ELSEVIER

Accident Analysis and Prevention 39 (2007) 657–670

ACCIDENT ANALYSIS & PREVENTION

www.elsevier.com/locate/aap

A crash-prediction model for multilane roads

Ciro Caliendo [a,*], Maurizio Guida [b], Alessandra Parisi [a]

[a] *Department of Civil Engineering, University of Salerno, 84084 Fisciano (SA), Italy*

[b] *Department of Information and Electrical Engineering, University of Salerno, 84084 Fisciano (SA), Italy*

Received 24 March 2006; received in revised form 5 October 2006; accepted 21 October 2006

Abstract

Considerable research has been carried out in recent years to establish relationships between crashes and traffic flow, geometric infrastructure characteristics and environmental factors for two-lane rural roads. Crash-prediction models focused on multilane rural roads, however, have rarely been investigated. In addition, most research has paid but little attention to the safety effects of variables such as stopping sight distance and pavement surface characteristics. Moreover, the statistical approaches have generally included Poisson and Negative Binomial regression models, whilst Negative Multinomial regression model has been used to a lesser extent. Finally, as far as the authors are aware, prediction models involving all the above-mentioned factors have still not been developed in Italy for multilane roads, such as motorways. Thus, in this paper crash-prediction models for a four-lane median-divided Italian motorway were set up on the basis of accident data observed during a 5-year monitoring period extending between 1999 and 2003. The Poisson, Negative Binomial and Negative Multinomial regression models, applied separately to tangents and curves, were used to model the frequency of accident occurrence. Model parameters were estimated by the Maximum Likelihood Method, and the Generalized Likelihood Ratio Test was applied to detect the significant variables to be included in the model equation. Goodness-of-fit was measured by means of both the explained fraction of total variation and the explained fraction of systematic variation. The Cumulative Residuals Method was also used to test the adequacy of a regression model throughout the range of each variable. The candidate set of explanatory variables was: length (L), curvature ($1/R$), annual average daily traffic (AADT), sight distance (SD), side friction coefficient (SFC), longitudinal slope (LS) and the presence of a junction (J). Separate prediction models for total crashes and for fatal and injury crashes only were considered. For curves it is shown that significant variables are L, $1/R$ and AADT, whereas for tangents they are L, AADT and junctions. The effect of rain precipitation was analysed on the basis of hourly rainfall data and assumptions about drying time. It is shown that a wet pavement significantly increases the number of crashes.

The models developed in this paper for Italian motorways appear to be useful for many applications such as the detection of critical factors, the estimation of accident reduction due to infrastructure and pavement improvement, and the predictions of accidents counts when comparing different design options. Thus this research may represent a point of reference for engineers in adjusting or designing multilane roads.

Keywords: Crash-prediction model; Multilane road; Negative Multinomial distribution; Traffic flow; Road geometry; Pavement friction; Weather

1. Introduction

Over the last few years numerous road-accident-prediction models have been developed to investigate the effects that various variables may have on the value of a pre-selected crash indicator. The most common crash indicators that have hitherto been used are the number of crashes per year (crash frequency) and the number of crashes per million vehicle-kilometres (crash rate). The fact that accidents might not be a linear function of traffic flow and section length generally induces one to use crash frequency as the more appropriate dependent variable for predictive models.

Many independent variables affect crash frequency, and these are generally related to traffic flow, section length, infrastructure geometric characteristics, pavement surface conditions, lighting, weather and driver behaviour. The value of certain variables is sometimes difficult to estimate, and the influence of such variables on road accidents may not be equally significant. Hence, from a wider set of independent variables researchers generally extract a reduced number of variables for inclusion in the proposed model. Traffic flow and road geometry, for example, have for years been recognized as the major causes of crashes, while only recently has human behaviour come to be increasingly acknowledged also as one of the predominant factors. Thus a

* Corresponding author. Tel.: +39 089 964140; fax: +39 089 964045.
E-mail address: ccaliendo@unisa.it (C. Caliendo).

0001-4575/$ – see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.aap.2006.10.012

vast literature exists regarding accident-prediction models based on independent traffic and road variables, whereas the human component variable is still seldom ever considered.

In keeping with the literature, prediction models have been developed by using analysis techniques that may basically be grouped into four main approaches, namely Multivariate Analysis (for references see Section 2), Empirical Bayes Method (Hauer, 2001; Miaou and Song, 2005; Ozbay and Noyan, 2006; Persaud et al., 1999), Fuzzy Logic (Adeli and Karim, 2000; Hsiao et al., 1994; Sayed et al., 1995) and Neural Network (Abdelwahab and Abdel-Aty, 2001; Chiou, 2006; Delen et al., 2006; Mussone et al., 1999). Furthermore, Principal Component Analysis (PCA) is also now being used as a preliminary tool for removing redundant variables (Caliendo and Parisi, 2005; Golob and Recker, 2003). However, among these techniques Multivariate Analysis has been successfully applied for many years so that nowadays it is the most widespread one adopted. The other statistical approaches mentioned above, albeit appearing to have a significant accident-forecasting capability, have only recently begun to be used, so that only few studies are available. Therefore a lot of detailed research has still to be done regarding their employment in the field of road safety. For this reason Multivariate Analysis has been pre-selected in this paper for its long and consolidated use in accident analysis.

With regard to this latter approach, the earlier models were based on Multiple Linear Regression, with the assumptions of normally distributed errors and homoscedacity. However, it was soon recognized that the nature of accident occurrence is such that it is better to model the process by using a Poisson distribution. The desired forms of relationships using the Poisson distribution have been developed by means of the technique of Generalized Linear Models (GLMs). However, albeit representing a significant advance in modelling capability, the Poisson distribution contains certain weaknesses, one of which regards the fact that the expected number of accidents per time unit is equal to the variance. In many accident analyses in contrast, the so-called phenomenon of "overdispersion" has been observed. Thus in order to take this parameter into account, the Negative Binomial or the Negative Multinomial distributions are often used when searching for accident-prediction models.

Many researchers have refined prediction models, and have also revealed their operating limits, as well as highlighting questions that still remain open. However, these models have been developed in countries where accident surveys, human behaviour, infrastructure and traffic characteristics differ from those in Italy. In addition, previous research has paid little attention to variables such as pavement surface friction and sight distance. Moreover, they generally refer to two-lane rural roads, whereas multilane roads have hitherto been investigated to a lesser degree.

As a result, there are at least four main reasons for justifying this paper. The first is motivated by the need to quantify the safety effects on the expected number of crashes for multilane rural roads of all the following variables: traffic flow, road geometry, sight distance, pavement surface friction and rain precipitation. In fact, there is an a priori reason to believe that crashes on multilane roads are associated with these variables in different ways as compared to accidents occurring on two-lane rural roads. Secondly, given the range of the above-mentioned statistical regression models (Poisson, Negative Binomial or Negative Multinomial distribution), an appropriate choice for modelling crash should be suggested. A third reason is for suggesting countermeasures appropriate for improving road safety. For example, it may be better understood whether geometry adjustment related to curves and junctions reduces crashes, and whether pavement resurfacing with the use of porous asphalt as surface course has a positive effect on road safety when it rains. Finally, given the gap of crash predictive models for Italian motorways, relationships that were not known before between total and severe crashes and the set of explanatory variables should be estimated. Consequently, this research might represent over the next few years a point of reference for engineers in adjusting or designing multilane roads.

Such then, is the context wherein the present work is set. The paper sets out to study crash data of accidents occurring in Italy on multilane roads. The objective is to identify a specific prediction model to estimate crash frequency as a function of traffic flow, infrastructure characteristics, pavement surface conditions (including whether wet or dry) and sight distance. For this purpose the results of a 5-year monitoring period on a four-lane median-divided motorway were analysed. Each carriageway was divided into segments with constant horizontal curvature and longitudinal slope. The number of crashes per year occurring on these geometric elements was assumed to represent the dependent variable to be related both to traffic flow and road factors in the model herein proposed. Both total crashes and injury crashes including fatalities were analysed in order to consider accident severity. The fact that the estimates here proposed combine fatal and injury accident data is a logical step when data are scarce.

Since preliminary analyses showed that a unique regression model (with zero curvature for tangents) would imply considerably overestimating (underestimating) the number of crashes occurring on curves (tangents), the methodology was applied separately to tangents and curves. Poisson, Negative Binomial and Negative Multinomial distributions were used to model the random variation of the number of crashes. The Maximum Likelihood Method was used for estimating prediction model parameters, and then the Generalized Likelihood Ratio Test (GLRT) was applied to evaluate the significant variables to be introduced into the final model. Subsequently the goodness-of-fit of the regression model was tested by means of both the explained fraction of total variation and the explained fraction of systematic variation. The Cumulative Residual method was also used to test the adequacy of a model throughout the range of each regression variable.

2. Literature review

As far as the authors are aware the first accident-prediction models for multilane roads were devised by Persaud and Dzbik (1993). Relationships between crash data and traffic flow, expressed both as average daily traffic (ADT) and hourly volume

(VH), were proposed. The analysis was based on generalized linear models. Results showed that crash rate increases with increasing traffic flow expressed both as ADT and VH. Accident risk on four-lane freeways was found to be lower than on freeways with more than four lanes, reflecting the fact that on freeways with more than four lanes, under the same traffic volume, free-flow conditions prevail, so that users have greater freedom for manoeuvre which major crash risk is associated with. Thus, traffic flow expressed as VH appears to be more appropriate than ADT for explaining accident phenomena since it takes into account congested or free-flow traffic conditions at the time of crashes. Unfortunately, accurate measures of VH are difficult to obtain, so that ADT is often used in accident-prediction models.

Knuiman et al. (1993) examined the effect of median width of four-lane roads on crash rate using a Negative Binomial distribution. The findings indicated that crash rate decreases with increasing median width. Furthermore, wider medians considerably reduce "crossover accidents" involving head-on crashes between opposing vehicles. As a result, a much greater positive effect on severe crashes than on property-damage-only crashes is expected.

Fridstrøm et al. (1995) related road accidents to four variables, namely traffic flow, speed limits, weather and lighting conditions. They considered Negative Binomial regression. The major point of interest of their work is the goodness-of-fit measure. Since goodness-of-fit measures based on the fraction of the explained variation as compared to the total variation (R^2 or R_D^2) are rather low or controversial, a new approach involving the explained variation R_D^2 as compared to the systematic component of variation P_D^2 rather than to total variation was developed. On the basis of the above-mentioned four independent variables, they were able to explain between 85 and 95% of the systematic component of variation.

Hadi et al. (1995) proposed several accident-prediction models with regard both to multilane roads and two-lane roads of rural or urban designation. The dependent variables were total crash rate or injury crash rate. The values of these accident indicators were estimated as a function of AADT and road environmental factors. Poisson and Negative Binomial regression models were considered. By examining the effect of traffic flow on the crash rate the conclusions reached were that crash rate increases with increasing AADT on roads having higher levels of traffic, while it decreases with AADT on roads with lower traffic volumes. This finding reflects the fact that in the presence of low traffic volumes, free-flow conditions exist so that by increasing AADT the users have more restricted freedom for manoeuvre with which a lesser crash risk is associated.

Persaud et al. (2000) presented one of the earliest studies for carrying out separate analyses for curves and tangents, albeit limited to two-lane roads. The dependent variable was crash frequency, while the independent variables were traffic flow and road geometry. Regression models were calibrated using a generalized linear modelling. A dummy variable for "flat" or "undulating" terrain was also used. For curves, crash frequency was found to increase with: AADT, section length (L) and curvature ($1/R$). For tangents, the number of accidents per year increases with AADT and L. A higher accident number on undulating terrain than on flat one was also shown.

Abdel-Aty and Essam Radwan (2000) used Negative Binomial distribution to predict crash frequency as a function of: AADT, degree of horizontal curvature, section length, lane, shoulder and median widths, and urban/rural designation. The attractiveness of this study is that the models were developed also to account for driver characteristics, including sex (male or female) and age (young, middle-aged and old). Results showed that crash frequency increases with AADT, degree of horizontal curvature and section length. Accident frequency decreases, in contrast, with lane, shoulder and median width.

For French interurban motorways, Martin (2002) described the relationship between crash rate and traffic volume per hour (VH) and the influence of traffic on crash severity. A Negative Binomial distribution was used. To model the probability of observing at least one injury crash in a crash, a logistic regression with a random component having a Binomial distribution was also used. The major point of interest emerging from this study is that the relationship between crash rate and VH was shown to be non-linear. Higher crash rate values, both for damage-only and for injury crashes, were found when VH was fewer than 400 vehicles/h. Crash rate also decreased rapidly with increasing VH, passing through a minimum (approximately 1000 and 1500 vehicles/h for two-lane and three-lane carriageways, respectively). After that the crash rate increased gradually with an increase in traffic.

Golob and Recker (2003) used linear and non-linear multivariate statistical analyses to determine how the type of accidents related to traffic flow, weather and lighting conditions. The study approach was based on the Principal Component Analysis (PCA) in order to identify the most significant variables from a set of original traffic flow variables, and a canonical correlation analysis (CCA) was used to relate the identified principal components both to weather and lighting conditions.

In a subsequent paper, Golob et al. (2004) evaluated the safety effects of changes in freeway traffic flow. The research was based on some of the same statistical methods used in the previous study, with the addition of further steps which were required for monitoring and forecasting purposes. Three crash characteristics were used in the analysis, namely crash type (rear end, sideswipe or hit object, number of vehicles involved); crash location (e.g. left lane, interior lanes, right lane, shoulder); crash severity (injuries and fatalities per vehicle).

Hauer (2004a) developed statistical road safety modelling by using the Negative Binomial distribution. The dependent variable was the number of accident per year, while the independent ones were geometric characteristics and traffic flow. First of all the author suggested guidelines for assigning the functional form to each variable in the model, and observed that the model equation should have both a multiplicative and an additive component. The multiplicative component is to account for the influence of variables that have a continuous role along a road (such as lane width or shoulder type), while the additive component is to account for the presence of hazardous points (such as driveways or narrow bridges). The most innovative aspect of this study was the introduction of an alternative tool for mea-

suring the goodness-of-fit of the predictive models, the so-called Cumulative Residuals (CURE) Method. This method consists of plotting the cumulative residuals as a function of the independent variable of interest, a good CURE plot being one oscillating around zero.

In a further study Hauer (2004b) applied the above-mentioned statistical model to estimate crash frequency on undivided four-lane urban roads. The proposed models evaluated the number of accidents per year and carriageway as a function of the following independent variables: AADT, percentage of trucks, degree and length of horizontal curves, grade of tangents and length of vertical curves, lane width, shoulder width and type, roadside hazard rating, speed limit, access points (e.g. signalized intersections, stop-controlled intersections, commercial driveways and other driveways), the presence and nature both of parking and two-way-left-turn-lanes. The findings showed that significant variables were: AADT, the number of commercial driveways and speed limit.

As it regards the effect of precipitation on crashes there is a general perception that this variable represents a road traffic hazard. But the findings are sometimes difficult to compare since they are based on: a variety of methods and variables, different accident types (damage-only accidents or fatal and injury accidents) and different time periods and data sources. However, most studies show that precipitation in the form of rain and snow causes more accidents as compared to dry conditions (Edwards, 1998; Eisenberg, 2004; Fridstrøm et al., 1995; Golob and Recker, 2003; Keay and Sommonds, 2005, 2006; Shankar et al., 1995). In particular it was found that the height of precipitation and, more relevantly, the time from the last precipitation are factors which affect crashes significantly.

This paper focuses on variables related to traffic flow, infrastructure geometry, pavement surface and rainfall. The next section describes the data set used and the process of preparing it for analysis.

3. Data description

A 5-year monitoring period extending from 1999 to 2003 was carried out on a four-lane median-divided motorway. This infrastructure was 46.6 km long, and the horizontal alignment contained tangents and circular curves without any transition curves. Vertical alignment consisted of gradients and circular curves.

During the period of observation, crash data, traffic flow, pavement surface conditions and rainfall data were collated. Accident data were extracted from the official reports of the Motorway Management Agency (MMA). For each accident a variety of details was recorded, including date and location of accident, horizontal alignment (tangent or curve), vertical alignment (upgrade or downgrade), weather and pavement surface conditions (dry or wet), type and severity of accidents, number of vehicles and persons involved, and a short description of the accident dynamics. Some 1916 accidents were considered in this study, 21 of which were fatal and 594 were injury accidents. Since fatalities appear to be too few to be analysed alone, fatal and injury crashes were considered collectively and are referred to as "severe" crashes hereinafter. 31.1% of all crashes and 33.4% of severe crashes occurred on curves, which represent 29.7% of the total length of the motorway. Table 1 gives accident count data observed during the 5-year monitoring period.

The database does not include accidents taking place on the ramps of junctions, on service areas, at tollbooths or on shoulders, since such accidents are not due to traffic flow and infrastructure characteristics. Pedestrians and bicycles are forbidden to use this infrastructure. Thus no pedestrian and bicycle were involved in accidents.

Besides, the monitored motorway is located in the South of Italy, connecting the cities of Naples and Salerno. In this flat area, meteorological data showed that very rarely did the temperature drop below freezing and that neither snow nor extremely fog were observed during the 5-year monitoring period (1999–2003). As a result, weather conditions such as ice, snow or fog cannot be considered as causes of crashes in the present case study. In contrast rain precipitations are frequent in this area.

Official reports on accidents do not contain information about drunk driving and this fact could represent a limitation. In fact very little data are presently available in Italy concerning the effect of the alcohol level in the blood on road safety. However, according to the Italian National Institute of Statistics (ISTAT) roads accidents due to alcohol abuse are very few (1.2%) so that it is reasonable to believe that the lack of this variable does not significantly limit the findings of this study.

As a result of the above considerations, this paper correlates total and severe crashes to traffic flow, horizontal and vertical motorway alignment, pavement surface characteristics and rainfall.

For the purpose of the subsequent statistical analysis, homogeneous road sections were first identified, i.e. segments for each carriageway having constant horizontal curvature and longitudinal slope. For these segments the following major variables did not change: width and number of lanes, type and width of shoulders, median width and type.

3.1. Horizontal and vertical alignment

Horizontal and vertical alignments of the monitored motorway were derived from a file containing a recent 3D aerial survey. The Autodesk Inc. AutoCAD® 2000 software was applied in order to measure the geometric characteristics (e.g. length of tangents and curves, horizontal and vertical curvature, longitudinal slope). Tangents with length ranging from 0.1 to 1.7 km and horizontal curves with radii from 0.2 to 8.0 km were computed. Furthermore, gradients with longitudinal slope ranging between −4.5 and +4.5% were estimated and vertical curves of circular type were defined.

3.2. Sight distance

In order to establish the role played by restricted visibility on accident occurrence, sight distances (SD) regarding the horizontal and vertical alignment were also determined.

On horizontal curves, a physical feature outside the travelled way such as the longitudinal safety barrier was con-

Table 1
Accident count data observed during the 5-year monitoring period

Year	Number of all (severe) accidents							Total number of vehicles travelling
	North direction			South direction			Year's total	
	Tangents	Curves	Year's total	Tangents	Curves	Year's total		
1999	140 (47)	57 (25)	197 (72)	111 (35)	36 (9)	147 (44)	344 (116)	54.3×10^6
2000	115 (29)	32 (10)	147 (39)	110 (46)	40 (14)	150 (60)	297 (99)	56.4×10^6
2001	119 (39)	78 (27)	197 (66)	142 (36)	69 (21)	211 (57)	408 (123)	56.1×10^6
2002	156 (52)	77 (25)	233 (77)	131 (40)	69 (30)	200 (70)	433 (147)	55.7×10^6
2003	155 (43)	68 (17)	223 (60)	141 (38)	70 (25)	211 (63)	434 (123)	56.0×10^6
Total	685 (210)	312 (104)	997 (314)	635 (195)	284 (99)	919 (294)	1916 (608)	

sidered as an obstruction limiting the driver's sight distance. On the vertical profile, the road surface at some point on a crest vertical curve was assumed as an obstruction limiting the driver's sight distance. The available sight distances were determined graphically at frequent intervals along the motorway both on the plane and vertical profile. Then the shorter sight lengths were introduced into the analysis for each direction of travel.

3.3. Traffic volume

Traffic flow was extracted from the traffic file of the MMA. This file contained only the daily number of vehicles entering a carriageway through the controlled access points located along the motorway for paying the toll. Since the daily number of vehicles leaving a carriageway through the exit points was not however recorded, an approximate procedure for estimating the annual average daily traffic (AADT) on each road section was followed. By noting that the traffic flow on the motorway is essentially that of commuters, it was assumed that the total number of vehicles leaving a carriageway in a year at a given exit point was equal to the number of vehicles entering the other carriageway in the same year through the corresponding controlled access point in the opposite direction. Thus, the total annual traffic flow at each carriageway input/output point was computed as the algebraic sum of the above-mentioned flows, assuming a positive sign for entrances and a negative one for exits. This total number of vehicles was then divided by the number of days in the year, and the annual average daily flows at each carriageway input/output point were estimated. The AADT on tangents and curves, located between two successive input/output points i and $i+1$, was computed as the sum of the first i above-mentioned traffic flows. AADT values ranging from about 17,600 to 47,400 vehicles per day were found.

3.4. Pavement friction coefficient

Road surface characteristics were defined by pavement friction and expressed in terms of Side Friction Coefficient (SFC), measured by means of a SCRIM equipment.

Every year MMA makes provision for evaluating the SFC in order to check that the pavement friction has not dropped below the threshold permitted for road safety. SFC is measured in continuous along the motorway both for the right and left lanes and for each direction of travel. For estimating the SFC value for each of 265 segments considered (147 for tangents and 118 for curves) in this paper the data provided by MMA was used and the following approximate procedure was applied. The SFC value for each segment of length l was calculated as the weighted mean $\mathrm{SFC} = \sum_{i=1}^{n} l_i \mathrm{SFC}_i / l$, where SFC_i is the pavement friction value on the sub-segment of length l_i (SFC_i being the average value of the friction measured on the two lanes of each travel direction). Using this procedure, SCF values ranging from 0.26 to 0.74 were estimated.

Summary statistics of the above independent variables are given in Table 2.

3.5. Rain

Rain precipitation data were derived from the Functional Hydrogeological Centre of Campanian Region. They consist of millimetres of rain per hour measured by eight weather stations located along the route of the motorway.

Of all the 1916 accidents registered on the motorway in the 5-year monitoring period, 273 are reported by MMA as having occurred on wet pavement. In order to evaluate a potential rain effect on the number of crashes, the amount of time the pavement is wet is estimated using the hourly rainfall data and assumptions about drying time. For this purpose for each of the 265 road segments (curve and tangent), the amount of time the pavement was wet in a year was estimated by summing up the hours of rainfall observed in that year as available from the records of the nearest weather station. Subsequently, this amount of time in hours was transformed into a time-equivalent number of "days with a wet pavement" in each year of the monitoring period. In other words, in this study conventional days are introduced, each day being totally "dry" or "wet", with a number of "dry" and "wet" days in a year proportional to the estimated amount of time the pavement of a road segment was, respectively, "dry" or "wet" during that year. Then, accidents occurred when the pavement surface was "dry" ("wet") are associated with days with a conventional surface status "dry" ("wet"). All remaining days, both "dry" or "wet", have zero accidents. Thus, a data set resulted which consists of the daily number of accidents on each road section in the 5-year monitoring period, along with the conventional daily status ("dry" or "wet") of the pavement surface.

Table 2
Summary statistics of independent variables—south direction carriageway

	Mean	Mode	Standard deviation	Minimum	Maximum
Length (km)	0.350	0.245	0.298	0.069	1.695
Longitudinal slope (%)	0.05	2.42	2.09	−4.37	4.26
Sight distance (km)	0.583	0.200	0.477	0.100	2.335
Curvature (km^{-1})	2.105	0.504	1.463	0.126	4.854
Side friction coefficient	0.473	0.515	0.060	0.286	0.670
AADT/10,000	2.748	1.874	0.972	1.764	4.741

Note. SFC and AADT vary along time and the table was calculated based on 665 observations.

Table 3
Summary statistics of rain data: day counts for all road segments from 1999 to 2003

	Wet pavement	Dry pavement	Row total
With 0 crashes	32,073	449,919	481,992
With at least 1 crash	270	1,628	1,898
Column total	32,343	451,547	483,890

This way of estimating the amount of time the pavement surface was wet is consistent with Brodsky and Hakkert's (1988) assumption about drying time. In fact, the resulting wet time is calculated by assuming on average both a raining time of 30 min and a drying time of 30 min, at any given hour.

Table 3 presents summary statistics for rain data. Day counts for all road segments and the 5-year monitoring period are given, as a function of two categorical variables, namely (day with no accidents; day with at least one accident) and (day with a dry pavement; day with a wet pavement).

As a consequence of the above analyses a data matrix was created containing the following column variables: number of accidents per year and carriageway occurring on tangents or curves as dependent variable; section length, curvature, annual average daily traffic, longitudinal slope, sight distance on curves, side friction coefficient and presence of junctions as independent variables.

In order to analyse the rain effect, a different data matrix was created containing the number of accidents per day as the dependent variable and the same previous independent variables plus a dummy (0, 1) variable for "dry" and "wet" pavement conditions, respectively.

4. Modelling accident counts

A somewhat natural way for describing the fluctuation of accident counts, say Y_i, which occur on a road section i during given time intervals (e.g. different years), is to assume that Y_i is a random variable (r.v.) with the Poisson probability law. Let λ_i be the expected number of accidents per unit of time on section i. It is well known that a Poisson r.v. Y_i has $\mathrm{Var}\{Y_i\} = E\{Y_i\} = \lambda_i$.

In many situations, however, it has been observed that accident counts appear to be "overdispersed" with respect to the theoretical variability consistent with the Poisson model. In order to account for this overdispersion, it is often assumed that the expected number of accidents per unit of time in the Poisson model is a r.v., $\lambda_i\theta$ say, where θ is assumed to be a Gamma variate with $E\{\theta\} = 1$ and $\mathrm{Var}\{\theta\} = 1/\varphi$. Under this assumption, it can be readily shown that Y_i is a Negative Binomial (NB) r.v. with $E\{Y_i\} = \lambda_i$ and $\mathrm{Var}\{Y_i\} = \lambda_i(1 + \lambda_i/\varphi)$, thus allowing for the variance of accident counts to be greater than the mean, provided that $1/\varphi > 0$. For this reason $1/\varphi$ (or φ itself) is often called the "overdispersion parameter".

Road sections do not have only accident counts, however, but also traits such as length, traffic flow, geometric design variables, environmental conditions, etc. Thus, the objective of statistical road modelling is to estimate the expected number of accidents on a given section as a function of its traits. In other words, it is assumed that there exists a "systematic", i.e. causal, component in accident counts, and "explanatory variables" (traits) may account for this non-random component. This implies defining a "regression model" where the explanatory variables (and possibly combinations thereof) act as "covariates".

Let $\mathbf{x}$ be a vector of k covariates and $\boldsymbol{\beta}$ a vector of k (unknown) coefficients. For both the Poisson and the NB model, a regression model of the expected number of accidents is defined by $\lambda = g(\mathbf{x}_i;\boldsymbol{\beta})$ where g(·) denotes a certain function. Not all functions can serve equally well, however. In fact, it is recognized (see, e.g. Hauer, 2004a,b) that the effect of variables that influence the probability of accident occurrence along a significant portion of a segment is more effectively represented by multiplicative terms, whereas the effect of variables that behave as point hazards are more effectively represented by additive terms. Thus, the regression model should in general have both a multiplicative and an additive portion. As to the multiplicative component, the exponential choice appears to be a natural one, in that it ensures that the expected number of accidents is always a positive number.

Longitudinal studies on accident counts generate multiple observations for the same road section at certain time intervals (typically 1 year). The yearly counts contributed by the same section form a "cluster". The potential problem for clustered counts data is that the observations of the same section may not be mutually independent. When independence does not hold, Poisson or Negative Binomial models are inappropriate. Let Y_{ij} be the r.v. describing the accident counts which occur on road section i in year j. In the case of dependence Guo (1996) suggested modeling each individual count Y_{ij} in the cluster i as a mixed Poisson r.v. with mean $\lambda_{ij}\theta_i$, where θ_i is assumed to be a Gamma variate with $E\{\theta_i\} = 1$ and $\mathrm{Var}\{\theta_i\} = 1/\varphi$. Note that in this model the random effect θ_i only varies across clusters, not the member of a cluster. Moreover, the model assumes that, conditional to θ_i, the count r.v.'s Y_{ij} ($j = 1, \ldots, n_i$) are mutually indepen-

dent. Then, on writing down the conditional joint probability function for cluster i, the unconditional joint probability function for Y_{ij} ($j = 1, \ldots, n_i$) is obtained by integrating over θ_i. This joint probability function is known as the Negative Multinomial (NM) distribution with $E\{Y_{ij}\} = \lambda_{ij}$, $\mathrm{Var}\{Y_{ij}\} = \lambda_{ij}(1 + \lambda_{ij}/\varphi)$ and $\mathrm{Cov}\{Y_{ij}, Y_{ik}\} = \lambda_{ij}\lambda_{ik}/\varphi$.

As for the NB model, the expected number of accidents λ_{ij} can be assumed to depend on a vector $\mathbf{x}_{ij}$ of covariates and a vector $\boldsymbol{\beta}$ of (unknown) coefficients, say $\lambda_{ij} = \lambda_{ij}(\mathbf{x}_{ij}; \boldsymbol{\beta})$.

A different tool to take into account the dependence in longitudinal studies is the Generalized Estimating Equation (GEE) model, which provides an extension of GLM models, in that it allows specifying several possible correlation structures (Abdel-Aty and Wang, 2006; Lord et al., 2005; Wang et al., 2006). The simpler Guo (1996) model is used in this paper, however, since we are not primarily interested in analysing the correlation structure.

Generalizations about the standard NB model and NM model have also been proposed, where parameter φ is not the same for all the sections under analysis, possibly depending on covariates and unknown parameters. In particular, in order to avoid the undesirable consequences that arise when road sections differ in length, Hauer (2001) suggested using a model where $\varphi_i = l_i\varphi$, thus assuming that the overdispersion parameter in each section is proportional to the section length l_i.

4.1. Estimation of model parameters

In order to obtain estimates, $\hat{\boldsymbol{\beta}}$ and $\hat{\varphi}$ say, of the unknown parameters $\boldsymbol{\beta}$ and φ, the likelihood function under the governing model (or equivalently its logarithm) is maximized. In the light of the invariance property of maximum likelihood (ML) estimation, $\hat{\lambda}_i = \lambda_i(\boldsymbol{x}_i; \hat{\boldsymbol{\beta}})$ ($\hat{\lambda}_{ij} = \lambda_{ij}(\boldsymbol{x}_{ij}; \hat{\boldsymbol{\beta}})$) is the ML estimate of the expected number of accidents per unit of time. In this paper a Fortran code was implemented which maximizes the likelihood function by using the double precision routine DBCONF of the IMSL® MATH/LIBRARY (1989).

4.2. Selecting the regression model

Given the data set of accident counts and section traits, the first step is to test the presence of "overdispersion", in order to discriminate between the Poisson model and the NB or NM models. On the basis of the ML estimates $\hat{\boldsymbol{\beta}}$ of the $\boldsymbol{\beta}$ regression coefficients under the Poisson model, the null hypothesis that $1/\varphi = 0$ against $1/\varphi > 0$ can be tested by the statistic:

$$Z = \frac{\sum_{i=1}^{N}\{(Y_i - \hat{\lambda}_i)^2 - Y_i\}}{\left\{2\sum_{i=1}^{N}\hat{\lambda}_i^2\right\}^{1/2}} \tag{1}$$

which is asymptotically distributed as a standard Normal r.v. (Dean and Lawless, 1989). Large positive values of Z indicate overdispersion, whereas large negative values indicate underdispersion.

Once the model type is chosen, and in order to decide which subset of the full set of potentially explanatory variables should be included in the regression model, a stepwise forward procedure based on the Generalized Likelihood Ratio Test (GLRT) was used.

4.3. Measuring goodness-of-fit

To measure the overall goodness-of-fit (g.o.f) in Linear Regression Models the so-called coefficient of determination, R^2, is often used.

In the case of Poisson and NB regression models, however, different measures of g.o.f. have also been suggested (see Fridstrøm et al., 1995). In fact, for these models the ML estimation method is usually used. To the extent that one wants to use R^2 statistic as a basis for testing g.o.f., the way the model parameters are estimated becomes relevant, since R^2 is maximized by ordinary Least Squares estimation but not by ML estimation.

Among g.o.f. indexes alternative to R^2, the most "natural" one appears to be the likelihood ratio g.o.f. statistic R^2_{D}. Let D^m be the "scaled deviance" of model m, that is $D^m = -2\ln(L^m/L^N)$, where L^N is the maximum of the likelihood function of the model in which there are as many parameters as there are observations (the so-called full or saturated model). Now, let D^0 denote the scaled deviance of the zero model, i.e. the model with only a constant term and an overdispersion parameter. In Fridstrøm et al. (1995) it is observed that D^0 can be assumed as a measure of the "total variation" present in the sample. Then, the R^2_{D} statistic defined by

$$R^2_{\mathrm{D}} = 1 - \frac{D^m/(N-m)}{D^0/(N-2)} \tag{2}$$

represents an obvious way for measuring the explanatory power of model m. In the same paper, however, it is argued that a more rational index of g.o.f. should measure the fraction of the explained variation as compared to the "systematic" component of variation alone, rather than to the total variation, which also contains the purely (inexplicable) random component. In fact, this allows computing g.o.f. measures that are equally informative, no matter how large the size of random variation may be.

For this purpose, a measure of the amount of systematic variation was introduced, which in a sense represents the upper bound on the amount of variation that one is able to explain. More especially, in the framework of the likelihood ratio g.o.f. statistic, the amount of systematic variation is estimated by

$$P^2_{\mathrm{D}} = 1 - \frac{E\{D^m\}/(N-m-1)}{D^0/(N-2)} \tag{3}$$

where $E\{D^m\}$ is the expected value of the scaled deviance of model m when it degenerates into the Poisson model. Thus, the fraction of "systematic variation" explained by the regression model is given by $R^2_{\mathrm{D}}/P^2_{\mathrm{D}}$.

A further goodness-of-fit tool used in this paper is the Cumulative Residuals (CURE) method described in Hauer (2004a). This method consists in plotting the cumulative residuals for each independent variable. The aim is to observe graphically how well the function fits the data set.

5. Estimation results

5.1. Modelling accident counts for curves

The data set for curves consists of annual number of total and severe accidents registered in $n = 5$ years from 1999 to 2003 on $N = 118$ segments, 59 for each carriageway. The candidate set of explanatory variables is: length (L), curvature ($1/R$), annual average daily traffic (AADT), sight distance (SD), side friction coefficient (SFC) and longitudinal slope (LS). Moreover, from Table 1 it appears that the number of accidents registered on curves during 1999 and especially during 2000 seems to be much smaller than during the remaining years. Since we were not able to assign a specific cause to this trend, dummy variables yr99, yr00, yr01 and yr02 are also considered to capture the potential non-random year effect. The reference year for these dummy variables is 2003. The log-linear regression model $\lambda_i(\mathbf{x}_i; \boldsymbol{\beta}) = \exp\left(\sum_{j=0}^{m} \beta_j x_{ji}\right)$ is assumed for the expected number of counts on section i.

5.1.1. Total accident counts

The stepwise procedure based on GLRT was implemented and the Poisson regression model was fitted as first. The value of the Z statistic (1) was calculated at convergence of the stepwise procedure. Since $Z = 5.052$, there is clear evidence that overdispersion is present and, consequently the NB model and the NM model were considered in the subsequent analysis. Both the assumption that the parameter φ is the same for all the road sections and the assumption that the overdispersion parameter in each section is proportional to the section length were taken into account. The acronyms NBH and NMH are used for this latter situation. When applying Poisson and NB regression models it was assumed that there are $N = 118 \times 5 = 590$ independent observations of accident counts from a Poisson and a NB distribution, respectively. When applying the NM regression model it was assumed that there are $N = 118$ independent clusters, each with $n = 5$ observations from a NM distribution.

For all the models the estimated coefficients have the expected sign. The number of accidents per year and carriageway occurring on curves increases with length, curvature and annual average daily traffic. These accidents decrease, in contrast, with sight distance, side friction coefficient and longitudinal slope. For all the regression models the significant variables affecting accidents are: L, $1/R$ and AADT, whereas SD, SFC and LS variables do not appear to be statistically significant at the 5% level. A non-random year effect on accident counts is also revealed, since variables yr00 and yr99 appear to be highly statistically significant.

Table 4 shows the parameter estimates at convergence of the stepwise procedure for all the models considered in the analysis. Models are ordered by increasing explanatory power, measured in terms of the explained fraction of both total and systematic variation.

The Negative Multinomial model appears to bring about a clear improvement in the fit of data over the Negative Binomial model. When the overdispersion parameter in each section is

Table 4
Parameter estimates and goodness-of-fit measures for curves

	All crashes					Severe crashes				
	Poisson	NB	NBH	NM	NMH	Poisson	NB	NBH	NM	NMH
Constant	0.01243	0.03523	−0.03459	0.09039	−0.07130	−1.33450	−1.33025	−1.39837	−1.32175	−1.45703
Log of the section length (km)	0.84796	0.86109	0.82431	0.88222	0.80311	0.90626	0.91221	0.88783	0.92575	0.86881
Curvature (km^{-1})	0.26772	0.26196	0.26640	0.25718	0.27017	0.32951	0.32744	0.33336	0.32702	0.33793
AADT/10,000	0.31986	0.32274	0.32674	0.31828	0.32660	0.38973	0.39095	0.39903	0.39709	0.40863
Year 1999 (dummy 0, 1)	−0.38248	−0.37102	−0.38406	−0.38278	−0.38151	0.28842	−0.26474	−0.26655	−0.28692	−0.28510
Year 2000 (dummy 0, 1)	−0.68724	−0.69053	−0.69335	−0.68692	−0.68686	−0.69605	−0.68694	−0.69321	−0.69578	−0.69566
Overdispersion parameter		3.623	14.491	4.146	18.254		2.625	9.548	3.443	13.384
Log likelihood	−490.93	−480.54	−479.57	−465.50	−463.60	−379.40	−376.53	−375.43	−372.23	−370.15
R_D^2, explained fraction of total variation	0.123	0.146	0.148	0.181	0.185	0.098	−0.106	0.110	0.121	0.129
R_D^2/P_D^2, explained fraction of systematic variation	0.458	0.542	0.550	0.671	0.687	0.616	0.671	0.696	0.767	0.814

assumed to be proportional to the section length, both NB and NM models increase their explanatory power. The fraction of total variation explained (R^2_D) by the regression model with the highest explanatory power, i.e. the NMH model, is 18.5%. However, it is worth noting that the fraction of systematic variation explained (R^2_D/P^2_D) amounts to 68.7%.

For the NMH model, the sequence of regression models and parameters developed by the stepwise procedure is shown in Table 5 (estimates at model convergence are in bold).

Goodness-of-fit was also examined with the Cumulative Residuals (CURE) methods. The plots of cumulative residuals against each variable were found to oscillate around zero and contained within $\pm 2\sigma^*$ standard deviations, as required.

5.1.2. Severe accident counts

The previous analyses, relative to all crashes occurring on curves, were repeated for severe crashes only. The findings are quite similar to the former ones. The significant variables affecting severe crashes are: L, $1/R$ and AADT. A non-random year effect, due to variables yr00 and yr99, is also revealed. In Table 4 parameter estimates and explained fraction of variation are given. The NMH model was found to be the model with the highest explanatory power. The fraction of total variation explained by the NMH model is 12.9%. However, it is worth noting that the fraction of systematic variation explained amounts to 81.4%.

5.2. Modelling accident counts for tangents

The data set for tangents consists of an annual number of total and severe accidents registered in $n = 5$ years from 1999 to 2003 on $N = 147$ segments, 73 and 74 segments for the two carriageways, respectively. The candidate set of explanatory variables is: length (L), annual average daily traffic (AADT), side friction coefficient (SFC), longitudinal slope (LS) and junctions. As for curves, Table 1 shows that a non-random year effect might be present also for tangents. Thus, dummy variables are also considered to capture the potential year effect. Moreover, since the presence of a junction represents a point hazard, the regression model $\lambda_i(\mathbf{x}_i;\boldsymbol{\beta}) = \exp(\beta_0 x_0)\left[\exp\left(\sum_{j=1}^{m-1}\beta_j x_{ji}\right) + \beta_m x_{mi}\right]$ with both a multiplicative and an additive component is assumed for the expected number of counts on section i. Note that this model is no longer in the class of Generalized Linear Models, since no link function exists which is linear in $\boldsymbol{\beta}$.

5.2.1. Total accident counts

The stepwise procedure based on GLRT was implemented and the Poisson regression model was fitted as first. The value of the Z statistic (1) was calculated at convergence of the stepwise procedure. Since $Z = 7.325$, there is clear evidence that overdispersion is present and, consequently the NB model and the NM model (with their generalizations) were considered in the subsequent analysis. When applying Poisson and NB regression models it was assumed that there are $N = 147 \times 5 = 735$ independent observations of accident counts from a Poisson

Table 5
Stepwise procedure: sequence of models and parameters for NMH model (curves—all crashes)

Step	φ	Constant	Log-length	Curvature	AADT	Yr2000	Yr1999	Slope	Friction	Sight	Yr2001	Yr2002	Log-LKH	GLRT	P value
1	9.382	0.13625											−509.31		
2	11.90	0.91513	0.59609										−500.61	17.4	3.03E−05
3	14.12	0.78127	0.73897	0.15967									−494.63	12.0	5.42E−04
4	18.20	−0.30633	0.80636	0.27726	0.34486								−483.19	22.9	1.72E−06
5	18.18	−0.23371	0.80742	0.27950	0.35244	−0.60843							−469.56	27.3	1.78E−07
6	**18.25**	**−0.07130**	**0.80311**	**0.27017**	**0.32660**	**−0.68686**	**−0.38151**						**−463.60**	**11.9**	**5.54E−04**
7	18.62	0.06550	0.81372	0.24766	0.33554	−0.68671	−0.37999	−0.19534					−462.76	1.67	1.96E−01
8	18.57	−0.06432	0.80088	0.26885	0.32329	−0.68690	−0.38205		−0.02615				−463.24	0.71	3.98E−01
9	18.59	0.25101	0.80430	0.26663	0.31542	−0.70106	−0.35773			−0.61921			−463.22	0.75	3.87E−01
10	18.25	−0.08520	0.80315	0.27027	0.32678	−0.67360	−0.36822				0.03938		−463.52	0.15	7.00E−01
11	18.25	−0.08198	0.80314	0.27025	0.32683	−0.67695	−0.37156					0.02954	−463.55	0.08	7.75E−01

and a NB distribution, respectively. When applying the NM regression model it was assumed that there are $N = 147$ independent clusters, each with $n = 5$ observations from a NM distribution.

For all the models the estimated coefficients have the expected sign. The number of accidents per year and carriageway occurring on tangents increases with length, annual average daily traffic and the presence of a junction, while it decreases with side friction coefficient and longitudinal slope. For all the regression models significant variables affecting accidents are: L, AADT and the presence of a junction, whereas SFC and LS variables do not appear to be statistically significant at the 5% level. A non-random year effect on accident counts is also revealed, since the variable yr00 appears to be highly significant.

Table 6 shows the parameter estimates at convergence of the stepwise procedure for all the models considered in the analysis. Models are ordered by increasing explanatory power.

The Negative Multinomial model appears to bring about a clear improvement in the fit of data over the Negative Binomial model. When the overdispersion parameter in each section is assumed to be proportional to the section length, both NB and NM models increase their explanatory power to an extent which is by far higher than for curves. The fraction of total variation explained (R^2_D) by the regression model with the highest explanatory power, i.e. the NMH model, is 25.9%. The fraction of systematic variation explained (R^2_D/P^2_D) amounts to 60.8%.

For the NMH model, the sequence of regression models and parameters developed by the stepwise procedure is shown in Table 7 (estimates at model convergence are in bold).

The Cumulative Residuals Plots also reflected the model fit.

5.2.2. Severe accident counts

The previous analyses, relative to all crashes occurring on tangents, were repeated for severe crashes only. The findings are similar to the former ones. The significant variables affecting severe crashes are: L, AADT and the presence of a junction. No systematic year effect is revealed, however. Table 6 gives parameter estimates and explained fraction of variation for severe crashes occurring on tangents. It is worth noting that the fraction of systematic variation explained by using the NMH model amounts to 82.3%.

5.3. Modelling rain effect on accident counts

In order to analyse the rain effect on the number of crashes, a data set was constructed which consisted of the number of accidents occurring in each day of the 5-year monitoring period from 1999 to 2003 along with the road pavement status, as was illustrated in Section 3. A regression analysis, based on the Negative Binomial model, was then applied to curves and tangents. The candidate set of explanatory variables for curves and tangents was the same as in Sections 5.1 and 5.2 plus a dummy (0, 1) variable which indicates whether in each day the conventional status of the pavement surface was "dry" or "wet, respectively.

Table 6
Parameter estimates and goodness-of-fit measures for tangents

	All crashes					Severe crashes				
	Poisson	NB	NBH	NM	NMH	Poisson	NB	NBH	NM	NMH
Constant	0.53501	0.53912	0.52609	0.50931	0.50347	−1.37427	−1.36559	−1.35319	−1.38939	−1.40044
Log of the section length (km)	0.86883	0.83745	0.86044	0.81380	0.85729	0.79270	0.77679	0.74731	0.77791	0.76232
AADT/10,000	0.23105	0.22096	0.23198	0.22554	0.23960	0.42610	0.41867	0.40574	0.42685	0.42575
AADT/10,000 × junctions (1 if present, 0 if absent)	0.22172	0.22066	0.23190	0.21715	0.22848	0.44149	0.44513	0.45428	0.47142	0.50628
Year 2000	−0.21343	−0.21404	−0.22190	−0.21327	−0.21344					
Overdispersion parameter		4.227	8.957	5.472	13.772		6.339	8.290	8.957	13.366
Log likelihood	−175.60	−151.116	−146.81	−142.04	−135.83	−513.37	−511.37	−510.01	−507.50	−506.09
R^2_D, explained fraction of total variation	0.203	0.236	0.243	0.250	0.259	0.205	0.207	0.210	0.216	0.219
R^2_D/P^2_D, explained fraction of systematic variation	0.475	0.556	0.571	0.587	0.608	0.762	0.778	0.789	0.811	0.823

Table 7
Stepwise procedure: sequence of models and parameters for NMH model (tangents—all crashes)

Step	φ	Constant	Log-length	AADT	Junction	Yr2000	Friction	Slope	Yr1999	Yr2001	Yr2002	Log-LKH	GLRT	*P* value
1	3.457	0.96813										−216.28		
2	7.502	1.33758	0.90237									−169.20	94.2	2.92E−22
3	11.09	0.39950	0.86380	0.29497								−148.95	40.5	1.96E−10
4	13.78	0.47637	0.85591	0.23404	0.24459							−139.50	18.9	1.38E−05
5	**13.77**	**0.50347**	**0.85729**	**0.23960**	**0.22848**	**−0.21344**						**−135.83**	7.34	6.74E−03
6	13.81	0.54376	0.85839	0.23813	0.21890	−0.21464	−0.07488					−135.82	0.02	8.88E−01
7	13.92	0.50505	0.86022	0.23927	0.22952	−0.21288		−0.02197				−135.40	0.86	3.54E−01
8	13.78	0.53827	0.85640	0.23508	0.22324	−0.23603			−0.09670			−135.08	1.50	2.21E−01
9	13.79	0.52941	0.85584	0.23888	0.22918	−0.24023				−0.11240		−134.81	2.04	1.53E−01
10	13.78	0.48707	0.85516	0.23588	0.24020	−0.18960					0.08981	−135.12	1.42	2.33E−01

5.3.1. Total accidents counts

A total of 273 accidents occurred when the road pavement surface was wet, 98 of which on curves and 175 on tangents. Thus, accidents on curves represented 35.9% of total accidents under wet conditions. A stepwise procedure was implemented and the same variables which were found to be significant on a yearly scale were confirmed to be significant also on a daily scale. In addition, wet pavement was found to be a highly significant factor in increasing the number of crashes. In fact, when road pavement surface is wet, the expected number of crashes increases by a factor 2.32 for tangents and by a factor 2.70 for curves relative to dry surface conditions.

5.3.2. Severe accident counts

Ninety-seven severe accidents occurred when pavement surface was wet, 39 of which on curves and 58 on tangents. Thus, accidents on curves represented 40.2% of severe accidents under wet conditions. A stepwise procedure was implemented and the same variables which were found to be significant on a yearly scale were confirmed to be significant also on a daily scale. In addition, wet pavement was found to be a highly significant factor in increasing the number of crashes. In fact, when the road pavement surface is wet, the expected number of crashes increases by a factor 2.81 for tangents and by a factor 3.26 for curves relative to dry surface conditions. These results also show that a wet surface proportionally increases severe crashes more than damage-only crashes.

It should be noted that the above estimates depend on assumptions about drying time, so that their accuracy might be limited. These results, however, appear to be in accordance with the literature (see, e.g. Eisenberg, 2004) in which relative risks of crash during rain varying from 1.6 to about 4 were reported. Table 8 shows the parameter estimates at convergence of the stepwise procedure for the Negative Binomial model.

6. Model equations

6.1. Curves

For multilane road sections containing curves, the accident-prediction models for each carriageway are:

- total crashes:

$$\hat{\lambda} = \exp[-0.07130 + 0.80311 \ln L + 0.27017 \times 1/R + 0.32660 \times \mathrm{AADT} \times 10^{-4}]$$

- severe crashes:

$$\hat{\lambda} = \exp[-1.45703 + 0.86881 \ln L + 0.33793 \times 1/R + 0.40863 \times \mathrm{AADT} \times 10^{-4}]$$

where $\hat{\lambda}$ is the predicted crashes/year and carriageway, L the curve length in kilometres, $1/R$ the curvature in km^{-1} and AADT is the annual average daily traffic in vehicles/day.

Table 8
Parameter estimates for the regression model with a surface status indicator

	All crashes		Severe crashes	
	Tangents	Curves	Tangents	Curves
Constant	−5.47635	−6.01594	−7.33916	−7.53979
Log of the section length (km)	0.87122	0.84834	0.80626	0.90707
Curvature (km^{-1})		0.26664		0.34003
AADT/10,000	0.24076	0.32652	0.42121	0.41839
Surface status (1 if wet, 0 if dry)	0.84363	0.99436	1.03324	1.18186
AADT/10,000 × junctions (1 if present, 0 if absent)	0.23873		0.44289	
Year 1999 (dummy 0, 1)		−0.37727		
Year 2000 (dummy 0, 1)	−0.19606	−0.66717		−0.61426
Overdispersion parameter	1.160	1.242	0.960	0.773

As an example, for a curve segment with $L = 0.275$ km ($\ln L = -1.291$), $1/R = 2\,km^{-1}$ ($R = 0.500$ km), AADT = 20,000 vehicles/day, one estimates $\hat{\lambda} = 1.089$ total crashes/year and $\hat{\lambda} = 0.338$ severe crashes/year.

6.2. Tangents

For multilane road sections containing tangents, the accident-prediction models for each carriageway are

- total crashes:

$$\hat{\lambda} = \exp(0.50347)[\exp(0.85729 \ln L + 0.23960 \times \text{AADT} \times 10^{-4}) + 0.22848 \times \text{AADT} \times 10^{-4} \times J]$$

- severe crashes:

$$\hat{\lambda} = \exp(-1.40044)[\exp(0.76232 \ln L + 0.42575 \times \text{AADT} \times 10^{-4}) + 0.50628 \times \text{AADT} \times 10^{-4} \times J]$$

where $\hat{\lambda}$ is the predicted crashes/year and carriageway, L the tangent length in kilometres, AADT the annual average daily traffic in vehicles/day and J is the junction (1 if present, 0 if absent).

As an example, for a tangent segment with $L = 1.2$ km ($\ln L = 0.1823$), AADT = 45,000 vehicle/day, in absence of a junction one estimates $\hat{\lambda} = 5.686$ total crashes/year ($\hat{\lambda} = 1.924$ severe crashes/year) and in presence of a junction one estimates $\hat{\lambda} = 6.064$ total crashes/year ($\hat{\lambda} = 2.486$ severe crashes/year). Note that, in presence of a junction the expected number of total crashes increases by a factor of 1.066, whereas the average number of severe crashes increases by a factor of 1.292.

7. Summary and conclusions

The writing of this paper was primarily motivated by the need to quantify, for the first time for Italian multilane roads, the safety effects on the expected number of total and severe crashes of all the following variables: traffic flow, road geometry, sight distance, pavement surface friction and rain precipitation, with a view to suggesting countermeasures for improving road safety. A further point of interest was to detect the more appropriate statistical regression tool for treating this kind of data.

On the basis of the 5-year monitoring period extending from 1999 to 2003 carried out on the four-lane median-divided motorway it may be concluded that the number of both total and severe accidents, per year and carriageway, occurring on curves increases with the length, the curvature and the annual average daily traffic. These accidents decrease, in contrast, with: sight distance, pavement friction and longitudinal slope. The set of significant variables affecting accidents is: L, $1/R$ and AADT, whereas SD, SFC and LS variables do not appear to be statistically significant at the 5% level. The results regarding the effect of the curvature could be used as a guide for designing horizontal curves. The Negative Multinomial regression model with overdispersion parameter proportional to section length proved to be the model with the highest explanatory power. The fraction of total variation explained for total (severe) crashes is 18.5% (12.9%), while the fraction of systematic variation explained amounts to 68.71% (81.4%).

For tangents, the number of both total and severe accidents per year and carriageway increases with: L, AADT and the presence of junctions, whereas it decreases with SFC and LS. The set of significant variables affecting accidents is: L, AADT and junctions, whereas pavement friction and longitudinal slope variables do not appear to be statistically significant at the 5% level. The Negative Multinomial regression model with overdispersion parameter proportional to section length proved to be the model with the highest explanatory power. For total (severe) crashes the fraction of total variation explained is 25.9% (21.9%), while the fraction of systematic variation explained amounts to 60.8% (82.3%). It is interesting to note that the presence of a junction on a road section seems more especially to increase the expected number of severe crashes. These latter findings indicate that road engineers should pay considerable attention to junctions. For example, serious consideration should be given to the use of acceleration and deceleration lanes that are of sufficient length to accommodate speed changes, and the weaving and maneuvering of traffic.

The comparison among Poisson, Negative Binomial and Negative Multinomial distributions showed that: (i) Poisson distribution is inappropriate for modelling the random variation of the number of crashes since there is clear evidence that overdispersion is present; (ii) Negative Multinomial distribu-

tion has a decidedly higher explanatory power than Negative Binomial distribution, thus supporting the hypothesis that the latter model is inappropriate when multiple observations for the same road section at different years are analysed; (iii) both Negative Binomial and Negative Multinomial model increase their explanatory power when the overdispersion parameter in each section is assumed to be proportional to the section length. As a result, Negative Multinomial distribution is suggested as the most appropriate statistical regression tool for modelling longitudinal crash data.

Rain was found to be a highly significant variable, affecting the expected number of total (severe) accidents by a factor 2.32 (2.81) for tangents and by a factor 2.70 (3.26) for curves. Thus, these results seem to show that a wet surface proportionally increases severe crashes more than damage-only crashes. Even if the accuracy of these estimates might be limited as approximations have to be used for exposure to a wet pavement, nevertheless these results clearly indicate that the added risk of an accident in wet pavement conditions is substantial. Besides, by comparing accident percentages it appears that rain tends to increase proportionally the number of accidents occurring on curves more than ones occurring on tangents. The higher number of accidents predicted on a wet pavement should suggest appropriate countermeasures such as pavement resurfacing with porous asphalt as surface course. In fact, porous asphalt prevents many wet-skidding or hydroplaning accidents, besides improving (as a consequence of the rapid removal of water) the visibility of pavement markings, which is generally recognized as an important factor in preventing accidents during rain precipitation.

In the light of the above results, the models developed for Italian motorways in this paper appear to be useful for many applications such as the detection of critical factors (an example in the present case study being the presence of a junction), the estimation of accident reductions due to infrastructure and pavement improvement, and the predictions of accident counts when comparing different design options. Thus, we are reasonably confident that this research may represent a point of reference for the engineers in adjusting or designing multilane roads.

With regard to the above-mentioned fraction of total variation explained it is worth noting that the rather low observed values of R^2_D are not surprising in relationships where the number of accidents is correlated only to traffic flow, roads factors and weather. In fact, there are several variables not considered in the analysis such as lighting, vehicle, and above all human behaviour, that are heavily influential on accidents counts. The major point of interest is "human behaviour" that many researchers consider to be the predominant factor in accidents. Unfortunately, information for quantifying the effects of human behaviour–road interface on accidents is currently wanting. The problem seems complex and only a multi-disciplinary approach with active coordination among disciplines in the field of both road and human behaviour may help. Thus the proposed accident models which for the most part relate to traffic flow and road geometry and rainfall, might be usefully integrated with human component variables, thereby increasing the explained fraction of total variation considerably. Thus, research needs to be addressed to an ever increasing degree through combined actions that take into account both these two predominant factors (man and road).

Acknowledgements

The authors would like to thank the Management Agency of Naples-Salerno Motorway (SAM) for providing data relating to accidents, traffic and pavement surface conditions, and the Functional Hydrogeological Centre of Campanian Region for providing rain precipitation data. The paper benefited from the input of two anonymous referees. Their comments proved invaluable.

References

Abdel-Aty, M.A., Essam Radwan, E.A., 2000. Modeling traffic accident occurrence and involvement. Accid. Anal. Prev. 32, 633–642.

Abdel-Aty, M.A., Wang, X., 2006. Crash estimation at signalized intersections along corridors: analyzing spatial effect and identifying significant factors. In: Proceedings of the TRB 2006 Annual Meeting, TRB 06-0009.

Abdelwahab, H.T., Abdel-Aty, M.A., 2001. Development of Artificial Neural Network Models to Predict Driver Injury Severity in Traffic Accident at Signalized Intersection. Transportation Research Record 1746, Paper No 01-2234, pp. 6–13.

Adeli, H., Karim, A., 2000. Fuzzy-wavelet RBFNN model for freeway incident detection. J. Transp. Eng. 126 (6), 464–471.

Autodesk Inc. AutoCAD® 2000. Copyright© 1999 Autodesk, Inc.

Brodsky, H., Hakkert, A.S., 1988. Risk of a road accident in rainy weather. Crash Anal. Prev. 20 (2), 161–176.

Caliendo, C., Parisi, A., 2005. Principal component analysis applied to crash data on multilane roads. In: Proceedings of Third International SIIV Congress, http://www.sed.siiv.scelta.com/bari2005/080.pdf.

Chiou, Y.C., 2006. An artificial network-based expert system for appraisal of two-car crash accidents. Accid. Anal. Prev. 38, 777–785.

Dean, C., Lawless, J.F., 1989. Tests for detecting overdispersion in Poisson regression models. JASA 84, 467–472.

Delen, D., Sharda, R., Besson, M., 2006. Identifying significant predictors of injury severity in traffic accidents using a series of artificial neural networks. Accid. Anal. Prev. 38, 434–444.

Edwards, J.B., 1998. The relationship between road accident severity and recorded weather. J. Saf. Res. 29 (4), 249–262.

Eisenberg, D., 2004. The mixed effects of precipitation on traffic crashes. Accid. Anal. Prev. 36, 637–647.

Fridstrøm, L., Ifver, J., Ingebrigtsen, S., Kumala, R., Krogsgård Thomsen, L., 1995. Measuring the contribution of randomness, exposure, weather, and daylight to the variation in road accident counts. Accid. Anal. Prev. 27, 1–20.

Golob, T.F., Recker, W.W., 2003. Relationship among urban freeway accidents, traffic flow, weather, and lighting conditions. J. Transp. Eng. 129, 342–353.

Golob, T.F., Recker, W.W., Alvarez, V.M., 2004. Toll to evaluate safety effect of changes in freeway traffic flow. J. Transp. Eng. 130 (2).

Guo, G., 1996. Negative multinomial regression models for clustered events counts. Sociol. Methodol. 26, 113–132.

Hadi, M.A., Aruldhas, J., Chow, L., Wattleworth, J.A., 1995. Estimating safety effects of cross-section design for various highway types using Negative Binomial regression. Transp. Res. Rec., 1500.

Hauer, E., 2001. Overdispersion in modelling accidents on road sections and in Empirical Bayes estimation. Accid. Anal. Prev. 33, 799–808.

Hauer, E., 2004a. Statistical road safety modeling. In: Proceedings of the 83rd TRB Annual Meeting, Washington, DC, USA, January 11–15.

Hauer, E., 2004b. Safety models for urban four-lane undivided road segments. In: Proceedings of the 83rd TRB Annual Meeting, Washington, DC, USA, January 11–15.

Hsiao, C.H., Lin, C.T., Cassidy, M., 1994. Application of fuzzy logic and neural networks to automatically detect freeway traffic incidents. J. Transp. Eng. 120, 753–773.

IMSL® MATH/LIBRARY, 1989. Fortran Subroutines for Mathematical Applications. IMSL.

Keay, K., Sommonds, I., 2005. The association of rainfall and other weather variables with road traffic volume in Melbourne. Aust. Accid. Anal. Prev. 37, 109–124.

Keay, K., Sommonds, I., 2006. Road accident and rainfall in a large Australian city. Accid. Anal. Prev. 38, 445–454.

Knuiman, M.W., Council, F.M., Reinfurt, D.W., 1993. Association of median width and highway accident rates. Transp. Res. Rec., 1401.

Lord, D., Washington, S.P., Ivan, J.N., 2005. Poisson, Poisson-gamma and zero-inflated regression models of motor vehicle crashes: balancing statistical fit and theory. Accid. Anal. Prev. 37, 35–46.

Martin, J.-L., 2002. Relationship crash rate and hourly traffic flow on interurban motorways. Accid. Anal. Prev. 34, 619–629.

Miaou, S.P., Song, J.J., 2005. Bayesian ranking of sites for engineering safety improvements: decision parameter, treatability concept, statistical criterion, and spatial dependence. Accid. Anal. Prev. 37, 699–720.

Mussone, L., Ferrari, A., Oneta, M., 1999. An analysis of urban collision using an artificial intelligence model. Accid. Anal. Prev. 31, 705–718.

Ozbay, K., Noyan, N., 2006. Estimation of incident clearance times using Bayesian Network approach. Accid. Anal. Prev. 38, 542–555.

Persaud, B., Dzbik, L., 1993. Accident prediction models for freeways. Transp. Res. Rec. 1401, 55–60.

Persaud, B., Lyon, C., Nguyen, T., 1999. Empirical Bayes procedure for ranking sites for safety investigation by potential for safety improvement. Transp. Res. Rec. 1665, 7–12.

Persaud, B., Retting, R.A., Lyon, C., 2000. Guidelines for the identification of Hazardous Highway Curves. Transp. Res. Rec. 1717, 14–18.

Sayed, T., Abdelwahab, W., Navin, F., 1995. Identifying accident-prone location using fuzzy pattern recognition. J. Transp. Eng. 121, 352–358.

Shankar, V., Mannering, F., Barfield, W., 1995. Effect of roadway geometrics and environmental factors on rural freeway accident frequencies. Accid. Anal. Prev. 27 (3), 542–555.

Wang, X., Abdel-Aty, M., Brady, P.A., 2006. Crash estimation at signalized intersections: significant factors and temporal effect. In: Proceedings of the TRB 2006 Annual Meeting, TRB 06-0009.