Comparison and analysis of road tunnel traffic accident frequencies and rates using random-parameter models

Ciro Caliendo[a,*], Maria Luisa De Gugliemo[a], Maurizio Guida[b]

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

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

Abstract

The paper provides an analysis of severe crashes (fatal and injury accidents only) occurred in 260 Italian road tunnels on the basis of random-parameters regression models. In particular, factors influencing "crash frequency" over a four-years monitoring period are investigated by means of a random-parameters Negative Binomial model (RPNB). This approach reveals that some regression model parameters should be considered as random variables, thus contributing to gaining new insights into the way the corresponding covariates influence accident frequency.

The presence of a year effect on the number of accidents was also investigated. To this aim, random-parameters models with random-effects were considered to account for temporal correlation in the data collected from the same road tunnel over successive time periods. In particular, a Negative Binomial model in which both the intercept and regression parameters are allowed to vary randomly (RPINB) was developed. It was found that in the present case study the unobserved heterogeneity is small, so as it is sufficient just using the random-intercept model.

This study also explores the use of crash rate instead of crash frequency as dependent variable of the regression model, in order to understand more especially the effect of this choice on the significance of the covariates. A random-parameters tobit regression model showed that the significance of some independent variables is in contrast with the results found upon the RPNB model based on crash frequency.

Keywords: Random-parameters Negative Binomial, Random-parameters Poisson, Crash frequency, Tobit regression, Crash rate.

(*) Address correspondence to Ciro Caliendo, Department of Civil Engineering, University of Salerno, via Giovanni Paulo II, 132, Fisciano, Salerno 84084, Italy. E-mail Address: ccaliendo@unisa.it (C. Caliendo), Tel.: +39+89+964140; fax: +39+89+964045

1. Introduction

A variety of statistical methods have been applied over the years to accident data analysis for developing prediction models and assessing the significance of factors influencing crashes. Lord and Mannering (2010), Savolainen et al. (2011), Mannering and Bhat (2014) provide a listing of the methods applied with their strengths and weaknesses, as well as methodological frontiers and future directions. With respect to the purposes of this paper random-effects and random-parameters regression models are here of special interest. This is to account for possible temporal correlations in accident data collected on the same tunnels over successive time periods or to capture the possible unobserved heterogeneity across accident observations both in absence than in presence of random-effects.

Random-effects models become necessary when spatial correlations (data collected from the same geographic region) or temporal correlations (data collected on the same observational unit over successive time periods) are suspected to be present. It is to be said that random-effects models are more especially appropriate for capturing respectively spatial correlations or temporal correlations, but not a combination of the two. Regression models to account for temporal correlation were proposed by Hausman et al. (1984) (the Random Effects Negative Binomial model, RENB) and by Guo, (1996) (the Negative Multinomial model, NM). Analyses of count data based on RENB model and/or NM model were considered in Shankar et al. (1998), Shankar and Ulfarsson (2003), Chin and Quddus (2003), Hauer (2004), Caliendo et al. (2007), Caliendo et al. (2013), Caliendo and Guida (2014). Further studies concerning temporal and spatial correlation models can be found also in Wang and Abdel-Aty (2006).

Random-parameters models have been recently considered in Milton et al. (2008), Anastasopoulos and Mannering (2009), El-Basyouny and Sayed (2009), Washington et al. (2011), Garnowski and Manner (2011), Anastasopoulos et al. (2012b), Venkataraman et al. (2013), Chen and Tarko (2014). These models attempt to account for possible unobserved heterogeneity (unobserved factors that may vary across accident observations) by allowing the regression model parameters to be random-parameters, in contrast with the assumption that parameters are considered to be constant across observations. In this respect, Greene (2007) has developed estimation procedures for incorporating random parameters in count-data models which are within the scope of this paper. Additionally, it is to be reminded that a special case in the class of random-parameters models is the so called random-intercept Negative Binomial model (RINB). In this model only the regression intercept is random, and this model is known in the literature to be equivalent to the random-effects Negative Binomial model (e.g., see Hilbe (2007) and Greene (2007) for greater in-depth knowledge).

In terms of methodological approaches, recent research has also shown the potential of the mixed logit model. This model (also called random-parameters logit model) may be used for better understanding the injury-severity distribution and might account for the unobserved heterogeneity that is to be probably present in using limited data as opposed to detailed crash-specific data. Applications of the mixed- logit model, the use of which is not within the scope of this paper, can be found, for example, in Milton et al. (2008), Anastasopoulos and Mannering (2011), Manner and Wunsh-Ziegler (2013), Ye and Lord (2014).

Crash models are usually exploited in order to investigate the effects that various variables may have on the value of a pre-selected crash indicator, assumed as dependent variable. In this respect, the most common crash indicators that have been hitherto used in road accident analyses are the number of crashes per year (crash frequency), and the number of crashes per 100-million vehicle-kilometres (crash rate). Independent variables that affect crashes are more especially related to traffic flow, section length, road geometric characteristics, and weather. In particular, the fact that the number of accidents might not be a linear function of traffic flow and road section length, generally induces one to use crash frequency instead of crash rate as the more appropriate dependent variable. However, the use of accident rate has considerable appeal because this indicator is widely used in aggregate-level accident reports, and provides a standardized measure of the relative safety of road segments that is more easily interpreted than crash frequency.

Accident frequencies have been widely investigated over years using a variety of statistical approaches, and more especially the Negative Binomial regression model with fixed parameters (NB). However, constraining the parameters to be constant, when they may possibly vary across observations, might lead to inconsistent and biased parameter estimates (Washington et al., 2011). Given the potential heterogeneity in accident-frequency data, a random-parameters Negative Binomial (RPNB) model might be more appropriate in some cases.

As an alternative to the aforementioned models based on accident frequencies, regression models based on accident rates might be taken into account. From the perspective of statistical modeling, crash rate is a continuous random variable, in contrast to crash frequency which is a discrete random variable. However, because accidents on specific road segments are assessed over necessarily limited time periods, the likelihood exists that many road segments will have no accidents reported during the observation period, which is an event not in keeping with the assumption of a continuous random variable. In this respect, an interesting model appears to be the tobit regression model (Tobin , 1958) that considers the dependent variable as a continuous random variable, left-censored at zero. This model has been used more especially in economics (Amemiya, 1973; Nelson, 1977; McDonal and Moffitt, 1980) or social sciences (Fair, 1978; Roncek, 1992; Smith and Brame, 2003). In the transportation area, and in particular with reference to road accident analysis, tobit model is not commonly used, and only few papers can be found in the literature. For example, Anastasopoulos et al. (2008) considered a tobit model with fixed-parameters to predict vehicle accident rates as a function of pavement condition, road geometry, and traffic characteristics. More recently Anastasopoulos et al. (2012a) presented a multivariate tobit model of road accident-injury-severity rates; and Anastasopoulos et al. (2012b) developed a random-parameters tobit model to account for unobserved heterogeneity in motor-vehicle accident rates. While these studies are exploratory in nature, they suggest that the use of tobit regression models needs to be supported by more studies. Therefore, another intent of this paper is to further analyze the potential of tobit regression model with random-parameters.

Considerable research has been carried out on accidents occurred on open roads, while accidents in road tunnels have been rarely investigated. However, if a motor vehicle accident occurs, the crash severity may be significantly higher in tunnels than on the corresponding open roads. In this respect, it is to be mentioned that, according to a monitoring study regarding Italian motorway tunnels (Caliendo and De Guglielmo, 2012), it has been found an average rate of 12 severe accidents/10^8 veh-km in road tunnels in contrast to 9 severe accident/10^8 veh-km on the corresponding motorways. Therefore, there are reasons to investigate motor-vehicle accidents in road tunnels in a deeper detail.

From the view point of statistical modeling, Caliendo et al. (2013) used the Bivariate Negative Binomial regression model, jointly applied to non-severe crashes (accidents involving material damage only) and severe crashes (fatal and injury accidents only) to analyse the frequency of accident occurrence in Italian motorway tunnels with unidirectional traffic only. Furthermore, for understanding the trend of severe crashes over a 4-years period the random-effects Negative Binomial (RENB) regression model and the Negative Multinomial model were also used. It is to be said that Caliendo and Guida (2014) have recently developed also a new Bivariate Regression model for the simultaneous analysis of severe and total crashes occurrence. However, in Caliendo et al. (2013) as well as in Caliendo and Guida (2014) heterogeneity across observations was not investigated in a great depth. Moreover, also accident rates were not examined in detail.

In the light of the above considerations, there are at least two main reasons which motivated the present paper. The first one is the need to develop some random-parameters models in analysing accident frequency both in absence and in presence of random-effects, which may contribute to gaining new insights into the way the covariates influence accident frequency. The second one is for studying the potential of the random-parameters tobit regression model, which is based on accident rate, and making a possible comparison with the random-parameters Negative Binomial model based on accident frequency. This is for understanding more especially if results based on accident rates are possibly in contrast with those based on accident frequencies.

For these purposes, on the basis of the aforementioned data-base concerning the number of severe crashes occurred in Italian motorway tunnels in a 4-years time period, a random-parameters Negative Binomial (RPNB) model is developed. Moreover, in order to account for temporal correlation in accident data, expressed in terms of severe crash frequency per year, another random-parameters Negative Binomial model in which both the intercept and regression coefficients are allowed to vary randomly (RPINB) is developed. It was found that the "inverse dispersion" parameter (φ) of the RPINB model diverged to infinity (i.e., the over-dispersion parameter α converged to zero), thus indicating that the outcomes of the RPINB model converged to those of the corresponding Poisson model. For this reason the Poisson model where both the intercept and regression parameters are allowed to vary randomly (RPIP) is commented on this paper. Finally, an analysis of severe-crashes rate based on the random-parameters tobit regression model is also proposed. Then, some differences emerged between the analysis based on crash rates and the ones based on crash frequencies are commented on.

2. Data description

The database consists of 765 severe accidents which occurred within 260 Italian motorway tunnels (232 two-lane and 28 three-lane) with unidirectional traffic only over a 4-years period (2006-2009). In two-lane tunnels 670 severe accidents were observed, while in three-lane tunnels 95 severe accidents were counted. More especially accident count data for year 2006, 2007, 2008, and 2009 were, respectively, 247, 181, 122, 120 for two-lane tunnels; and 32, 33, 16, 14 for three-lane tunnels.

Annual average daily traffic (AADT) values ranging from 4500 to 40,760 vehicles per day were found for two-lane tunnels and between 5030 and 32,260 vehicles per day for three-lane tunnels. The percentage of trucks was 14-31% and 17-23% for two- and three-lane tunnels, respectively. Two-lane tunnels have lengths ranging from 387 m to 3254 m, while three-lane tunnel have lengths between 524 m and 4725 m. The presence of sidewalk was recorded in 147 tunnels. More detailed information on accident counts, summary statistics of tunnel length, AADT, and percentage of trucks, are reported in Caliendo et al. (2013) who used the same database of severe crashes.

Summary statistics of tunnel length, AADT and percentage of trucks are given in Table1.

Table 1. Summary statistics of independent variables

Variable		Mean	Mode	Standard deviation	Minimum	Maximum
Length (km)	Two-lane tunnels	1.188	0.529	0.637	0.387	3.254
	Three-lane tunnels	0.957	0.632	0.791	0.524	4.725
AADT /10,000 (veh/day)	Two-lane tunnels	1.512	0.613	0.837	0.450	4.076
	Three-lane tunnels	2.179	2.364	0.665	0.503	3.226
Percentage of trucks (%)	Two-lane tunnels	21.5	16	4	14	31
	Three-lane tunnels	22.1	23	2	17	23

It is to be said that 63 out of 260 tunnels had no severe accidents reported during the analysis period.

3. Functional form for traffic volume

A preliminary exploratory data analysis was carried out in order to assign a suitable functional form to *AADT* per lane ($AADT_L$), which represents the most influential variable in predicting the number of accidents. To this aim (see Caliendo et al., 2013, or Caliendo and Guida, 2014) we used the approach proposed by Hauer and Bamfo (1997). In particular by following the Integrate-Differentiate method in its generalization to more than one explanatory variable, the Empirical Integral Function (EIF) for the traffic volume per lane was derived in the case of severe accidents. Since this function showed to have two inflection points: one at an

$AADT_L$ of about 5000 vehicles per day and another one at an $AADT_L$ of about 13,000 vehicles per day, simple forms such as x^{β} and $e^{\beta x}$ appeared to be inadequate for describing the dependence of the expected number of severe accidents on traffic volume over the entire range of the observed $AADT_L$ values in the present situation. In fact, it was found that the relative error of prediction when using the power model, although smaller than that for the exponential model, was still high in the regions ($AADT_L$ < 5000) and ($AADT_L$ > 13,000), with underestimation in the former one and overestimation in the latter one. Thus, in order to get a better fit to tunnel data, also in the present paper we propose the following functional form for the $AADT_L$ variable: $g(AADT_L) = AADT_L^{\beta_1}\ e^{\beta_2 D_1}\ e^{\beta_3 D_2}$ where D_1 is a dummy variable which is 1 when the $AADT_L$ is fewer than 5000 vehicles per day and 0 otherwise, while D_2 is a dummy variable which is 1 when the $AADT_L$ is greater than 13,000 vehicles per day and 0 otherwise. When using this model, the predicted number of accidents computed in the said intervals of traffic volume practically coincides with the observed number of accidents. Moreover, it was also found that, under this model, the CURE (Cumulative Residuals) line is entirely contained within two standard deviations ($\pm 2\sigma^{*}$); which according to Hauer and Bamfo (1997) indicates that the functional form assumed for the $AADT_L$ variable fits the data well along the entire range of values assumed by this variable. Thereby, in the current paper, which uses the same dataset of severe accidents relative to 260 Italian motorway tunnels, random-parameters Negative Binomial models are carried out on the basis of the aforementioned modified power function.

One possible explanation of the trend of crash frequency versus $AADT_L$ might be attributable to the fact that free-flowing conditions ($AADT_L$ < 5,000 vehicles/day) or almost congested flow-traffic situations ($AADT_L$ > 13,000 vehicles/day) might be found. In particular when the traffic volume is low, as $AADT_L$ increases, the frequency of lane changing and overtaking movements increase so that more traffic conflicts and consequently more crashes are expected. In almost congested conditions, as $AADT_L$ increases, the drivers' freedom of manoeuvre (changing lanes and/or overtaking) begins to become ever more limited, and consequently a minor increase of crashes may be expected.

4. Set of independent variables

The candidate set of independent variables is: tunnel length (L), annual average daily traffic per lane ($AADT_L$), percentage of trucks ($\%T_r$), number of lanes (N_L), presence of sidewalk (SW). In addition the two dummy variables D_1 (1 if $AADT_L \leq 5000$ vehicles per day; 0 otherwise) and D_2 (1 if $AADT_L \geq 13{,}000$ vehicles per day; 0 otherwise) are also considered.

Moreover, when one wants to capture also the trend of severe crashes over time during the 4-years monitoring period, the dummy variables year 2007, year 2008, and year 2009 are also introduced, by assuming 2006 as reference year.

The dependent variable is crash frequency (number of severe crashes per year) in the case of Negative Binomial model and crash rate (number of severe crashes per 100-milion vehicle-kilometers) in the case of tobit regression model. It is to be reminded that, in the latter case, tunnel length (L) and $AADT_L$ are to be excluded from the set of independent variables, so that for the tobit model the set of potential independent variables is: $\%T_r$, N_L, SW, D_1, and D_2.

It is to be stressed that the scope of the paper is to investigate crashes occurred in road tunnels. In this respect, some variables such as those concerning weather, lighting, and junctions or intersections are in general not present in contrast with crashes occurred on open roads. In other terms, often a lower number of variables are available for investigating crashes occurred in tunnels compared to those of open roads.

5. Data analysis based on Negative Binomial regression models

5.1 Methodology

A very basic way for describing the fluctuation of accident counts, say Y_i, occurring on a road section *i* during given time intervals, is to assume that Y_i is a Poisson random variable, which implies that $E(Y_i) = \lambda_i$ and $Var(Y_i) = \lambda_i$. In many situations, however, accident counts appear to be "over-dispersed", i.e. $Var(Y_i) > E(Y_i)$. Hence, the Negative Binomial (NB) model

$$f(y_i) = \frac{\Gamma(y_i + \varphi)}{y_i!\Gamma(\varphi)} \left[\frac{\lambda_i}{\lambda_i + \varphi} \right]^{y_i} \left[\frac{\varphi}{\lambda_i + \varphi} \right]^{\varphi} \tag{1}$$

which has $E(Y_i) = \lambda_i$ and $Var(Y_i) = \lambda_i \left(1 + \lambda_i / \varphi\right)$, is often used, thus allowing for the variance of accident counts to be greater than the mean, provided that $1/\varphi > 0$. Then, in order to take into account the effect of explanatory variables, a regression model of the expected number of accidents is defined by typically using the log-linear function $\lambda_i = \exp(\mathbf{x}_i^T \boldsymbol{\beta})$, where $\boldsymbol{\beta}$ is a vector of fixed, even if unknown, coefficients. In presence of unobserved factors which may vary across road sections, however, a better approach is considering the $\boldsymbol{\beta}$ parameters as random variables, by assuming

$$\boldsymbol{\beta}_i = \boldsymbol{\beta} + \boldsymbol{\delta}_i$$

where $\boldsymbol{\delta}_i$ is a random variable with some probability density function $f(\delta)$, for example a normal variable with zero mean and constant standard deviation.

Let $L(Y_i \mid X_i, \delta)$ be the likelihood function for tunnel *i* $(i = 1, \ldots, N)$, given δ, and the vector of independent variables X_i. The marginal likelihood of all observed data is given by

$$L = \prod_{i=1}^{N} \int_{\delta} L(Y_i \mid X_i, \delta) \, f(\delta) \mathrm{d}\delta \tag{2}$$

from which unknown parameters can be estimated by a maximization procedure. It is to be noted, however, that conventional procedures based on numerical maximization of eq. (2) are, in most cases, unfeasible when the number of unknown parameters is not very small. Thus, Green (2007) has developed an efficient simulation procedure, based on the Halton draws (see Halton (1960) for greater depth of knowledge), which allows to estimate the model parameters even in presence of a high number of random parameters.

5.2. Estimation results of the random-parameters Negative Binomial (RPNB) model

The data set consists of the number of severe accidents registered in $n = 4$ years from 2006 to 2009 in $N = 260$ tunnels. The explanatory variables are: L, $AADT_L$, $\%Tr$, NL, SW, D_1, and D_2.

Table 2 presents the estimation results of the random-parameters Negative Binomial (RPNB) model based on the number of severe crashes registered in the 4 years period (crash frequency can be easily computed dividing the expected number of crashes by 4 years). The random-parameters model was estimated by using 2000 Halton draws. Since some of the potential covariates could actually have very little or even no effect on accident counts, a procedure based on the Likelihood Ratio Test (LRT) was used in order to decide which subset of the full set of potentially explanatory variables should be included in the regression model (the likelihood-ratio statistic $\Lambda = 2\left[l(\hat{\boldsymbol{\beta}}, \hat{\varphi}) - l(\hat{\boldsymbol{\beta}}', \beta_k = 0, \hat{\varphi}) \right]$ where $l(\hat{\boldsymbol{\beta}}, \hat{\varphi})$ is the log-likelihood of the regression model containing all the covariates, and $l(\hat{\boldsymbol{\beta}}', \beta_k = 0, \hat{\varphi})$ is the log-likelihood of the regression model with the k-th covariate out, is asymptotically distributed as a chi-squared distribution with 1 degree of freedom). In this respect, all variables were found to be significant at a significance level much less than 0.05 (this reflects at least a 95% confidence), except the percentage of trucks ($\%Tr$) and the presence of sidewalk (SW) that do not appear to be statistically significant in the RPNB model. Among the statistically significant variables, the coefficients that were found to be random are associated to the variables L and D_1. The estimated "inverse dispersion" parameter φ ($\varphi = 1/\alpha$) is large, thus possibly indicating that part of unexplained heterogeneity is captured by random parameters. However, it is to be said that in an exploratory analysis we have found that the random-parameters Negative Binomial (RPNB) model is not statistically superior to the fixed-parameters NB model in the case of tunnels investigated. One possible explanation of this finding, which appears to be in contrast with the results of studies based on accidents occurred on open roads (where in general a major difference between random-parameters and fixed-parameters models is found), might be that tunnels do not have maybe as much unobserved heterogeneity as open roads. In fact, tunnels are closed structures with confined space where some factors that can cause accidents and/or contribute to unobserved heterogeneity are indeed to be controlled. For example there is absence of rain, snow, ice, wind, and fog; except that in the entrance and exit zones of tunnels. No difference exists between night and day because lighting conditions are constant within tunnels. Moreover hazardous points as junctions or intersections in general are not present in tunnels. However, this needs to be better investigated by means of other studies also in tunnels different from Italian ones.

In the present paper we discuss only the random-parameters Negative Binomial (RPNB) model in order to focus more especially on the factors that influence motor-vehicle accidents occurred in road tunnels and save space. Additionally, it is to be said that the interpretation of regression coefficients that are found to be random contributes, however, to a deeper understanding of the way in which the corresponding covariates influence crashes occurred in road tunnels.

The number of severe accidents appears to be positively associated with tunnel length (L), average daily traffic per lane ($AADT_L$), number of lanes (NL), and percentage of trucks ($\%Tr$). Moreover, a multiplicative coefficient greater than 1 ($e^{\beta_2 D_1}$) is associated to a small traffic volume ($AADT_L$ < 5000 vehicles per day), whilst a multiplicative coefficient lower than 1 ($e^{\beta_3 D_2}$) is associated to a high traffic volume ($AADT_L$ > 13,000 vehicles per day); (for more in-depth knowledge about the influence of different $AADT_L$ intervals on crashes see Caliendo et al. (2013) and/or Caliendo and Guida (2014)). However, tunnel length resulted in a random parameter that is normally distributed, with mean 0.484 and standard deviation 0.588, so that 79.5 percent of its distribution is greater than 0 and 20.5 percent less than 0. This suggests that a large majority of road tunnels result in an increase in accident occurrences as length increases, but a few ones result in a decrease. Similarly, the indicator variable D_1 resulted in a random parameter that is normally distributed, with mean 0.353 and standard deviation 0.673, so that 70.0 percent of its distribution is greater than 0 and 30.0 percent less than 0. This suggests that a majority of road tunnels result in an increase in severe accident occurrences when traffic volume is small, but a few ones result in a decrease.

We observe that in the present case study the continuous variables enter the model in logarithmic form. In such a case, the estimated coefficient can also be interpreted as the percent increase in the expected number of accidents corresponding to a 1 percent increase in the variable. Thus, from Table 2, it can be readily seen that the variable that has the strongest effect on the number of accidents is $AADT_L$. In fact, a 1 percent increase in annual average daily traffic per lane on average leads to a 2.01 percent increase in the expected number of accidents in the 4-year period. Similarly, a 1 percent increase in tunnel length on average leads to a 0.484 percent increase in the expected number of accidents; and a 1 percent increase in the percentage of trucks on average leads to a 0.436 percent increase in the expected number of accidents.

Table 2. Estimation results of random-parameters Negative Binomial (RPNB) model based on the 4-year number of severe crashes

Variables	RPNB		
	Point Estimate	Standard Error	LRT statistic
Constant	2.09599	0.46310	
*Log of length (km)	0.48430	0.11320	15.38
Standard deviation of parameter distribution: 0.58760			
Standard error of standard deviation:0.09521			
Log of AADT per lane /10,000	2.01479	0.20901	81.48
* D1 (1 if AADT per lane <5000; 0 otherwise)	0.35333	0.25119	7.72
Standard deviation of parameter distribution: 0.67320			
Standard error of standard deviation:0.13258			
D2 (1 if AADT per lane >13,000; 0 otherwise)	-0.50690	0.22040	7.23
Log of percentage of trucks/100	0.43569	0.27822	1.64
Number of lanes (0 if two lanes, 1 if three lanes)	0.42335	0.15613	5.23
*Sidewalk (1 if present, 0 if absent)	-0.04211	0.13203	0.30
Standard deviation of parameter distribution: 0.19418			
Standard error of standard deviation:0.06890			
Overdispersion parameter ($1/\alpha$)	13.1519	5.69355	
Number of observations		260	
Log likelihood function LL(β)		-484.34800	
Log likelihood with constant only LL(0)		-580.20329	
* random parameters			

Therefore, the results obtained confirm that longer tunnels are (in large majority) associated with a greater number of severe accidents, possibly due to the drivers' diminishing concentration with increasing length (which reflects an increased exposure). However, it is to be stressed that the coefficient associated to the tunnel length is less than 1, so as the relationship between the number of crashes and tunnel length is nonlinear with a less than proportional increase. Moreover, in free-flowing conditions (small traffic volumes), the frequency of lane changing and overtaking movements increases so that more severe accidents are expected. Likewise as the percentage of trucks increases more severe accidents are expected to occur. Finally, severe accidents are also expected to increase in correspondence of an increase in the number of lanes, which is attributable to the greater opportunities for lane change and consequently to more traffic conflicts.

In order to investigate on the predictive accuracy of the random parameters Negative Binomial (RPNB) model proposed in this paper, we have numerically estimated the total number of crashes for the 260 tunnels investigated obtaining an estimate of 768 severe crashes over a 4-year period (2006-2009) against 765 severe crashes observed in field. Thus, this model seems to fit well the crashes observed in the tunnels investigated.

5.3. Estimation results of the random-parameters model with random-effects

This analysis is based on the number of severe crashes observed in each tunnel in each year from 2006 to 2009. It is to be said that the number of severe crashes occurring in the same tunnel in different years cannot be considered stochastically independent observations, since for example they share at least the effect of the geometrical traits (not considered in this study) of that road tunnel. Thus, it is reasonable to assume that they form a *cluster* of observations sharing common, even if unknown, *random effects*. Random-effects models analyses data *by clusters*, thus ensuring that the presence of correlation among observations pertaining to the same road tunnel in different years is taken into account. It is understood that random-effects are the same within each panel of cluster, but they differ across clusters. However, the regression parameters are the same within and across clusters. In other words, random-effects models do not consider that the regression parameters, which can be fixed within a cluster (same tunnel), can vary across clusters. In this respect, random-parameters models represent a big advantage because they can account for this information (e.g., in LIMDEP statistical package (Greene, 2007) a panel-data vector (PDS) is included for taking into account that the estimated random parameters can be fixed within clusters and vary between them).

In the light of the above considerations, in order to account for both temporal correlations and unobserved heterogeneity in crash data the Negative Binomial model where both the intercept and regression parameters are allowed to vary randomly (RPINB) was considered. The dummy variables year 2007, year 2008, and year 2009 (assuming as reference year 2006) were added to capture the year effect on severe accidents counts.

In an exploratory analysis, however, we found that the "inverse dispersion" parameter φ (φ = 1/α) of the aforementioned RPINB model diverged to infinity, thus indicating the convergence of this model to the corresponding Poisson model. This suggested that, at least in the present case study, the Poisson model where both the intercept and parameters are allowed to vary randomly (RPIP) could be used as an alternative to the RPINB model. Moreover, we also found that the RPIP model is not statistically superior to the random-intercept Poisson (RIP) model (i.e., to the random-effects Poisson model). In the present paper, however, we discuss only the random-parameters Poisson (RPIP) model.

Table 3 shows that all variables are significant at a level much less than 0.05, except the percentage of trucks (*%Tr*) and sidewalk (*SW*) that do not appear to be statistically significant. Note that, in addition to the regression intercept assumed a priori random, also the parameters associated to *L*, D_1, and *SW* were found to be random. The *L* resulted in a random parameter normally distributed, with mean 0.482 and standard deviation 0.547, so that 81.1 percent of its distribution is greater than 0 and 18.9 percent less than 0. The indicator variable D_1 resulted in a random parameter that is normally distributed, with mean 0.457 and standard deviation 0.547, so that 79.8 percent of its distribution is greater than 0 and 20.2 percent less than 0. This analysis confirms the presence of a significant year effect, consisting in a systematic reduction in severe crashes over time in Italian motorway tunnels. This reduction of the severe crash occurrence might be attributable to: implementation and/or reinforcement of some facilities in tunnels, increasing installation on the Italian motorways of an electronic system for checking the average speed of vehicles, introduction of the driving licence with a demerit point system (see Caliendo et al. (2013) for a greater in-depth knowledge about the year effect; where we also showed that the random-effects Negative Binomial (RENB) model is statistically superior to fixed-effects one).

Table 3. Estimation results where both the intercept and parameters are allowed to vary randomly in the Poisson (RPIP) model

Variables	RPIP		
	Point Estimate	Standard Error	LRT statistic
* Constant	0.94672	0.39353	
Standard deviation of parameter distribution: 0.28138			
Standard error of standard deviation: 0.03570			
* Log of length (km)	0.48166	0.09660	15.69
Standard deviation of parameter distribution: 0.54744			
Standard error of standard deviation: 0.08069			
Log of AADT per lane /10,000	2.03744	0.19233	89.42
* D1 (1 if AADT per lane <5000; 0 otherwise)	0.45704	0.23255	7.29
Standard deviation of parameter distribution: 0.54736			
Standard error of standard deviation: 0.11678			
D2 (1 if AADT per lane >13,000; 0 otherwise)	-0.52615	0.15449	9.93
Log of percentage of trucks/100	0.36440	0.23315	1.28
Number of lanes (0 if two lanes, 1 if three lanes)	0.41827	0.13685	5.36
* Sidewalk (1 if present, 0 if absent)	-0.06276	0.11646	0.21
Standard deviation of parameter distribution: 0.24082			
Standard error of standard deviation: 0..06129			
year 2007	-0.26570	0.10960	8.50
year 2008	-0.70650	0.13895	48.91
year 2009	-0.71325	0.12517	46.46
Number of observations	1040		
Log likelihood function LL(β)	-997.28054		
Log likelihood function LL(β)[(1)]	-997.21528[(1)]		
Overdispersion parameter ($1/\alpha$)[(1)]	8.341E+06[(1)]		
Log likelihood with constant only LL(0)	-1132.23540		

* random parameters

[(1)] Binomial Negative model having both the intercept and parameters that are allowed to vary randomly (RPINB)

It is to be said that a reason why the RPIP and RIP models come to the same results might be that tunnels, being closed structures with controlled conditions, might not have as much unobserved heterogeneity as open roads. It is worth mentioning that, in a recent paper on work zones data, also Chen and Tarko (2014) found that the random-parameters Negative Binomial model gives quite similar results to the corresponding random-intercept model.

6. Tobit regression models

6.1. Methodology

The accident rate is to be considered as a continuous random variable (rv), in contrast to accident count which is a discrete one. Since, however, the time of observation is finite, a cluster in zero is often observed for the accident rate, due to road segments with no accidents. In this respect, a suitable candidate to address the abovementioned situation may be the tobit model (Tobin, 1958), conceived for the case where there is a latent variable Y_i^* which is observable only when above a given threshold y_L (in particular, $y_L = 0$).

Specifically, the tobit model for the latent variable Y_i^* is formulated as

$$Y_i^* = \beta' X_i + \varepsilon_i \tag{3}$$

thus assuming that Y_i^* linearly depends on the independent vector of variables X_i via the vector of parameters β. In addition, there is a normally distributed error term ε_i, with zero mean and constant variance σ^2 across *i*. Then, the observable variable Y_i, *i.e.* the accident rate in the present application, is defined as

$$\begin{cases} Y_i = Y_i^* & \text{if } Y_i^* > 0 \\ Y_i = 0 & \text{if } Y_i^* \le 0 \end{cases} \tag{4}$$

The likelihood function for tobit model over *N* road sections is

$$L = \prod_{i=1}^{N} \left[1 - \Phi(\beta' X_i / \sigma)\right]^{1-I(Y_i)} \left[\phi\left((Y_i - \beta' X_i)/\sigma\right)\right]^{I(Y_i)} \quad (5)$$

where $\Phi(\cdot)$ is the standard normal distribution function, $\phi(\cdot)$ is the standard normal density function, σ is the standard deviation of the error term, and $I(\cdot)$ is the indicator function which takes value 0 if $Y_i = 0$ and 1 otherwise.

Note that the expected value of the observable variable Y_i for all cases is

$$E(Y_i \mid X_i) = \beta' X_i \Phi(z) + \sigma\phi(z) \quad (6)$$

where $z = \beta' X_i / \sigma$, and $\Phi(z)$ represents the probability of Y_i being above zero. Instead, the expected value for cases above zero only is given by

$$E(Y_i \mid X_i; Y_i > 0) = \beta' X_i + \sigma \frac{\phi(z)}{\Phi(z)} \quad (7)$$

In order to evaluate the effect of an independent variable on the expected value, the first-order partial derivatives of eq. (6) or (7) with respect to X_k can be used. In particular, McDonald and Maffin (1980) found that

$$\frac{\partial E(Y_i \mid X_i)}{\partial X_k} = \Phi(z) \frac{\partial E(Y_i \mid X_i, Y_i > 0)}{\partial X_k} + E(Y_i \mid X_i, Y_i > 0) \frac{\partial \Phi(z)}{\partial X_k} \quad (8)$$

where

$$\frac{\partial \Phi(z)}{\partial X_k} = \frac{\beta_k \phi(z)}{\sigma}$$

and, from eq. (7),

$$\frac{\partial E(Y_i \mid X_i, Y_i > 0)}{\partial X_k} = \beta_k \left[1 - z\frac{\phi(z)}{\Phi(z)} - \left(\frac{\phi(z)}{\Phi(z)}\right)^2\right] \quad (9)$$

In the model discussed above, parameters β_k are assumed to be fixed (even if unknown) constants. However, when heterogeneity due to unobserved factors is possibly present among road sections, a better approach is considering the β_k parameters as random variables, by assuming

$$\beta_{ki} = \beta_k + \delta_{ki} \qquad k = 1,...,p \qquad i = 1,...,N$$

where δ_{ki} is a random variable with some probability density function $f(\delta_k)$, for example a normal variable with zero mean and constant standard deviation σ_k across *i*. As such, the marginal likelihood is given by

$$L = \prod_{i=1}^{N} \int_{\delta} L(Y_i \mid X_i, \delta) f(\delta) \mathrm{d}\delta \qquad (10)$$

from which parameters β_k and σ_k $k = 1,...,p$ can be estimated by a maximization procedure. For this case too, Green (2007) has developed a procedure which allows to estimate the model parameters even when the number of random parameters is large.

6.2. Estimation results of random-parameters tobit model

Table 4 presents the estimation results of the analysis based on severe crash rate, when using the tobit model with random-parameters.

For this model severe accident rate is negatively associated with D_1, D_2, and *SW*; while it is positively associated with percentage of trucks ($\%T_r$) and number of lane (*NL*). Only the parameter associated to the independent variable D_1 resulted to vary randomly. Moreover, the number of lanes (*NL*) and the indicator variable of a high traffic level per lane (D_2) are not statistically significant variables, whereas percentage of trucks ($\%T_r$) and the presence of sidewalk (*SW*) are here statistically significant.

Table 4. Estimation results of random-parameters tobit model based on severe accident rates per lane

Variables	Random-parameter tobit		
	Point Estimate	Standard Error	LRT statistic
Constant	89.40460	21.49088	
* D1 (1 if AADT per lane <5000; 0 otherwise)	-26.67970	5.21823	28.38
Standard deviation of parameter distribution: 34.8216			
Standard error of standard deviation: 3.34164			
D2 (1 if AADT per lane >13,000; 0 otherwise)	-1.54928	15.18658	0.04
Log of percentage of trucks/100	34.31410	13.38183	6.70
Number of lanes (0 if two lanes, 1 if three lanes)	3.06836	6.38422	0.19
Sidewalk (1 if present, 0 if absent)	-19.96890	4.76867	14.43
σ	28.40140	1.09930	
Number of observation (N)	260		
Log likelihood function LL(β)	-1003.43897		
Log likelihood with constant only LL(0)	-1030.19755		
* random parameter			

Note that the above results are in contrast with the aforementioned results based on crash counts. This seems to indicate that modeling accident occurrence by using crash rate instead of crash counts may result in very different significance of the regression covariates. One possible explanation is that, when using crash rate, two very important factors which significantly influence crash occurrence such as length (L) and annual average daily traffic per lane ($AADT_L$) no longer act as independent variables in the regression model, but they directly act on the value of the dependent variable, i.e. the crash rate, in a *linear* way. However, since the dependence of crash occurrence on L and $AADT_L$ is usually far from being linear (in particular, in the present case study the crash occurrence multiplicatively depends on the square root of L and on the square of $AADT_L$), this reveals that the significance of the remaining variables on crash rate could not be correct in the tobit model.

7. Summary and Conclusions

The paper explores the use of the random-parameters Negative Binomial model (RPNB) to account for heterogeneity across accident observations by allowing model parameters to randomly vary in contrast with the fixed-parameters Negative Binomial (NB) model.

Using 4 years of severe crash data from Italian motorway tunnels with unidirectional traffic only, and assuming crash counts (number of severe crashes in the 4-year period) as dependent variable, the following results may be drawn. Severe crashes are positively associated with: tunnels length (*L*), annual average daily traffic per lane ($AADT_L$), and number of lanes (*NL*). In contrast the presence of the sidewalk (*SW*) and the percentage of trucks *(%Tr*) are not found statistically significant factors. It was found that the coefficients associated to the significant variables *L* and D_1 should be considered as random variables. The interpretation of random coefficients has contributed to a fuller understanding of the way in which the corresponding covariates influence crash frequency.

The RPNB model was found to be statistically quite similar to NB one, which appears to differ form results relative to accidents occurred on open roads. One possible explanation is that, being tunnels closed structures with controlled conditions, they might not have as much heterogeneity as open roads. This, however, need to be better investigated by means of other studies in tunnels different from Italian ones.

To account for temporal correlations in crash data, the Negative Binomial model where both the intercept and parameters are allowed to vary randomly (RPINB) was considered. Since, however, the "inverse dispersion" parameter φ of the RPINB model was found to diverge to infinity, the Poisson model where both the regression intercept and parameters are allowed to randomly vary (RPIP) was adopted. It was found that the coefficients associated to the variables *L* and D_1 should be considered as random variables, which added more information about the way in which the corresponding covariates influence crash frequency.

This model confirmed the presence of a significant year effect, consisting in a systematic reduction in severe crashes over years. The RPIP model, however, was found to be statistically quite similar to the intercept-random Poisson (RIP) model.

The fact that, in the present case study, random-parameters models are found to be statistically quite similar to fixed-parameters ones might be attributable to the combination of different factors. In particular, in addition to the fact that tunnels might not have as much heterogeneity as open roads, this might have been caused also by the number of independent variables available for the analysis, the sample size, the number and type of parameters that are found to be random compared to fixed ones. In this respect, however, it is to be stressed that also Garnowki and Manner (2011) found that the use of a random-parameters model compared to a fixed-parameters one might lead to a small improvement in the model fit; moreover Chen and Tarko (2014) showed that the random-parameters negative binomial model with random-effects (both the intercept that parameters are allowed to vary randomly) gives results similar to those of the random-intercept model. However, this is not, in our opinion, a justification for “a priori” using an approach which cannot account for regression parameters varying randomly across observations. Therefore studies should be addressed towards unobserved heterogeneity to make additional developments possible in research. Future research should also be addressed to analyzing the effect of distributions of random-parameters different from the normal one (e.g., uniform, log-normal, Weibull, triangular, etc.) on the significance of the regression model.

The use of crash rate (number of crashes per 100-million vehicle-kilometres) instead of crash frequency as dependent variable in statistical modelling is also investigated, in order to understand more especially if a different significance of the covariates might be expected. In this respect, a random-parameters tobit regression model that use accident rate as dependent continuous variable left-censored at zero was developed. In contrast to the analysis based on accident counts, it resulted that number of lanes (*NL*) and high level of traffic per lane (D_2) are not statistically significant variables, whereas the percentage of trucks (%T_r) and the presence of sidewalk (*SW*) is statistically significant. In the light of this findings, it is to be stressed that modeling accident occurrence by using crash rate instead of crash counts may result in different significance of covariate in regression models. This appears to be attributable to fact that the variables *L* and $AADT_L$ are no more involved as independent variables in the regression model, but they act on the value of the dependent variable (i.e. the crash rate) in a *linear way*. Given that the dependence of crash occurrence on *L* and $AADT_L$ is usually not linear, this could lead to a misinterpretation of the influence of the remaining variables on crash rate. Therefore, it is suggested that some caution has to be used when using crash rates to detect significant variables.

It is to be said that in the current study we have modeled crash frequencies and rates without taking into account of their injury-severities. Therefore a possible direction of expansion of the present work could be an analysis based on the injury-severity distribution. Another possible extension of this paper in future studies should be testing the influence of trucks on crashes occurred in tunnels using dummy variables for truck percentage.

References

Amemiya, T., 1973. Regression analysis when the dependent variable is truncated normal. Econometrica Vol.41, No. 6, 997-1016.

Anastasopoulos P.C., Tarko, A. P., Mannering F.L., 2008. Tobit analysis of vehicle accident rates on interstate higways. Accident Analysis and Prevention, Vol. 40, 768-775.

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

Anastasopoulos P.C., Mannering F.L., 2011. An empirical assessment of fixed and random parameter logit modes using crash- and non-crash-specific injury data. Accident Analysis and Prevention, Vol. 43, 1140-1147.

Anastasopoulos P.C., Mannering F.L., Shankar V.N., Haddock J.E., 2012a. A multivariate tobit analysis of highway accident-injury-severity rates. Accident Analysis and Prevention, Vol. 45, 110-119.

Anastasopoulos P.C., Mannering F.L., Shankar V.N., Haddock J.E., 2012b. A study of factors affecting highway accident rates using the random-parameters tobit model. Accident Analysis and Prevention, Vol. 45, 628-633.

Caliendo C., Guida M., Parisi A., 2007. A crash-prediction model for multilane roads. Accident Analysis and Prevention, Vol. 39, 657-670.

Caliendo C., De Guglielmo, M. L., 2012. Accident rates in road tunnels and social costs evaluation. SIIV- 5th International Congress- Sustainability of Road Infrastructures. In: Procedia-Social and Behavioral Sciences, Vol. 53, 166-177. http://www.sciencedirect.com/science/article/pii/S1877042812043327

Caliendo C., De Guglielmo M.L., Guida M., 2013. A crash-prediction model for road tunnels. Accident Analysis and Prevention, Vol. 55, 107-115.

Caliendo, C., Guida, M. (2014). A new bivariate regression model for the simultaneous analysis of total and severe crashes occurrence. Journal of Transportation Safety & Security. Vol. 6, Issue 1,78-92.

Chen E., Tarko A.P., 2014. Modeling safety of highway work zones with random parameters and random effects models. Analytic Methods in Accident Research. Vol.1, 86-95.

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

El-Basyouny, K., Sayed, T., 2009. Accident prediction models with random corridor parameters. Accident Analysis and Prevention, Vol. 41, No 5, 1118–1123.

Fair, R.C., 1978. A theory of extramarital affairs. Journal of Political Economy, Vol. 86, No. 1, 45-61.

Garnowski, M., Manner, H., 2011. On factors related to car accidents on German Autobahn connectors. Accident Analysis and Prevention 43(5), 1864-1871

Greene, W., 2007. Limdep, Version 9.0. Econometric Software Inc., Plainview, NY.

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

Halton, J., 1960. On the efficiency of certain quasi-random sequences of points in evaluating multi-dimensional integrals. Numerische Mathematik 2, 84-90.

Hauer, E., & Bamfo, J. (1997). Two tools for finding what function links the dependent variable to the explanatory variables. *In Proceedings of the International Cooperation on Theories and Concepts in Traffic Conference.* Retrieved from: http://www.oocities.org/hauer@rogers.com/Pubs/ICTCT97TwoTools.pdf

Hauer, E., 2004. Statistical road safety modeling. In: 83rd TRB Annual Meeting, Washington,D.C., USA, January 11-15.

Hausman J., B.H. Hall, Z. Griliches, 1984. Econometric models for count data with an application to the patents-R&D relationship. Econometrica, Vol. 52, No. 4, 909-938.

Hilbe, J.M., 2007. Negative Binomial regression. Published in the USA, New York, by Cambridge University Press. www.cambridge.org

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

Manner, H., Wunsch-Ziegler, L., 2013. Analyzing the severity of accidents on the German

Autobahn. Accident Analysis and Prevention, Vol. 57, 40-48.

Mannering, F.L., Bhat C.R., 2014. Analytic methods in accident research: Methodological frontier and future directions. Analytic Methods in Accident Research. Vol.1, 1-22.

McDonald, J.F., Moffitt, R.A., 1980. The uses of tobit analysis. The Review of Economics and Statistics, Vol. 62, Issue 2, 318-321.

Milton, J., Shankar, V., Mannering, F., 2008. Highway accident severities and the mixed logit model: an exploratory empirical analysis. Accident Analysis and Prevention, Vol. 40, No 1, 260–266.

Nelson, F., 1977. Censored regression models with unobserved, stochastic censoring thresholds. Journal of Econometrics, Vol.6, 309-327.

Roncek, D.W., 1992. Learning more from tobit coefficients: extending a comparative analysis of political protest. American Social Review, Vol. 57, No. 4, 503-507.

Savolainen P.T., Mannering F.L., Lord D., Quddus M.A., 2011. The statistical analysis of highway crash-injury severities: A review and assessment of methodological alternatives. Accident Analysis and Prevention Vol. 43, 1666– 1676.

Shankar V.N., Albin R.B., Milton J.C., Mannering F.L., 1998. Evaluating median crossover likelihoods with clustered accident counts: an empirical inquiry using the random effects Negative Binomial model. Transportation Research Record 1635, 44-48.

Shankar V.N., Ulfarsson G.F., 2003. Accident Count Model Based on Multiyear Cross-Sectional Roadway Data with Serial Correlation. Transportation Research Record 1840, 193-197.

Smith, D.A., Brame, R., 2003. Tobit models in social science research: some limitations and a more general alternative. Soc. Methods Res., Vol. 31, No. 3, 364-388.

Tobin, J., 1958. Estimation of relationships for limited dependent variables. Econometrica, Vol. 26, No. 1, 24-36.

Venkataraman N., Ulfarsson G.F., Shankar V.N., 2013. Random parameters models of interstate crash frequencies by severity, number of vehicles involved, collision and location type. Accident Analysis and Prevention Vol.59, 309-318.

Wang, X., Abdel-Aty, M., 2006. Temporal and spatial analyses of rear-end crashes at signalized intersections. Accident Analysis and Preventio 38 (6), 1137-1150.

Washington S.P., Karlaftis M.G., Mannering F.L., 2011. Statistical and Econometric Methods of Transportation Data Analysis. Chapman & Hall/CRC, printed in USA.

Ye, F., Lord, D., 2014. Comparing three commonly used crash severity models on sample size requirements: multinomial logit, ordered probit, and mixed logit. Analytic Methods in Accident Research. Vol.1, 72-85.