashes in the Netherlands with crash prediction models

J.W.H. (Jan Hendrik) van Petegem [a,*], Fred Wegman [b]

[a] SWOV — Institute for Road Safety Research, P.O. Box 1090, Postcode 2260 BB Leidschendam, Netherlands
[b] Delft University of Technology, Civil Engineering and Geosciences, The Netherlands

ARTICLE INFO

*Article history:*
Received 29 November 2013
Accepted 5 March 2014
Available online 24 April 2014

*Keywords:*
Crash prediction model
Run-off-road
Road design
Safety zone
Road safety

ABSTRACT

*Problem:* About 50% of all road traffic fatalities and 30% of all traffic injuries in the Netherlands take place on rural roads with a speed limit of 80 km/h. About 50% of these crashes are run-off-road (ROR) crashes. To reduce the number of crashes on this road type, attention should be put on improving the safety of the infrastructure of this road type. With the development of a crash prediction model for ROR crashes on rural roads with a speed limit of 80 km/h, this study aims at making a start in providing the necessary new tools for a proactive road safety policy to road administrators in the Netherlands. *Method:* The paper presents a basic framework of the model development, comprising a problem description, the data used, and the method for developing the model. The model is developed with the utilization of generalized linear modeling in SAS, using the Negative Binomial probability distribution. A stepwise approach is used by adding one variable at a time, which forms the basis for striving for a parsimonious model and the evaluation of the model. The likelihood ratio test and the Akaike information criterion are used to assess the model fit, and parameter estimations are compared with literature findings to check for consistency. *Results:* The results comprise two important outcomes. One is a crash prediction model (CPM) to estimate the relative safety of rural roads with a speed limit of 80 km/h in a network. The other is a small set of estimated effects of traffic volume and road characteristics on ROR crash frequencies. *Practical applications:* The results may lead to adjustments of the road design guidelines in the Netherlands and to further research on the quantification of risk factors with crash prediction models.



## 1. Introduction

### 1.1. Background

About 50% of all crashes resulting in fatalities and 30% of all crashes resulting in severe injuries[1] in the Netherlands take place on rural main roads with a speed limit of 80 km/h (further referred to as rural roads). Since these roads account for only a relatively small portion of the total road network, these roads are seen as the most unsafe roads in the Netherlands. Furthermore most of the crashes take place on a road section and about 50% of these crashes are run-off-road (ROR) crashes.

An important aspect in reducing the amount of crashes on rural roads is the improvement of the safety of both existing and new roads. This is the task and the responsibility of road administrators. There are some obstacles, however, that need to be overcome to enable an efficient and proactive policy by road administrators on the improvement of traffic safety on rural roads.

First of all, road administrators face a limited budget for the implementation of safety measures. Second, traditional black spots fall short in supporting a proactive traffic safety policy since they are unable to identify potential hazardous road stretches in case of low crash frequencies. Such is the case on these roads with an average of 0.5 crashes per kilometer. Third of all, there is a lack of knowledge with respect to the quantitative relation between traffic safety and the cross sectional road design elements and traffic safety measures. This becomes apparent in the Dutch road design guidelines of rural roads, where only few of the guidelines are quantitatively related to crashes.

### 1.2. Crash prediction models

To enable an efficient allocation of tight resources, new knowledge and instruments are needed for the detection of unsafe road sections and for the estimation of the effects of safety (design) measures. Crash prediction models (CPMs) can be used to estimate the crash frequency on road section related to the traffic flow, road length, and risk factors such as cross sectional design elements (further referred to as road characteristics). These models can help in investigating the quantitative relations between road characteristics and crashes; they can be used for identifying relative unsafe road section in a network and for estimating the effects of the implementation of safety measures on a road section.

* Tel.: +31 70 3173 302; fax: +31 70 3201 161.
*E-mail address:* Jan.Hendrik.van.Petegem@swov.nl (J.W.H.(J.H.) van Petegem).

[1] A severe injury is defined as an injury with a maximum abbreviated injury score (MAIS) score of MAIS = 2 or higher. A severe accident is defined as an accident where the MAIS of one of the actors is MAIS = 2 or higher

http://dx.doi.org/10.1016/j.jsr.2014.03.003
0022-4375/

The development of CPMs is, however, not a straightforward exercise. It is known that the road length and traffic flow (the exposition to danger) are major components in explaining the crash frequency on a road. A general accepted theory of how crash frequencies relate to traffic flow, however, does not seem to exist in the field of crash analysis. Such is also stated by Wood, Mountain, Connors, Maher, and Ropkins (2013).

Furthermore, problems can be expected with regard to the research database. High quality data in large quantities for road characteristics, traffic flow, and crashes are seldom readily available and compatibility issues arise when combining different datasets of road characteristics and traffic flow (Schermers & Duivenvoorden, 2010; Van Petegem, 2012). This has also been a problem in this research. Only a limited set of relevant road characteristics was available for the development of the CPM.

### 1.3. Aims of this research

The development of CPMs on rural roads in the Netherlands is still at its infancy. The research described in this paper therefore aims at taking the development of CPMs in the Netherlands one step further. The focus thereby is put on the Dutch practice where roadside crashes on rural roads are of major concern and the development of a CPM on ROR crashes is presented.

The paper aims at describing a basic framework for the development of CPMs in practice. This means that the research is aimed at the use of proven and readily available statistical test and regression analysis procedures in statistical software packages like SAS enterprise, SPSS, and STATA. Although it is acknowledged that some of the developments in regression analysis on crashes can lead to more accurate models, readily available techniques are preferred because of convenience in practice.

The structure of the rest of this paper is as follows. First, a problem description on ROR crashes is given. The following section describes the data available for the model development. Subsequently the model framework and the method for the development and evaluation of the CPM are presented. Finally the results on the development of a CPM on ROR crashes in the Netherlands is presented and discussed.

## 2. Problem description

The most urgent traffic safety problem on rural roads are ROR crashes. This becomes apparent when looking at the share of fatal and severe ROR crashes compared to all other crash types. The share of fatal and severe ROR crashes on rural roads amounts to about 50% of all crash types. Furthermore, although the number of deadly injury crashes has gradually declined about 50% and the number of severe injury crashes has gradually declined about 30%, respectively, in the period 1998–2008, the share of ROR crashes has remained the same (Van Petegem, 2012).

The majority of these ROR crashes in the Netherlands are from single vehicle crashes. In two in-depth studies where all possible information was collected on ROR crashes (traffic conditions, the immediate environment, the road users involved in the crash, and the sustained injuries) in two regions (Zeeland and The Hague Region) for a period of eight months (Davidse, 2011; Davidse, Doumen, Duijvenvoorde, & Louwerse, 2011), single vehicle crashes accounted for 94% of all ROR crashes. These crashes can be relatively easily identified in the crash data of the Netherlands. ROR crashes with multiple vehicles involved are, however, more difficult to identify as those crashes can be registered in different ways. It is uncertain how a crash is registered when a side impact crash also leads to a ROR crash. Because of the low share of ROR crashes with multiple vehicles and the uncertainty in the registration, only single vehicle ROR crashes are analyzed.

The majority of the population of rural roads under consideration are single carriageway roads, with two driving directions, one traffic lane per driving direction, without a median and a speed limit of 80 km/h. Because homogeneity is an important aspect in the analyses (e.g., a fixed safety zone will have a different impact at different speed limits), other road configurations with regards to those aspects are excluded. Since no variation is allowed on these elements within this research, they are also excluded from the crash characteristics.

The remainder of this paper presents an assessment on the relations between road design and ROR crash risk, based on a literature review.

### 2.1. Crash characteristics

The in-depth study of Davidse et al. shows that it is a confluence of events and general factors that forms the cause of ROR crashes. Major factors related to road design that can be identified from this study are curves, obstacles, driver (risk) awareness, and time and space for correction (Davidse, 2011; Davidse et al., 2011).

Curves impose an additional risk compared to straight road sections when small curve radiuses are in place and curves are not marked clearly. In the in-depth study of Davidse et al., 50% of all ROR crashes occurred in or directly after the curve. Two studies on crash risks due to curves in Sweden (Brüde, Larsson, & Thulin, 1980; Hedman, 1989) and the United States (Zegeer, Reinfurt, Neuman, Stewart, & Council, 1990; as in Dijkstra, 1998) provide similar findings about the relation between the curve radius and crash risks. It was found that the crash risk rapidly grows with curve radiuses smaller than 1000 m. The relative risk of a curve radius of 400 m was found to be about 40% higher than a curve radius of 1000 m. The Dutch road design guidelines on curves, however, are based on the physics of friction resulting in a minimum guideline of a curve radius of 300 m with a positive superelevation of 2.5% (CROW, 2002). Although the studies from Sweden and the United States don't take superelevation into account in the analysis, it might well be that small curve radiuses in the Netherlands are part of the causation of a large share of ROR crashes in curves.

Obstacles are also an important factor in the resulting injury from a ROR crash. In the Dutch road design guidelines (CROW, 2002), an obstacle is described as an object, vegetation, or other roadside element that causes severe damage and injury to the vehicle and occupants on impact. Some examples are poles, trees, ditches, slopes, and canals. The Dutch guidelines recommend a safety zone without any obstacles of 6 m and a minimum of 4.5 m, based on encroachment studies from the sixties from Hutchinson and Kennedy (1966) and a research to the crash risk of trees in the eighties (Schoon & Bos, 1983). In practice, however, these guidelines are often not met. A narrow safety zone was the cause of the injury in 40% of the ROR crashes in the in-depth study of Davidse et al. What is more, all fatal and severe injuries were caused by obstacles.

The human factor also plays an important role in the crash cause. In almost every case study in the in-depth study of Davidse et al., the (risk) awareness was negatively influenced, contributing to the crash cause, by one (or a combination) of the following factors: alcohol, distraction, fatigue, emotional state, and lack of experience (novice). Such factors can lead to accidental lane departures, slow awareness and anticipation to curves, slow responses or panic responses to lane departure and too high speeds for the circumstances. By means of visual road guidance and haptic feedbacks, the infrastructure can notify drivers for lane departures and (sharp) curves and as such help drivers with lane keeping and curve anticipation.

A final important factor is a fail-safe design, providing time and space for the correction of, for example, swerving and accidental departures. This aims at providing a driver enough time and space to correct the error by safely steering back on the road or safely coming to standstill in the verge by the cross sectional road design. This design principle relates to an array of factors, including a sufficient safety zone and road design elements that can stimulate (risk) awareness. Other important factors include the road width, lane width, and hard and soft shoulder width and friction. Those elements should provide vehicles the grip and space needed to safely redirect the vehicle or to safely come to a standstill.

### 2.2. Road design elements and quantifying the relation with crash risk and frequencies

A list of relevant road design elements in which a relation to ROR crashes is expected can be extracted from the crash characteristics:

- Road width, lane width and shoulder width
- Shoulder pavement or friction
- Safety zone
- Curve radius and curvature
- Road guidance (road markings, lighting, road side or curve markings)
- Haptic feedbacks (like rumble strips on shoulders, medians and preceding curves).

The Dutch road design guidelines cover most of these elements, prescribing minimum, recommended, and maximum designs standards. These design guidelines are, however, not obligatory, and road administrators may choose to deviate from the guidelines in practice.

The most common reasons to deviate from the guidelines when designing new roads or restructuring existing roads are tight budget and space. To assess the risks incurred by deviating from the guidelines and to assess the urgency and benefits of safety improvements, there is a need to quantify the relationship between crash risk and the different design standards prescribed in the design guidelines and deviations from these guidelines. This is missing, however, in the current guidelines. As a result, road administrators miss essential knowledge needed to balance the road safety effects (crash risks incurred by deviations of the guidelines) to other interests in the design and decision making process (like budget allocation).

The development of CPMs in the Netherlands aims at making the relation between road design and crash risk more explicit. This, in turn, can serve as a basis for better informed decision making in the road design.

## 3. Data

The database used for the model analysis in this research covers the rural roads of the two provinces Drenthe and Gelderland and consists of crash data, traffic flow data, and road characteristics data.

The roads that are considered are so called 'N-roads.' Every N road is divided by a hectometre (hm) layout system. This system is used to identify road sections by its road number, Nxxx, and an hm number. The Dutch crash database (BRON) uses this system to appoint an hm road section to a crash wherever possible.

The research database is also structured by the hm system, dividing the roads in 100 meter road sections. The database counts a total of 9,471 road sections, accounting for 947 km of rural roads. The traffic flow, road characteristics, and crash counts are registered by each 100 m road section. The three data types are described in the following paragraphs. The reports of Schermers and Duivenvoorden (2010) and Van Petegem (2012) give a more elaborate description and analysis of the data.

### 3.1. Road characteristics data

The road characteristics data are collected in the year 2007. Nothing is known about alterations of these road characteristics before and after the registration, Structural changes, however, are usually planned alongside with major maintenance works, which has a cycle of about 20 years. It is therefore assumed that by limiting the historical crash and traffic flow data for only 3 years, structural and small alterations to these roads will be minimal and insignificant.

Based on the road characteristics data and relevance (Section 2), a homogeneous selection of roads was selected. A total of 7,347 road sections were selected from the original database.

Furthermore, the road characteristics data provide sufficient data on the safety zone, road side barriers, road lighting, and curvature to be incorporated in the model analysis. The curvature in the database, however, does not relate to the geometric definitions, which are measured in gon. The curvature is subjectively measured by driving these roads, differentiating between a strong curvature, a mild curvature, and straights (on approximation). The other variables of interest were, however, not all properly covered by the database. Due to limitations to the database, the road design elements lane, shoulder and total road width, shoulder pavement, vertical alignment, road markings, and haptic feedbacks attributes could not be included in the model analyses.

### 3.2. Crash data

The crash data in the database are the sum of a 5 year period; 2004–2008. Although an even spread around the year of registration of road characteristics would normally be preferred if possible, an uneven spread is chosen because of a significant decline of the rate of registration of crashes after 2008. The total amount of ROR injury crashes on the selected roads is 324, which means that the average crash rate in 5 years is equal to 0.044.

The crash data in the database are extracted from BRON, the Dutch national crash database. Crash data are registered by the police when they are present at the crash site. BRON is also coupled with a spatial database of the Dutch national road network (NWB). Based on the BRON data and the coupling with NWB, a strict selection from only single vehicle crashes is made, of which the vast majority is a ROR crash.

A problem found in the crash registration in BRON was that not all crashes on the considered N-roads contained an hm number. This can happen, for instance, when the police do not register the hm number because of a missing hm or just simply forget to register the hm number.

By performing a spatial join in a GIS, an hm registration number was added to all crashes on the N roads without an initial hm registration. Although this is important for the sheer volume of crashes in the research database, it does introduce a spatial error. Crashes without an initial hm registration are coupled to the NWB based on the locational attributes of municipality, street name, and road number and are usually placed at the middle of a road section in the NWB with the same locational attributes. The closest hm number to the middle of this road section is then added to the crash. The real location (with a 100 m radius), however, is not known. Because the average section length with the same locational attributes is relatively small, on average only minor changes in traffic flow and road design attributes are expected on the complete road section. It is therefore assumed that, on average, the error of linking a crash to the wrong road characteristics will also be relatively small. The choice therefore has been made to include crashes in the database with an hm number based on the spatial join.

### 3.3. Traffic flow data

The traffic flow data in the database are the annual average daily traffic of the same period as the crash data; 2004–2008. The data are collected by the provincial road administrators. The density of measuring points is, however, insufficient to account for all lower level connections to the rural 80 km/h road network. This introduces uncertainty to the traffic counts. It is assumed that the traffic flow to and from these lower level roads is relatively small compared to the traffic flow on the rural 80 km/h roads.

## 4. Model framework

### 4.1. A general model formulation

A CPM aims at estimating the expected number of crashes on a road in a certain time frame (crash frequency) by its length, the traffic flow (the first part of independent variables), and several risk factors (factors that influence the risk on a crash, the second part of independent variables). The crash frequency is called the dependent (variable). The

other components are incorporated in the CPM as independent model variables, also called predictors. The two major predictors of all CPMs in crash analysis are the traffic flow and road length. Both predictors represent a part of the relation between the exposition to risk (amount of kilometers driven) and crashes: When the road length and or traffic flow increases, more crashes can be expected.

With regards to the road length this relation is often presumed to be linear, and as a result often found to be incorporated in the denominator of the expected crash frequency of the CPM. The relation with the traffic flow is assumed to be non-linear, giving a better fit in the model calibration process, although a generally accepted theory on how the traffic flow relates to accident frequencies does not exist (Wood et al., 2013). Nevertheless, there seems to be a consensus on a general model formulation in the field of crash analysis, which can be found in many publications on CPMs (AASHTO, 2010; Reurings et al., 2006).

$$E(Y) = \alpha Q^{\beta} L^{\gamma} e^{\sum \delta_j x_j} \quad (1)$$

where:

| | |
|---|---|
| $E(Y)$ | expected number of crashes, or expected crash frequency |
| $\alpha$ | constant, to be estimated by the calibration process (exponent of the intercept) |
| Q | annual average daily traffic flow (AADT) |
| L | road length or road section length |
| $X_j$ (1,2,…, n) | risk factors related to the crash risk, such as road characteristics which describe the cross sectional road design or road geometry. |
| $\beta$, $\gamma$, $\delta_j$ | model parameters of the different model variables, to be estimated by the calibration process |

### 4.2. Risk factors

The second part of the CPMs is the risk factors, which can consist of numerous components, among which road characteristics are most often found. Other risk factors can be, for instance, behavioral aspects or surrounding conditions like the weather.

The array of risk factors found in CPM studies varies a lot. Although many of the road design elements as stated in Section 2.2 can be found in parts in some models (as can be found for example in: Chin & Quddus, 2003; Lee & Mannering, 2002; Reurings et al., 2006), models with a complete set of stated relevant road design elements do not exist. Several reasons can be mentioned why a 'complete' model does not exist. For example:

- The lack of data — In many cases there is too little data (or no data at all) available on road characteristics to find statistically significant parameter estimations of the risk factors. In addition, models with a higher complexity (with many predictors) more data is than simpler models (with few predictors) to find statistically significant parameter estimations of the risk factor.
- A lack of variation in the data — When a risk factor has little variation, it will be difficult to find a significant effect of the variation, unless the effect is very high.
- Dependency between the explaining variables — If two variables strongly correlate, incorporating both variables (instead of one) may lead to problems in fitting the model.

This is why CPMs often cover a specific crash type, a specific road type, and/or road design elements. In this research the considered road design elements in the model development are limited by the availability of data on these elements and the size of the network.

### 4.3. Regression analysis

CPMs are developed based on regression analysis to obtain estimates of the model parameters and asses the fit of the model to the data. In the development of CPMs, the regression analysis is often based on Generalized Linear Models (GLM), assuming a Negative Binomial (NB) probability distribution for the crash counts on a road i, and the link function to be the log link.

Crash data are a type of count data, for which the Poisson distribution is the basic distribution (Mccullagh & Nelders, 1989). In the development of CPMs, however, models often suffer from overdispersion when the Poisson distribution is chosen (e.g., Lord & Mannering, 2010; Wood et al., 2013). Overdispersion can result in an overestimation of the significance of parameter estimations and thus a risk of falsely accepting parameter estimates as significant, which is clearly unwanted. The NB distribution can be used to correct for overdispersion (Hilbe, 2011) and is therefore preferred over the Poisson distribution in the development of CPMs.

The link function is a component of the GLM that specifies the relation between a linear predictor $\eta_i = \sum x_{ij}\delta_j$ and the expected value $E(Y_i) = \mu_i$. The link function is specified as $g(\mu_i) = \eta_i$. In the case of the NB distribution, the so called canonical link function is the log link function with $g(\mu_i) = \log \mu_i = \eta_i$ (BRON). This means that $\mu_i = e^{\eta_i} = e^{\sum x_{ij}\delta_j}$.

The functional specification of the general model (1) is a derivation of these specifications of the GLM of the NB distribution. This can easily be shown by rewriting the general model formulation:

$$E(Y) = \alpha Q^{\beta} L^{\gamma} e^{\sum \delta_j x_j} = e^{\beta \log Q + \gamma \log L + \sum \delta_j x_j + c}. \quad (2)$$

A maximum likelihood estimation (MLE) is used to obtain the estimates of the model parameters. The likelihood is a measure of the model fit. The higher the likelihood, the better the model fit. The MLE procedure maximizes the likelihood value of the model, by varying the model parameters (Bijleveld & Commandeur, 2012; Dobson, 2002). This research uses the GENMOD procedure from SAS.

There are other regression analysis techniques and probability distribution functions developed and evaluated to obtain better model fits, like random effect models (Chin & Quddus, 2003; Wood et al., 2013) and general Lindley models (Lord & Geedipally, 2011). They are more complicated to use since they are not yet implemented as standard procedures in statistical software like SAS Enterprise, SPSS, and Stata. This makes a regression analysis based on GLM, assuming a NB distribution is still a convenient choice in developing CPMs.

## 5. Method

The key to the development of CPMs are obtaining a parsimonious model, evaluating the fit of the model to the data, and the parameter estimates. With regards to these three aspects, a stepwise development of the CPM is chosen in this research. The first step of the model development is the construction of a null (or intercept-only) model, with no other variables except the intercept. In the next steps, a new model is constructed with only one additional variable per step. This method is explained in the following three paragraphs.

### 5.1. Parsimony

Parsimony is a principle in statistics (and science in general) that states that the simpler explanation to describe a phenomenon equally well should be favored above more complex possible explanations. A more complex model is, however, preferred if it presents a better explanation of the phenomenon. Translated to the development of CPMs, the model that gives the best fit to the crash data, using the least amount of predictors, should be the preferred model.

Several reasons can be mentioned in favor of striving for a parsimonious model. First of all, every additional predictor adds to the degrees of freedom (df) of the model. Adding the df to a model, an additional predictor will always improve the fit of the model. If in extreme the degrees of freedom would be equal to the data points, the fit of the model would be perfect. Chances are high that the predictors represent coincidental relations to the predictor or random errors in the data. This means that the model overfits the data. Meaning that when an additional predictor represents a coincidental relation with the dependent or random error in the data, the additional fit gained by the predictor is based on coincidence and the explanation of the model on the real relation between the dependent, and the predictors become worse.

Secondly, adding predictors increases the risk of correlations between the predictors. If two or more predictors strongly correlate, the individual contribution of each predictor with the dependent becomes unclear. Therefore, independent predictors are strongly preferred.

Also, adding predictors makes the model more complex to understand. It is, for instance, easier to understand a relation in two dimensions than in four or more dimensions.

Finally, every predictor represents real data. This means that for every additional predictor, more data needs to be gathered to be able to use the model. A model that explains the dependent well, by using only a minimum set of predictors, will therefore be cheaper than a more extensive model.

### 5.2. Model fit

The model fit describes how well the model is able to represent the real data. To evaluate the fit of the model, the likelihood ratio test (LRT) is used as well as the Akaike information criterion (AIC).

At every step in the construction of the model, the model fit is examined and compared to the other models, based on the likelihood ratio test (LRT) and Akaike information criterion (AIC). The LRT is a statistical (chi-squared) test that compares the fit of two models and states if the difference in the likelihood of the two models is statistically significant or not (Dobson, 2002). In every step, the LRT is used to compare the new model to the null model and to the model of the previous step. The comparison to the null model states whether or not the model gives a significant improvement in estimating the number of crashes on a road in comparison to the average crash number. The comparison to the previous model states whether or not the added predictor gives a statistical improvement in the model fit, compared to the model without this additional predictor. If not, the predictor should be eliminated from the model to obtain a parsimonious model and prevent the risk of overfitting.

Besides the LRT, the AIC is evaluated on every model. The AIC represents a relative score to the model fit. The AIC also compares the likelihoods of two models, but adds a penalty to the amount of predictors (or df) in the model as every predictor adds to the risk of overfitting the data. If the likelihood improves more than the penalty incurred by the added predictor, a better AIC score is gained. A lower AIC represents a better model. Therefore the model with the lowest AIC, given a set of models, represents the best model (Bijleveld & Commandeur, 2012; Burnham & Anderson, 2010). Caution is needed, however, since the AIC is only a relative score. Hence, if all models are rubbish, the best model is still rubbish as well. Therefore the AIC should never be used on its own.

### 5.3. Parameter estimates

The third step in evaluating the models is the evaluation of the parameter estimates. The first evaluation is based on the significance testing of the predictors. This tests the null hypothesis that the predictor should be zero. If p is smaller than 0.05 the null hypothesis is rejected and we can assume with 95% certainty that the predictor has a relation with the dependent variable (or crash frequency on a road).

Secondly, the parameter estimate should be evaluated on its sign and value, based on expectations coming from literature and or expert opinions. If the parameter estimate is counterintuitive to the expectations, it is likely that something is wrong with the parameter estimate that represents a coincidental relation or error in the data. Therefore, predictors should only be added to the model if the estimate can be evaluated on beforehand to determine expectation; this will prevent approving predictors that represent errors in the data or coincidental relations.

Thirdly, the precision describes the uncertainty about the value of the parameter estimates. The 95% confidence interval (CI) describes the range of possible values of the parameter estimate that would be found in 95 of the cases when 100 samples were to be drawn from a population. The smaller the range of the confidence interval, the higher the precision and the closer the real value to the parameter is expected.

### 5.4. Correlations of the predictors

The final step of the model evaluation is checking for correlations between the predictors. As stated earlier, correlating predictors makes a model more complex to understand and leads to a situation in which the relation between the dependent and the individual predictors is unclear. Moreover, if two predictors strongly correlate, one of the predictors will be sufficient to incorporate the effect of both predictors in the model.

To check for correlations between the predictors, a correlation matrix can be calculated. The matrix shows the correlation between two predictors for all the predictors in the model. The correlation between two predictors lies between −1 and 1. The sign shows if the correlation is positive or negative and the absolute value between 1 and 0 shows how strongly the predictors correlate, with 0 meaning no correlation and 1 a 100% correlation. This correlation also affects the confidence intervals. If predictor B is correlated to A and A has a wide confidence interval, the confidence interval of B will also be wide.

## 6. Model results and discussion

### 6.1. Results

This section presents the model analyses and results. The CPMs presented in this section are, however, an adaptation to the general model formulation as stated in Section 4. The relation between the traffic flow and the crash frequencies in this adapted CPM is modeled by two components: the traffic flow Q and a scaling factor of the traffic flow Qs that is calculated as Q/1000 (for numerical reasons of the parameter estimate which otherwise would be very small). The reason for doing so relates to a decline in the absolute amount of E(Y) at large values for Q, which is shown in Fig. 1.

Since a power function of Q cannot decline after an initial growth of E(Y), Qs was added to the model to act as a scale factor to bend the curve of Q to the actual lower amount of E(Y) by higher values of Q to improve the fit of the model. By performing an LRT, comparing the model with and without Qs, a significant improvement in the model fit was indeed

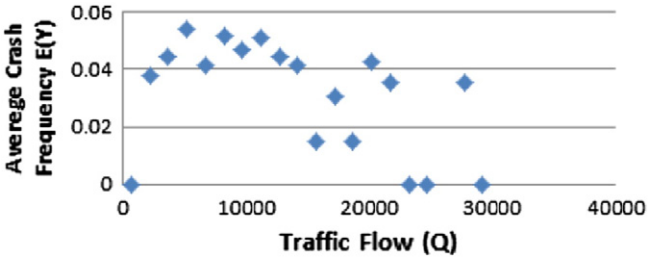


**Fig. 1.** Plot of the average ROR crash frequencies against the traffic flow.

found (with a p value of 0.02 for the LRT), which was reflected by the AIC as well. The parameter estimate of Qs was also found to be significant and lets Qs indeed act as a scaling factor for Q. The adapted CPM is given by:

$$E(Y) = Q^{\beta} e^{\theta \cdot Qs + \sum \delta_j x_j + I}. \quad (3)$$

The best model (including the risk factors), which was selected based on the AIC value of the models is:

$$E(Y) = Q^{1,21} \cdot e^{-0,14 \cdot Qs + 0.40 \cdot obstacle2m - 0.71 \cdot barrier + 1.08 \cdot scurvature + 0,18 \cdot mcurvature - 13,02}. \quad (4)$$

The LRT comparing this model to the null model was highly significant with p being smaller then 0.001. All correlations between the predictors of this model are 0.1 or smaller (except for Q and Qs). Thus, the predictors are considered independent. The parameter estimates, confidence intervals, significance of all predictors, and the exponents of the dummy predictors are presented in Table 1.

Important findings of this model are found by the risk factors. The obstacle predictor is a dummy variable that states whether or not an obstacle within 0–2 m on one or both road sides is present. The exponent of the predictor shows us that on road sections with an obstacle present within 2 m of the road, 50% more ROR crashes are expected (all other variables being equal) than on other roads. This is similar to findings presented by Lamm, Psarianos, and Mailaender (1999). The curvature predictor is a categorical predictor. It states that for roads with a strong curvature, almost three times more ROR crashes are expected compared to straight roads. This relates to the higher risk of small curves on crashes as mentioned in Section 2 and the analysis of Lamm on curvature (Lamm et al., 1999). Both of these estimates are highly significant.

The barrier predictor indicates that the presence of a roadside barrier reduces the expected amount of ROR crashes by 50%, compared to road sections with an obstacle in the safety zone. This relation is, however, only indicatory because the certainty is only 91%. The parameter estimate of the medium curvature is not tested for significance but kept in the equation as part of the categorical variable curvature.

Other predictors that were tested are a categorical variable on the safety zone, distinguishing the presence of distances of obstacles to the road for 0–2 m, 2–5 m, 5–7 m and 7 m or more, and the presence of lighting poles. However, the predictors on the obstacle distance of 2–5 m and 5–7 did not test significant with p values of 0.7 and 0.3. The model given by Eq. (4), excluding these two predictors, showed to have a lower (thus better) AIC value. The predictor of the presence of lighting poles also showed to be insignificant with a p value of 0.16. These predictors were therefore eliminated from the model.

The insignificant testing of the safety zone predictors of 2–5, and 5–7 m was believed to be a problem of an insufficient number of road sections within the database. Both categories of the safety zone of 0–2 m and 7 m or more account for 40% of the road sections each. The categories of 2–5 m and 5–7 m only accounted for 15% and 5% of the database. As it is assumed that obstacles in these categories actually do affect the crash risk, it is also assumed that the lack of significant findings on these variables is due to a lack of data.

The lack of significance of the barrier predictor (only indicatory), was also believed to be a problem of a lack of data, since a barrier was present on only 5% of the road sections in the database. Another issue in this case is that barriers might have a positive effect on severe crashes (including deadly crashes), but a negative effect on light injury crashes. The database did not contain enough severe crashes to build a statistically significant model based on just the severe injury crashes.

**Table 1**
Parameter estimates of the CPM of Eq. (4).[a]

| Predictor | Exponent | Parameter estimate | Left 95% CI | Right 95% CI | p-Value |
|---|---|---|---|---|---|
| Intercept | 2.22E−06 | −13.02 | −20.59 | −6.08 | 0.001 |
| Q | | 1.21 | 0.33 | 2.17 | 0.009 |
| Qs | 0.87 | −0.14 | −0.27 | −0.026 | 0.022 |
| Obstacle2m | 1.49 | 0.40 | 0.15 | 0.65 | 0.001 |
| Barrier | 0.49 | −0.72 | −1.68 | 0.06 | 0.100 |
| Scurvature | 2.94 | 1.08 | 0.24 | 1.86 | 0.008 |
| Mcurvature | 1.20 | 0.18 | −0.06 | 0.44 | 0.138 |

[a] p-Values are estimated based on likelihood statistics and not on Wald statistics. p-Values are therefore equal to a type 3 analysis in which a LRT is performed on the presence and absence of each predictor in the full model.

### 6.2. Discussion

It was found that Qs as a scale parameter of Q significantly improved the model fit, by lowering the expected crash frequency for higher traffic flows. This means that at higher traffic flows, not only is a lower individual crash risk expected (such is modeled by the power function of Q), but that a lower number of crashes is also expected. One explanation could be that roads where higher traffic flows are expected at the design stage are designed safer. Such was stated by road administrators in the Netherlands in conversations. This would mean that CPMs would not be able to estimate changes in the expected crash risk by changes in the traffic flow only. In case the road design is not altered, the model might not represent the changes in the expected crash frequency by changes in the traffic flow. After all, higher or lower traffic flows in the CPM might represent different road designs, which are not differentiated in the dataset since it only provides limited information on the road design. Furthermore, as the normal power function and the alternated model formulation both indicate a decline on the risk in a crash for higher traffic flows, this issue might have affected other models as well. Further research on this issue is needed to check if this relation indeed is a true explanation for Qs and other CPMs.

## 7. Summary and conclusions

The paper has presented a basic framework of the development of a crash prediction model on run-off-road crashes in the Netherlands. Relevant road characteristics on analyzing run-off-road crashes are identified based on both Dutch and international literature. The data available for this research are briefly described with some of it limitations. This study shows that the available data limits the possibilities on the development of crash prediction models. Because of the limitations of the available data, only the safety zone, roadside barriers, lighting, and curvature could be evaluated on their relation with run-off-road crashes by the developed crash prediction model.

A model framework and method for the model development and evaluation of crash prediction models is presented as well. These consisted of a formulation of a general model, regression analysis techniques, and basic instruments for developing and evaluating the model. This framework is focused on the development of crash prediction models in practice and as such presents a framework that can be used with software like SAS enterprise and SPSS, without the need of extensive knowledge on programming, statistics, and numerical problem solving and optimization processes. This paper suggest the use of generalized linear modeling procedures available in SAS enterprise (GENMOD) and SPSS (GENLIN), specified by the Negative Binomial distribution function and log link function for the development of the crash prediction model. A stepwise model building process is suggested for the development of a parsimonious model and for the assessment of the individual contribution of predictors to the model fit with the use of the likelihood ratio test and Akaike information criterion. Finally, several criteria are suggested to check for the quality of the parameter estimates of the predictors.

The results of the model development comprise a crash prediction model with an alteration of the general model formulation to obtain a better model fit and estimates of the relation between the number of

crashes on a road for the safety zone and curvature. It was found that on roads with a small safety zone of maximum 2 m, 50% more run-off-road crashes were expected than on other roads (all other variables being equal). A strong curvature was found to add 3 times more run-off-road crashes than on straight roads. Additionally the parameter estimate of road side barriers indicated a reduction of 50% of run-off-road crashes compared to roads with a small safety zone.

## Acknowledgments

This paper is a follow up of my master thesis which I conducted in collaboration between the Technical University of Delft, engineering company Wittenveen+Bos and the SWOV — Institute for road safety research. I want to express my gratitude to my thesis committee members for their time, effort and advice in my thesis research which has formed the basis of this paper: Dr. Ir. A. Dijkstra, Ir. E. Verbree, Ir. J. Verspuij, Prof. Ir. F.C.M. Wegman and Ir. P.B.L. Wiggenraad

## References

AASHTO (2010). *Highway Safety Manual* An introduction to the Highway Safety Manual + Volumes 1, 2 and 3 (1st ed.). Washington, D.C. American Association of State Highway and Transportation Officials AASHTO (1500 pp., ref.; HSM-1 — ISBN 1-56051-477-0 (Three-volume-set)).

Bijleveld, C. C. J. H., & Commandeur, J. J. F. (2012). *Multivariate analyse: een inleiding voor criminologen en andere sociale wetenschappers: Uitgeverij Lemma b.v.*

Brüde, U., Larsson, J., & Thulin, H. (1980). *Trafikolyckors samband med linjeföring (VTI Meddelande No. 235).* Linköping: Statens Väg- och trafik institut VTI.

Burnham, K. P., & Anderson, D. R. (2010). *Model selection and multi-model inference: A practical information-theoretic approach.* New York: Springer.

Chin, H. C., & Quddus, M. A. (2003). Applying the random effect negative binomial model to examine traffic accident occurrence at signalized intersections. *Accident Analysis & Prevention, 35*(2), 253–259.

CROW (2002). *Handboek wegontwerp wegen buiten de bebouwde kom: gebiedsontsluitingswegen (164c).* Ede: CROW kenniscentrum voor verkeer, vervoer en infrastructuur.

Davidse, R. J. (2011). *Bermongevallen: karakteristieken ongevalsscenarios en mogelijke interventies (R-2011-24).* Leidschendam: SWOV — Stichting Wetenschappelijk Onderzoek Verkeersveilighied.

Davidse, R. J., Doumen, M. J. A., van Duijvenvoorde, K., & Louwerse, W. J. R. (2011). *Bermongevallen in Zeeland: karakteristieken en oplossingsrichtingen (R-2011–20).* Leidschendam: SWOV.

Dijkstra, A. (1998). *Oriëntatie op kwantitatieve relaties tussen elementen van het wegontwerp en indicatoren voor verkeersonveiligheid: literatuurstudie buitenlands onderzoek. (R-98-49).* Leidschendam: Stichting Wetenschappelijk Onderzoek Verkeersveiligheid SWOV.

Dobson, A. J. (2002). *An introduction to generalized linear models* (2nd ed.). Chapman & Hall/CRC Press Company.

Hedman, K. -O. (1989). Road design and safety. *Proceedings of strategic highway research program and traffic safety on two continents in Gothenburg, Sweden, 27–29 September, 1989* (pp. 225–238). Linkoping: VTI.

Hilbe, J. M. (2011). *Negative binomial regression.* Cambridge University Press.

Hutchinson, J. W., & Kennedy, T. W. (1966). *Medians of divided highways: Frequency and nature of vehicle encroachments (Bulletin 487).* Illinois: University of Illinois Engineering Experiment Station.

Lamm, R., Psarianos, B., & Mailaender, T. (1999). *Highway design and traffic safety engineering handbook.* New York: McGraw-Hill.

Lee, J., & Mannering, F. (2002). Impact of roadside features on the frequency and severity of run-off-roadway accidents: An empirical analysis. *Accident Analysis & Prevention, 34*(2), 149–161.

Lord, D., & 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 & Prevention, 43*(5), 1738–1742.

Lord, D., & Mannering, F. (2010). The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives. *Transportation Research Part A: Policy and Practice, 44*(5), 291–305.

Mccullagh, P., & Nelders, J. A. (1989). *Generalized linear models* (2nd ed.). Suffolk: Chapman & Hall/St Edmundsbury Press.

Reurings, M., Janssen, T., Eenink, R., Elvik, R., Cardoso, J., & Stefan, C. (2006). *Accident prediction models and road safety impact assessment: A state-of-the-art.* Brussels: European Commission, Directorate-General for Transport and Energy (TREN).

Schermers, G., & Duivenvoorden, C. W. A. E. (2010). *Een SWOV-database wegkenmerken.* Leidschendam: SWOV.

Schoon, C. C., & Bos, J. M. J. (1983). *Boomongevallen: een verkennend onderzoek naar de frequentie en ernst van botsingen tegen obstakels in relatie tot de breedte van de obstakelvrije zone (R-83-23).* Leidschendam: SWOV.

Van Petegem, J. W. H. (2012). *Een modelonderzoek naar bermongevallen: Delft.* Technische Universiteit Delft TUD, Faculteit Civiele Techniek (Master Thesis, 177 pp., 50 ref.).

Wood, A. G., Mountain, L. J., Connors, R. D., Maher, M. J., & Ropkins, K. (2013). Updating outdated predictive accident models. *Accident Analysis & Prevention, 55*, 54–66.

Zegeer, C., Reinfurt, D., Neuman, T., Stewart, R., & Council, F. (1990). *Safety improvements on horizontal curves for two-lane rural roads: Informational guide.* Chapel Hill: University of North Carolina UNC, Highway Safety Research Center HSRC.

**Jan Hendrik van Petegem** is a researcher at the SWOV. His research is focused on the relation between road design and road safety. Before his work at the SWOV he graduated for his master's degree at the section Transport & Planning at the faculty of Civil Engineering of the Technical University of Delft.

**Fred Wegman** is a full professor Traffic Safety at the Delft University of Technology. He is the former managing director of the SWOV Institute for Road Safety Research in the Netherlands. His main fields of research at the University are safe road design, safety in relation to ITS in road traffic, road safety data and 'evidence based and data driven' road safety strategies.