Accident Analysis and Prevention 43 (2011) 1864–1871

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

On factors related to car accidents on German Autobahn connectors

Martin Garnowski[a], Hans Manner[b,*]

[a] FedEx Express Europe, Middle East, Indian Subcontinent and Africa, Cologne, Germany
[b] Department of Economic and Social Statistics, University of Cologne, Meister-Ekkehart-Str. 9, 50937 Cologne, Germany

ARTICLE INFO

Article history:
Received 7 January 2011
Received in revised form 15 April 2011
Accepted 24 April 2011

Keywords:
Highway connectors
German Autobahn
Accident causes
Negative binomial regression
Random parameters

ABSTRACT

We make an attempt to identify factors that explain accidents on German Autobahn connectors. To find these factors we perform an empirical study making use of count data models with fixed and random coefficients. The findings are based on a set of 197 ramps, which we classify into three distinct types of ramps. For these ramps, accident data is available for a period of 3 years (January 2003 until December 2005). The negative binomial model with some random coefficients proved to be an appropriate model in our cross-sectional setting for detecting factors that are related to accidents. The most significant variable is a measure of the average daily traffic. For geometric variables, not only continuous effects were found to be significant, but also threshold effects indicating the exceedance of certain values.

1. Introduction

The traffic on German highways, the so-called “Autobahn”, has been increasing drastically over the past years and is expected to grow further in the future, due to Germany’s central geographical position in Europe. The increase in traffic surpasses not only the economic growth, but also the speed of construction of roads. If the road network is not expanded significantly, the increasing number of vehicles on German Autobahns will certainly lead to an increasing number of accidents. Due to limitations in the potential expansions of the Autobahn, particularly in the short run, an important task is to identify accident factors and their influence on accident probabilities. This information could give suggestions for low-cost, short-term improvements for the prevention of accidents on existing Autobahn segments. One of the most dangerous situations for car drivers on Autobahns is the weaving out of the flow of traffic via a road connector. In the years 2003–2005 nearly 8000 accidents happened on road connectors on Autobahns in the administrative district Düsseldorf, which has an extremely dense Autobahn network and is the region we focus on in this study. Due to the safety-standards on Autobahns “only” 10 of these accidents were fatal, however, the economic damage caused by accidents is remarkable.

Several studies found that about 90% of all accidents are at least partially caused by human failure, see, e.g. Treat et al. (1977). As driver behavior is influenced by the whole environment, the aim of road construction should be to construct road sites that prevent or forgive human errors. The problem with road connectors is that each of them is constructed differently according to distinct traffic volumes or geographical constraints. The key question is which factors cause drivers to make mistakes. The aim of our study is to find a model that explains the number or the probability of accidents at various types of Autobahn connectors. This is a statistical problem. However, due to the nature of the problem, the use of standard linear regression models is inappropriate, as argued by Jovanis and Chang (1986) and Miaou and Lum (1993). The variable of interest, namely the number of accidents during a given time interval, suggests the use of count data models.

Miaou and Lum (1993), who investigated the relationship between truck accidents and roadway geometries, and Pickering et al. (1986) used the Poisson regression model to study accident data. Hauer et al. (1988), on the other hand, introduced the more appropriate negative binomial model to find that traffic flow and various road characteristics have a significant effect on the number of accidents on signalized intersections in Toronto. Another study applying the negative binomial model to determine the causes of car accidents is Shankar et al. (1995), who analyzed accidents on a section of the Interstate 90 near Seattle. Both Poisson and negative binomial models require a cross sectional setting. Chin and Quddus (2003) found that panel count data models have the advantage that they are able to deal with spatial or temporal effects in contrast to cross sectional count data models. They analyzed different types of accidents on signalized intersections in Singapore using a set of variables containing geometric variables, traffic volume variables and regulatory controls. Another paper applying panel data techniques to study accident data is Shankar et al. (1998). As accident data typically tends to have more zero-observations

* Corresponding author. Tel.: +49 221 470 4130; fax: +49 221 470 5084.
E-mail addresses: mgarnowski@fedex.com (M. Garnowski), manner@statistik.uni-koeln.de (H. Manner).

0001-4575/$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.aap.2011.04.026

than are predicted by standard count data models, zero-inflated models have been introduced into traffic accident research. For example, Shankar et al. (1997) investigated accidents on arterials in Washington with two years of accident data and concluded that zero-inflated models have a great flexibility in uncovering processes affecting accident frequencies on roadway sections. For run-off-roadway accidents on a section of a highway in Washington State Lee and Mannering (2002) got promising results applying zero-inflated models in contrast to standard models. However, Washington et al. (2003) and Lord et al. (2005) provide arguments against the use of zero-inflated models in the analysis of accident data. Recently, Anastasopoulos and Mannering (2009) and El-Basyouny and Sayed (2009) introduced count data models with random parameters to account for unobserved heterogeneity and found that these models perform very well. For more on regression models with count data we refer to, among others, Cameron and Trivedi (1986); Lord et al. (2005); Lord and Mannering (2010) and Kibria (2006).

None of the above-mentioned studies analyzes data on highway connectors, but the statistical techniques and explanatory variables that they use are similar to the ones used here. We make an attempt to find an appropriate model for our dataset of 3 years of accidents on Autobahn connectors in the administrative district Düsseldorf (approximately a fifth of the area of North Rhine-Westphalia). In our analysis, we consider more than 60 Autobahn connectors with 197 ramps in an area of approximately 2300 km^2, using traffic data and geometric variables both in continuous form and allowing for threshold effects, which are represented by dummies indicating the exceedance of certain threshold values.

The remainder of the paper is organized as follows. In the next section we describe our methodology. In Section 3 we introduce and explain our dataset, followed by the presentation of the empirical results in Section 4. Finally, Section 5 concludes our paper.

2. Methodology

As our variable of interest, the number of accidents on highway connectors, is a *count variable*, linear regression models are not an appropriate tool for our analysis. Instead, we make use of count data regression models that have been designed for the specific purpose of modeling discrete count variables. The benchmark model for count data is the Poisson regression model, which is derived from the Poisson distribution. We assume a cross sectional setting with n independent observations, the ith of which being $(y_i, \mathbf{x}_i)$, where y_i is the number of occurrences of the event of interest and $\mathbf{x}_i$ is a vector of regressors that determine the number of accidents y_i. The Poisson regression model is defined by

$$f(y_i|\mathbf{x}_i, \beta) = \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!}, \qquad y_i = 0, 1, 2, \ldots \tag{2.1}$$

where $\lambda_i > 0$ is the intensity or rate parameter of observation i. Equation (2.1) measures the probability of y_i occurrences of an event during a unit of time. In this model, the mean and the variance are the same, which is called the *equidispersion*-property of the Poisson distribution. The intensity parameter λ_i is assumed to depend on the regressors through

$$\lambda_i = \exp(\mathbf{x}_i'\beta), \tag{2.2}$$

where the log-linear dependence of λ_i on $\mathbf{x}_i$ assures that the intensity parameter is always positive. It is crucial that the conditional mean equation is correctly specified and that the assumption of equidispersion is satisfied. In the case of overdispersion, maximum likelihood estimation (MLE) t-statistics are inflated, which can lead to too optimistic conclusions about the statistical significance of regressors. The assumption that y_i is Poisson distributed can be relaxed considerably as studied in Gourieroux et al. (1984a,b). Given a correctly specified mean, the pseudo MLE based on a density from the linear exponential family (LEF) is consistent. This allows the assumption of equidispersion to be relaxed either by allowing for specific variance functions or by leaving the form of the variance unspecified. In the latter case standard errors can be obtained by a robust sandwich or bootstrap estimator.

As the assumption of equidispersion is unlikely to hold in reality, a natural extension of the model is to allow for unobserved heterogeneity. Unobserved heterogeneity arises when the covariates do not account for the full amount of individual heterogeneity. An extension of the Poisson model that allows for unobserved heterogeneity and overdispersion is the negative binomial regression (NB) model. The NB model can be obtained by writing

$$\lambda_i = \exp(\mathbf{x}_i'\beta + \varepsilon_i), \tag{2.3}$$

where $\exp(\varepsilon_i)$ follows a gamma distribution with mean 1 and variance α. For this reason it is also called the Poisson-gamma model. The density of the NB distribution is given by

$$f(y_i|\mathbf{x}_i, \beta, \alpha) = \frac{\Gamma(\alpha^{-1} + y_i)}{\Gamma(\alpha^{-1})y_i!}\left(\frac{\alpha^{-1}}{\alpha^{-1} + \lambda_i}\right)^{\alpha^{-1}}\left(\frac{\lambda_i}{\alpha^{-1} + \lambda_i}\right)^{y_i}. \tag{2.4}$$

The variance of this distribution is:

$$\mathrm{V}[y_i|\lambda_i, \alpha] = \lambda_i(1 + \alpha\lambda_i) > \lambda_i.$$

Thus, for $\alpha > 0$, this model allows for overdispersion. The NB regression is also estimated by MLE and, as it is also a member of LEF, it is robust to distributional misspecifications. However, if the model is misspecified the maximum likelihood standard errors are in general inconsistent and either robust sandwich or bootstrap standard errors should be used.

One possibility to allow for heterogeneity across observations (possibly caused by unobserved factors) is to let all or some of the parameters be random. Random parameter count data models for accident data have been proposed by Anastasopoulos and Mannering (2009) and El-Basyouny and Sayed (2009). The random parameters are written as

$$\beta_i = \beta + \varphi_i, \tag{2.5}$$

where φ_i is a random variable with density $g(\cdot)$. The most popular choice is the normal distribution with mean 0 and variance σ^2, which we also use in this paper. Conditional on the random components, the intensity parameters are given by $\lambda_i|\varphi_i = \exp(\mathbf{x}_i'\beta)$ and $\lambda_i|\varphi_i = \exp(\mathbf{x}_i'\beta + \varepsilon_i)$ for the Poisson and negative binomial regression, respectively. The log-likelihood of the random parameter model can be obtained by integrating out φ_i from the joint density of y_i and φ_i:

$$\ln \mathrm{L} = \sum_{i=1}^{n} \ln \int_{\varphi_i} g(\varphi_i) f(y_i|\varphi_i) d\varphi_i. \tag{2.6}$$

As this integral cannot be evaluated analytically and numerical integration is computational infeasible when the number of random parameters goes beyond one or two, Anastasopoulos and Mannering (2009) suggest to evaluate it by simulation. However, instead of using pseudo random numbers the integral above is evaluated using so called scrambled Halton sequences. Halton sequences are non-random sequences that cover the domain of integration more uniformly than random numbers and lead to a more precise evaluation of the integral with fewer draws. We refer to Train (1999) and Bhat (2001, 2003) for details on Halton sequences and their use in simulated maximum likelihood. Note that we treat a parameter as random when its estimated variance is significantly different from zero.

In order to decide between the competing models, it is important to test for overdispersion in the data. Besides comparing the

sample mean and variance, a simple formal test can be performed by noting that the NB model reduces to the Poisson model when $\alpha = 0$. Thus, the null hypothesis of equidispersion can be tested by estimating the NB and Poisson models and performing a likelihood ratio (LR) test for $H_0 : \alpha = 0$. Since α is restricted to be positive the LR statistic asymptotically has probability mass of a half at zero and a half $\chi^2(1)$ distribution above 0. The critical value is then $\chi^2_{1-2\delta}(1)$ when testing at level δ.

The fit of competing models can be measured by the (McFadden) pseudo-R^2 given by

$$1 - \frac{\ln \mathsf{L}_1}{\ln \mathsf{L}_0} \tag{2.7}$$

where $\ln \mathsf{L}_1$ is the log-likelihood of the full model and $\ln \mathsf{L}_0$ is the log-likelihood of the model without any regressors. Additionally, potentially non-nested models can be compared by looking at information criteria. The most popular ones are the Akaike Information Criterion (AIC) proposed by Akaike (1973) and defined as

$$\mathsf{AIC} = -2 \ln \mathsf{L} + 2k, \tag{2.8}$$

and the Bayesian Information criterion (BIC) proposed by Schwarz (1978) and given by

$$\mathsf{BIC} = -2 \ln \mathsf{L} + (\ln n)k, \tag{2.9}$$

where k is the number of parameters in the model. The BIC places a larger penalty on additional regressors and thus leads to the selection of more parsimonious models.

The estimated coefficients of the above models can be interpreted as semi-elasticities and, in contrast to linear regression coefficients, the response does not stay constant with varying regressors. For dummy variables the conditional mean of the dependant variable is $\exp(\hat{\beta}_j)$ times larger if the dummy variable is one. The marginal effect is different for each observation i, but we consider both the average marginal effect ("Avg") and the marginal effect evaluated at the average ("At Avg"), which are calculated as

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\partial \mathsf{E}[y_i|\mathbf{x}_i]}{\partial x_{ij}} = \frac{1}{n}\sum_{i=1}^{n} \beta_j \exp(\mathbf{x}_i'\beta) \tag{2.10}$$

and

$$\left.\frac{\partial \mathsf{E}[y_i|\mathbf{x}]}{\partial x_j}\right|_{\bar{\mathbf{x}}} = \beta_j \exp(\bar{\mathbf{x}}'\beta), \tag{2.11}$$

respectively, where $\bar{\mathbf{x}}$ denotes the vector of sample means of the regressors.

3. Data

In this section we describe important terms and introduce our dataset. Further details about the data are available from the authors upon request.

3.1. Clarification of terms

We distinguish between the terms connector, ramp and curve. Fig. 1 shows a schematic picture of a connector for two Autobahns. This connector consists of 8 ramps and each ramp consists of at least one curve. One has to distinguish between different types of connectors, namely *clover leaves* which connect two Autobahns as in Fig. 1 and *diamond interchanges* which connect an Autobahn with a minor road as in Fig. 2. The schematic pictures present only two possible shapes of connectors as their form varies due to construction constraints.

The ramps can be categorized as tangents, loops, egress-ramps and drive-up ramps. Drive-up ramps are not considered here, because many variables that can be gathered for the other types of ramps cannot be found for this type. Furthermore, significantly fewer accidents occur on these ramps due to a lower permitted driving speed. Thus, we only consider ramps on which cars leave a specific Autobahn and change onto another Autobahn or a minor road. For the sake of simplicity tangent ramps, loop-ramps and egress-ramps will be called T-ramps, O-ramps and E-ramps, respectively.

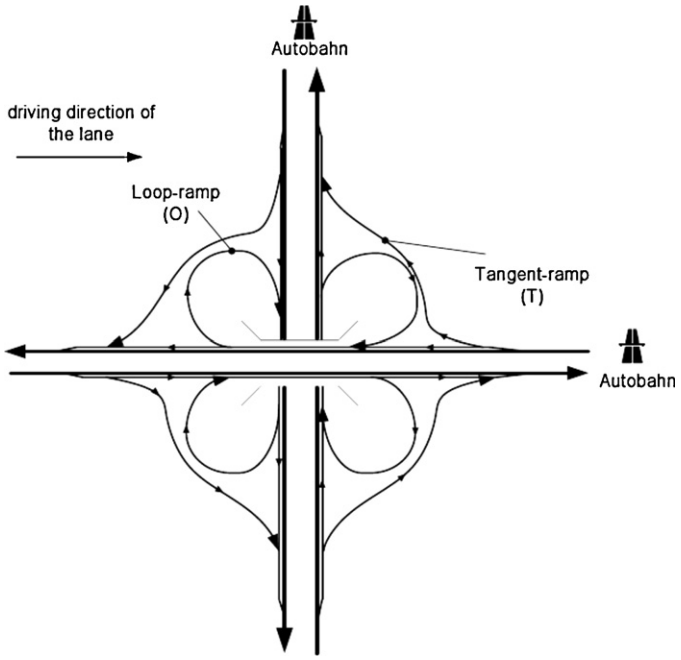

Fig. 1. Clover leaf.

3.2. Data description

Our data contains the following information: details on individual accidents, traffic flow data and geometrical properties of the ramps. We describe each type in turn.

Accident data: The raw data on accidents was provided by the "Autobahnpolizei Düsseldorf", the highway police for the district Düsseldorf. The dataset contains detailed information for all reported accidents on Autobahns in the administrative district in the time period from January 2003 to December 2005. This amounts to a total of 39,032 accidents (12,887 in 2003, 13,433 in 2004 and 12,712 in 2005). This detailed information includes the exact time of the accident, its location, type of vehicle, number of vehicles involved, the severity and type of the accident, information on the driver, sight conditions, and road sleekness. Out of nearly 40,000 reported accidents, we filtered out the accidents that happened on ramps of the various connectors. Given the constraints imposed by the datasets we had a total of 197 ramps under investigation of which 95 were E-ramps, 33 were O-ramps and 69 were T-ramps. After the filtering process a total of 3048 accidents were analyzed. Due to the fact that we are interested in modeling the aggregated counts over a certain period of time at a particular location, the detailed information on the accidents is disregarded in this study.

Table 1 shows some descriptive statistics for accidents on the different types of ramps. The significant difference in the mean of accident numbers for the different types of ramps is eye-catching and suggests the existence of heterogeneity across ramp types.

A phenomenon that is often present in accident data analysis is the predominance of zero-observations, which calls for the use of zero inflated models, i.e. count data models that explicitly account for the presence of a large number of zero-observations. Lord et al.

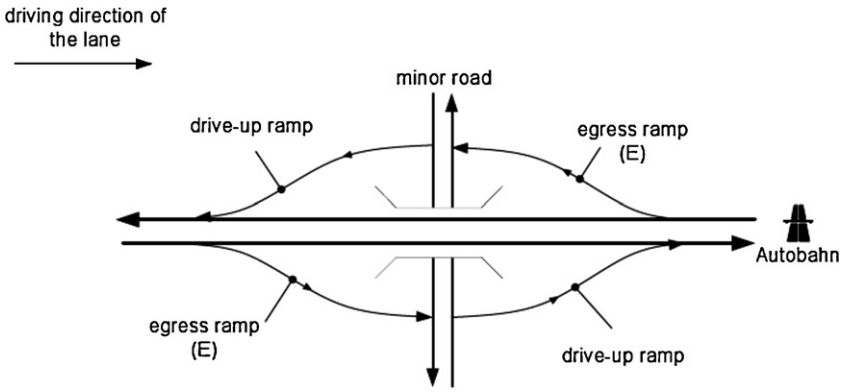

Fig. 2. Diamond interchange.

Table 1
Descriptive statistics, accidents, whole sample period.

Variable	All ramps	E-ramps	O-ramps	T-ramps
Min number of accidents on a ramp	0	0	0	0
Max number of accidents on a ramp	103	76	99	103
Mean number of accidents per ramp	15.32	10.97	15.86	21.03
Mode	4	4	2	4
Median	9	7	11	12
Standard deviation	18.53	12.01	18.56	23.84
Variance	343.33	144.20	344.48	568.44
Inter quartile range	15	12	18	25
Total number of accidents	3048	1042	555	1451
Number of ramps	197	95	33	69
Number of zero-accident ramps	10	5	1	4

(2005) concluded that excess zeros indicate an inappropriate choice of time scale. Referring to Table 1, for neither ramp-type the mode is zero. We have only 10 zero-observations out of the 197 ramps (5 E-ramps, 1 O-ramp and 4 T-ramps). On the other hand, on 35 ramps, 25 or more accidents happened in our sample period of three years. Furthermore, Cheng and Washington (2005) found that three years of crash-history data provides an appropriate crash history duration, which initially motivated our choice of the data period. Therefore we decided against the use of zero inflated models in our analysis.

Traffic flow data: The traffic flow data was provided by the "Landesbetrieb Straßenbau Nordrhein-Westfalen", the institution responsible for the planning, construction and maintenance of the Autobahn in the area of North Rhine-Westphalia.

The traffic volume is counted automatically by induction loops that additionally recognize the length of passing cars and categorizes them into the groups of cars smaller or larger than 7.50 m. We will use the term passenger cars for the former and trucks for the latter group.

The raw dataset contained daily data for all induction loops on Autobahns in the administrative district Düsseldorf for the period from the 4th of March 2005 to the 7th of March 2006, from which the traffic volume on the connectors of interest was extracted. Note that the sample period is not exactly the same as that of the accident data. However, we believe that the available data captures the essential information on the amount of traffic and can be used without any reservations.

For our cross sectional analysis we calculated the average daily traffic (abbreviated ADT) for each ramp. For ramp i the ADT is defined as:

$$\mathrm{ADT}_i = \frac{1}{T_i}\sum_{t=1}^{T_i} \mathrm{TV}_{i,t}, \tag{3.1}$$

where $\mathrm{TV}_{i,t}$ is the traffic volume on ramp i on day t and T_i is the number of daily observations available for ramp i. We also calculated the percentage of trucks on each ramp. Table 2 shows some descriptive statistics for the variable ADT. It is imminent that there is a large spread in the distribution of traffic volume on the various ramps.

Geometry data: The geometry data was collected manually by using aerial photos of the ramps of interest. Details on how the variables were constructed are available from the authors upon request. Table 3 presents a list of the geometry variables along with their descriptive statistics.

Next, a number of dummy variables were constructed that are shown in Table 4.

Finally, we have information on the surface types on the ramps. Three different types of surfaces have to be distinguished, namely, melted or mastic asphalt (MA), found on 57 ramps, mastic asphalt using chipping (MAC), found on 136 ramps, and asphaltic concrete

Table 2
Descriptive statistics ADT, whole sample period.

Average daily traffic (ADT)	
All vehicles ($\mathrm{ADT_{Total}}$)	
Mean	6630.8
Maximum	28,178.47
Minimum	579.26
Standard deviation	4579.36
Passenger cars ($\mathrm{ADT_{PC}}$)	
Mean	5173.95
Maximum	25,094.72
Minimum	472.57
Standard deviation	3582.56
Trucks ($\mathrm{ADT_{Trucks}}$)	
Mean	1456.85
Maximum	15,844.4
Minimum	51.94
Standard deviation	2051.49

Table 3
Descriptive statistics of geometry variables.

Variable	Minimum	Maximum	Mean	Stand.Dev.
Length of the ramp (m)	49.56	1868.15	335.12	243.95
Length of the deceleration lane (m)	0.00	902.24	258.00	158.08
Total width of the lanes on the ramp (m)	3.30	8.19	4.93	1.13
Width per lane (m)	3.02	7.16	4.33	0.73
Width of the shoulder lane (m)	0.00	4.81	2.14	1.40
Radius steepest curve (m)	28.43	1428.47	164.78	195.56
Total deflection angle (°)	0.00	305.15	115.78	92.75
Absolute total deflection angle (°)	4.32	305.15	148.29	83.97
Angle of the steepest curve (°)	3.76	302.00	119.86	86.47
Length of the steepest curve (m)	30.05	1259.25	205.49	155.90
Number of lanes on the ramp	1	2	1.17	0.38
Number of inflection points on the ramp	0	4	0.59	0.86
Position of the steepest curve on the ramp	1	4	1.45	0.72
Number of curves on the ramp	1	6	1.72	0.93
Number of lanes on the autobahn leaving	1	3	2.48	0.51

Table 4
Descriptive statistics dummy variables.

Variable	Mean	Stand.Dev.
Right Autobahn lane becomes decel. lane	0.06	0.24
A curve gets steeper	0.24	0.43
A curve gets less steep	0.13	0.34
Incline on the ramp	0.50	0.50
Decline on the ramp	0.54	0.50
Trees inside	0.87	0.33
Trees outside	0.78	0.42
Crossing lane at the access of the ramp	0.22	0.42
Crossing lane at the exit of the ramp	0.22	0.41
Median between Autobahn and decel. lane	0.38	0.49

(AC) found on 4 ramps. Without going into details, according to Richter and Heindel (2004) the main advantages of the different surface types are: MA is known to have a good grip, with the trade-off of being a loud surface type. AC has the worst grip, but it is the cheapest of the three surface types. Roads build out of MAC are very durable and thus MAC is perfect for roads with high traffic density.

3.3. Missing information

There are a number of potential variables that may have been useful for the analysis but could not be collected. First of all, we did not have any information on traffic signs located on the connectors. On German Autobahns as well as on ramps there is no mandatory speed limit. However, on several ramps there are speed limit signs. "Slippery road"-signs can also be found on several ramps.

Next, the aerial photos we used for the collection of our geometry variables did not give us any information on standing guardrails on the ramps. However, the effect of guardrails on accidents is a complex issue since guardrails may increase the frequency of property damage, but reduce the severity of accidents.[1]

Another shortcoming of our dataset was the short period of our traffic flow variable, not allowing us for a substantiated panel data approach and hence making us ignore information on seasonality. Since a panel approach was not conducted, weather variables have not been used in our research. As the area of the connectors investigated is quite small (around 2300 km^2), we think there are no significant differences in weather conditions for aggregated data. Nevertheless, in a panel approach it might have been interesting to take weather into consideration, not only to explain cross-sectional differences, but also to explain seasonality.

Finally, the information we had on the road surface was also quite rudimentary. Without doubt, not only the type of the surface, but also the age of the surface has an impact on the grip of the road and thus an influence on the accident frequency.

[1] We would like to thank an anonymous referee for pointing this out.

4. Empirical results

In this section we present the results of our empirical analysis. Section 4.1 presents the estimation result. Their interpretation can be found in Section 4.2.

Before we present our results we would like to note the following. First, Table 5 presents all variables we use along with their abbreviations. Second, for the regressions, the unit of measurement was changed to thousands of cars for all ADT-variables and to hundreds of meters for the radius in order to avoid unreadably small regression results.

Table 5
Abbreviations of variables in the dataset.

Variable	Abbrevation
Number of accidents	nb_acc
Length of the ramp	length_ramp
Length of the deceleration lane	length_decel
Total width of the lanes on the ramp	width_lanes
Width per lane	width_per_lane
Width of the shoulder lane	width_should
Radius steepest curve	radius
Total deflection angle	angle_tot
Absolute total deflection angle	angle_abs
Angle of the steepest curve	angle_steepest
Length of the steepest curve	length_steepest
Number of lanes on the ramp	nb_lanes
Number of inflection points on the ramp	nb_infl
Position of the steepest curve on the ramp	pos_steepest
Number of curves on the ramp	nb_curves
Number of lanes on the Autobahn leaving	nb_autob_lanes
Dummy – Right Autobahn lane becomes decel. lane	D_autob_decel
Dummy – A curve gets steeper	D_steeper
Dummy – A curve gets less steep	D_less_steep
Dummy – Incline on the ramp	D_incline
Dummy – Decline on the ramp	D_decline
Dummy – Trees inside	D_trees_in
Dummy – Trees outside	D_trees_out
Dummy – Crossing lane at the access of the ramp	D_cross_access
Dummy – Crossing lane at the exit of the ramp	D_cross_exit
Dummy – Median between Autobahn and decel. lane	D_median
Average daily traffic passenger cars	ADT_pc
Average daily traffic trucks	ADT_trucks
Truck percentage	truck_perc
Threshold Dummy – Length deceleration lane	$T\_length\_decel_{threshold}$
Threshold Dummy – Width of the lanes	$T\_width\_lanes_{threshold}$
Threshold Dummy – Position of the steepest curve	$T\_pos\_steepest_{threshold}$
Threshold Dummy – Radius of the steepest curve	$T\_radius_{threshold}$

Table 6
Estimation results.

Variable	Fixed parameters model		Random parameters model	
	Coefficient	Std. Errors	Coefficient	Std. Errors
Constant	−0.9091	0.4791	−1.2067	0.4690
ln(ADT_pc)	0.7363	0.0988	0.8130	0.1030
truck_perc	0.8954	0.4194	1.2078	0.4114
ln(angle_abs)	0.2685	0.0840	0.2838	0.0780
D_steeper	0.3435	0.1661	0.2364	0.1776
σ_{RC}			*0.5989*	*0.1925*
T_length_decel$_{180}$	0.4831	0.1409	0.4352	0.1382
σ_{RC}			*0.5071*	*0.1381*
T_width_lanes$_{3.90}$	0.4333	0.1824	0.4057	0.1942
T_pos_steepest$_1$	0.3115	0.1470	0.2246	0.1527
σ_{RC}			*0.5012*	*0.1765*
Dispersion parameter	0.6267	0.0719	0.2909	0.1160
ln L	−689.86		−685.99	
AIC	1397.7		1395.9	
BIC	1427.3		1435.4	
Pseudo-R^2	0.0672		0.0725	

4.1. Estimation results

We consider the negative binomial (NB) regression both with fixed and with (possibly) random coefficients (RC). A coefficient is treated as random when its estimated standard deviation σ_{RC} is significantly different from 0. Since the negative binomial model nests the Poisson regression model when $\alpha = 0$ the null hypothesis of a Poisson Model, and hence the null hypothesis of equidispersion, can be tested using the LR test as described in Section 2. We computed this LR statistic for all estimated models and it turned out that the test has p-values of virtually zero for all the model specifications we considered. An exception is the RC model with random intercept, in which case the fit of the Poisson regression is almost identical to the fit of the NB model. This is not surprising as the NB model and the Poisson model with random intercept are very similar differing only in the error distributions.

Preliminary estimates of the NB regression reveal that from our initial set of variables only ADT_pc, ln(angle_abs), truck_perc and D_steeper are statistically significant. Alternatively, the radius can be used instead of ln(angle_abs), as these two variables have a correlation of about −0.76. In the next step of our analysis, we test for the different functional forms of the regressor ADT_pc in order to allow for nonlinearities. We consider two variations, first by adding a quadratic term to the initial model and second by working with the natural logarithm of ADT_pc. Both choices clearly improve the model fit as indicated by the pseudo-R^2 and information criteria. However, there is no clear evidence whether the squared form is to be preferred over the logarithmic form. In order to get an additional evaluation criterion we conducted the LR-test for non-nested models proposed by Vuong (1989). The null hypothesis of equivalence of the models cannot be rejected in favor of any of the specifications. However, as it has one parameter less to estimate we continue working with the logarithmic form.

Next, we want to investigate the heterogeneity across the different ramp types suggested by the descriptive statistics. Ideally, a separate regression should be estimated for each ramp type, but given the small number of observations for the individual ramp types the results would not be reliable. Therefore we consider dummy variables for each ramp type (skipping the intercept to avoid multicollinearity) and perform a likelihood ratio test to see if the model fit improved. The improvement in fit is not significant so we dropped the dummies from the model.

As none of the remaining variables were found to have additional significant effects in explaining accident frequencies we investigate whether there are possible threshold effects. By this we mean the effect when variables are larger than some (predetermined) value and we consider these to capture potential nonlinearities. To identify the thresholds we create dummy variables that for values exceeding the supposed threshold value and zero otherwise. By varying the threshold value and comparing the information criteria and the pseudo-R^2 we try to determine the actual value of the threshold. The notation of the threshold variables is as follows: A "T" in front of the variable indicates that this is a dummy measuring the threshold and the index-number shows the value of the threshold. This means that the variable is 1 for values larger than the threshold and zero otherwise. Three variables seem to have a threshold effect, namely the length of the deceleration lane (length_decel), the total width of the lane(s) (width_lanes) and the position of the steepest curve (pos_steepest).

We would like to note two things concerning our final model: (i) the variable width_lanes measures the total width of the officially accessible lanes without accounting for the width of a possible shoulder lane. We also investigated a possible threshold effect of the total width of the official road together with the shoulder lane (i.e. the whole possibly accessible road). However no significant effect could be found. (ii) the variable T_pos_steepest$_1$ takes on the value one if the first curve is not the steepest curve on the ramp.

The following step in the analysis is to allow for random coefficients. We estimate the model using 500 Halton draws to simulate the log-likelihood function. Initially, all model parameters are allowed to vary randomly, but we treat parameters as fixed when the estimated variance of the random coefficient σ_{RC} is not significantly different from 0. Only the variables D_steeper, T_length_decel$_{180}$ and T_pos_steepest$_1$ have random coefficients. Furthermore, in the random coefficient model no additional variables from our dataset have a significant (fixed or random) effect on the number of accidents.

Table 6 presents the estimation results of our model with fixed and random coefficients. First of all, it is noticeable that the random coefficient specification only leads to a small improvement in the model fit, which is in contrast to the findings of Anastasopoulos and Mannering (2009). Nevertheless, a likelihood ratio test suggests that the improvement in fit is significant with a p-value of about 0.05. The BIC, on the other hand, indicates that the simpler version of the model with fixed parameters is more suitable. Furthermore, the estimated dispersion parameter α of the binomial distribution is much smaller for the random coefficient model, which is likely due to the fact that the heterogeneity of the random coefficients picks up some of the heterogeneity originally captured by the dispersion parameter. The parameter estimates are mostly very similar across the models, with the exception that the parameter of the variable truck_perc changes when allowing for random coefficients. How-

Table 7
Marginal effects.

Variable	Fixed parameters model			Random parameters model		
	Coefficient	Avg	At Avg	Coefficient	Avg	At Avg
ln(ADT_pc)	0.74	11.14	8.89	0.81	10.53	8.32
truck_perc	0.90	13.55	10.82	1.21	15.64	12.36
ln(angle_abs)	0.27	4.06	3.24	0.28	3.67	2.90
D_steeper	0.34	5.20	4.15	0.24	3.06	2.42
$T\_length\_decel_{180}$	0.48	7.31	5.84	0.44	5.64	4.45
$T\_width\_lanes_{3.90}$	0.43	6.56	5.23	0.41	5.25	4.15
$T\_pos\_steepest_1$	0.31	4.71	3.76	0.22	2.91	2.30

ever, for the variables that have random coefficients, the estimates are a bit lower compared to their fixed counterparts.

4.2. Interpretation

In Table 7 we report both the average marginal effect, as well as the marginal effect evaluated at the averages of the regressors. We only interpret the model with random coefficients, since the overall model fit is better and the parameter estimates are quite close for most of the variables.

The variable that has the strongest effect on the number of accidents is the measure of average daily traffic. Since the variable ADT_pc enters in logarithmic form it means that a 1% increase in average daily traffic on average leads to a 0.81% increase in the expected number of accidents. An increase of truck_perc by 0.1 units, hence an increase of the truck ratio by 10 percentage points leads to an increase in the expected number of accidents by 12.08%. The mean effect "Avg" suggests that the average effect of a 10 percentage point increase in the ratio of trucks yields 1.56 more expected accidents per ramp in a time period of three years. The "Avg" estimates for this model are about 26% higher than those of the "representative" ramp given in "At Avg", which is due to the convex exponential mean function. The same phenomenon can therefore be observed for the other variables. The remaining variables are more interesting from an engineering perspective. The estimated coefficient of the variable ln(angle_abs) implies that a 1% increase in the absolute total deflection angle leads to a 0.27% increase in the expected number of accidents. Next, a ramp on which a curve gets steeper on average has an $\exp(0.24) = 1.27$ times higher expected number of accidents than a ramp on which no curve is getting steeper. In absolute terms you expect on average 3.06 more accidents over a period of three years when a curve on a ramp gets steeper. The distribution of the random coefficient implies that in 35% of the cases the effect of a curve getting steeper actually reduces the number of accidents, but for the majority of cases this has a positive effect.

The estimates of the remaining variables are rather counterintuitive. The estimate of 0.4352 for the coefficient of the variable $T\_length\_decel_{180}$ suggests that $\exp(0.43) = 1.54$ times more accidents occur on ramps with a deceleration lane larger than 180 m in contrast to ramps with a deceleration lane that is smaller than 180 m. However, 20% of the parameter distribution is less than 0 and thus of the expected sign. Similarly, the expected number of accidents on ramps with lane widths exceeding 3.9 m is 1.5 times higher than on ramps with lane widths of less than 3.9 m. Finally, the estimate of the threshold variable $T\_pos\_steepest_1$ suggests that we can expect the number of accidents to be 1.25 times higher, if the first curve is not the steepest on the ramp. In this case, about 33% of the distribution is smaller than zero. One reason for these three counterintuitive results might be that an unsafe looking ramp causes the driver to be more aware. Another explanation is an omitted variable bias, since it is likely that there are stricter speed restrictions (or warning sings) on these unsafe ramps, which in would turn decrease the number of accidents. However, part of the possible problem of omitted variable bias is handled by allowing for random parameters.

5. Conclusion

The aim of this paper was to find an appropriate statistical model that helps to explain accident frequencies on Autobahn connectors in Germany. The negative binomial regression model with random coefficients turned out to be the most appropriate tool. Our dataset contains detailed information on accidents on 197 connectors, data on traffic flow of passenger cars and trucks, and a set of nearly 30 geometry variables. Although the three types of ramps, egress-ramps (n =95), tangent-ramps (n =69) and loop-ramps (n =33), have quite distinct characteristics, dummy variables indicating ramp types are not significant and separate models cannot be estimated due to the small number of observations in the sub-populations.

The most important variable explaining accident frequencies is the average daily traffic. The fraction of trucks is also an important factor. Beyond that, measures of the curvature, like the radius of the steepest curve or a dummy indicating that a curve gets steeper, turned out to be relevant. The variables measuring the length of the deceleration lane, the width of the lanes and the position of the steepest curve only have an effect when exceeding certain thresholds.

The final question is whether and how our findings can be used to improve the safety of existing Autobahn connectors and to give recommendations in the construction of new ones. The accident factors identified in this study cannot be used for simple short term improvements, which is not surprising, as the easiest and cheapest solution one can think of is putting up additional warning signs and speed restrictions. Since we had no data available on these, and it is likely that these measures were already taken on relatively dangerous connectors, we have to leave the recommendation of simple solutions to future research that includes such information. However, if connectors are built from scratch, our results might be helpful. The finding that curves getting steeper yield higher accident frequencies is definitely a result that future planning should not disregard. The radius effect might also be interesting for the design of ramps. This can be simplified as: *the steeper a curve, the more accidents can be observed*. The significant positive parameter of the variable absolute total deflection angle can be interpreted as: *the simpler a ramp is constructed, the less accidents can be expected*.

In future research we need to deal with the data limitations encountered. Information on speed limitations, vertical grades on the ramps or exit approaches, warning signs and pavement friction is likely to add significant explanatory power to the model. Furthermore, one may try to expand the dataset and consider temporal and dynamic effects, possibly in a panel framework. Finally, the small number of observations is another limitation of the current research, as it was not possible to estimate models for the different ramp types separately.

Acknowledgments

The authors would like to thank Jean-Pierre Urbain, Frowin Schulz and three anonymous referees for valuable comments. The authors would also like to thank the Landesbetrieb Straßenbau Nordrhein-Westfalen and the Autobahnpolizei Düsseldorf for providing the raw data.

References

Akaike, H., 1973. Information theory and an extension of the maximum likelihood principle. In: Petrov, B., Csaki, F. (Eds.), Second International Symposium on Information Theory. , pp. 267–281.

Anastasopoulos, P.C., Mannering, F.L., 2009. A note on modeling vehicle accident frequencies with random-parameters count models. Accident Analysis and Prevention 41, 153–159.

Bhat, C., 2001. Quasi-random maximum simulated likelihood estimation of the mixed multinomial logit model. Transportation Research Part B 17, 677–693.

Bhat, C., 2003. Simulation estimation of mixed discrete choice models using randomized and scrambled halton sequences. Transportation Research Part B 37, 837–855.

Cameron, A.C., Trivedi, P.K., 1986. Econometric models based on count data: comparisons and applications of some estimators. Journal of Applied Econometrics 1, 29–53.

Cheng, W., Washington, S.P., 2005. Experimental evaluation of hotspot identification methods. Accident Analysis and Prevention 37, 870–881.

Chin, H., Quddus, M., 2003. Applying the random effect negative binomial model to examine traffic accident occurrence at signalized intersections. Accident Analysis and Prevention 35, 253–259.

El-Basyouny, K., Sayed, T., 2009. Accident prediction models with random corridor parameters. Accident Analysis and Prevention 41, 1118–1123.

Gourieroux, C., Monfort, A., Trognon, A., 1984a. Pseudo maximum likelihood methods: applications to poisson models. Econometrica 52, 701–720.

Gourieroux, C., Monfort, A., Trognon, A., 1984b. Pseudo maximum likelihood methods: theory. Econometrica 52, 681–700.

Hauer, E., Ng, J., Lovell, J., 1988. Estimation of safety at signalized intersections. Transportation Research Record 1185, 48–61.

Jovanis, P.P., Chang, H.-L., 1986. Modeling the relationship of accidents to miles traveled. Transportation Research Record 1068, 42–51.

Kibria, B.M.G., 2006. Applications of some discrete regression models for count data. Pakistan Journal of Statistics and Operation Research 2, 1–16.

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

Lord, D., Mannering, F.L., 2010. The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives. Transportation Research Part A 44, 291–305.

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

Miaou, S., Lum, H., 1993. Modeling vehicle accidents and highway geometric design relationships. Accident Analysis and Prevention 25, 689–709.

Pickering, D., Hall, R., Grimmer, M., 1986. Accidents on rural t-junctions. Tech. rep., Research Report 65. Transportation and Research Laboratory, Department for Transport, Crowthrone, Berkshire, United Kingdom.

Richter, D., Heindel, M., 2004. Straßen- und Tiefbau. Mit lernfeldorientierten Projekten. Teubner.

Schwarz, G., 1978. Estimating the dimension of a model. Annals of Statistics 6, 461–464.

Shankar, V., Albin, R., Milton, J., Mannering, F., 1998. Evaluating median cross-over likelihoods with clustered accident counts: an empirical inquiry using the random effects negative binomial model. Transportation Research Record, 1635.

Shankar, V., Mannering, F., Barfield, W., 1995. Effect of roadway geometrics and environmental factors on rural freeway accident frequencies. Accident Analysis and Prevention 27, 371–389.

Shankar, V., Milton, J., Mannering, F., 1997. Modeling accident frequencies as zero-altered probability processes: an empirical inquiry. Accident Analysis and Prevention 29, 829–837.

Train, K., (1999). Halton sequences for mixed logit. Working paper. University of California, Department of Economics, Berkeley.

Treat, J., Tumbas, N., McDonald, S., Shinar, D., Hume, R., Mayer, R., Stanisfer, R., Castellan, N., 1977. Tri-level study of the causes of traffic accidents. Report No. DOT-HS-034–3-535–77 (TAC).

Vuong, Q., 1989. Likelihood ratio tests for model selection and non-nested hypothesis. Econometrica 57, 307–334.

Washington, S., Karlaftis, M.G., Mannering, F.L., 2003. Statistical and Econometric Methods for Transportation Data Analysis. CRC, Chapman and Hall.