Accident Analysis and Prevention 111 (2018) 147–154

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

Identifying traffic accident black spots with Poisson-Tweedie models

Birgit Debrabant[a], Ulrich Halekoh[a,*], Wagner Hugo Bonat[b], Dennis L. Hansen[c,d], Jacob Hjelmborg[a], Jens Lauritsen[c,d]

[a] Department of Public Health, University of Southern Denmark, Odense, Denmark
[b] Department of Statistics, Paraná Federal University, Curitiba, Brazil
[c] Accident Analysis Group, Department of Ortopedics, Odense University Hospital, Odense, Denmark
[d] Department of Clinical Medicine, University of Southern Denmark, Odense, Denmark

ARTICLE INFO

Keywords:
Black spot detection
Traffic accidents
Hospital admission data
Poisson–Tweedie distribution

ABSTRACT

This paper aims at the identification of black spots for traffic accidents, i.e. locations with accident counts beyond what is usual for similar locations, using spatially and temporally aggregated hospital records from Funen, Denmark. Specifically, we apply an autoregressive Poisson–Tweedie model, which covers a wide range of discrete distributions and handles zero-inflation as well as overdispersion. The estimated power parameter of the model was 1.6 ($SE = 0.06$) suggesting a distribution close to the Pólya-Aeppli distribution. We identified nine black spots consistently standing out in all six considered calendar years and calculated by simulations a probability of $p = 0.03$ for these to be chance findings. Altogether, our results recommend these sites for further investigation and suggest that our simple approach could play a role in future area based traffic accident prevention planning.

1. Introduction

We present a case study of black spot detection for traffic accidents, based on six years of hospital admissions data for traffic accidents on the island of Funen, Denmark. The main goal of black spot detection is to identify specific sites, e.g. intersections or road segments, as candidates for traffic safety improvements.

This is an active area of research, see e.g. Thomas and DeRobertis (2013), De Pauw et al. (2014), Vandenbulcke et al. (2014). The concern for traffic accident prevention stems from the fact that traffic accidents are estimated to be the eighth leading cause of death at the moment and are predicted to be the third leading cause of death by 2030 (WHO, 2013).

The data for the present study originated from records of all traffic-related injuries in the Funen region for the period 2002–2007, using hospital admissions data from all three hospitals. No study has yet been done in Denmark using this kind of data, and previous decisions regarding traffic safety improvements have been based on accident records by the police. Although hospital data do not contain those accidents where only material damage occurred, police records, on the other hand, tend to substantially under-represent vulnerable road users such as pedestrians and cyclists.

It has also been documented (Lauritsen et al., 2002) that for the region covered (Funen) more than 90% of treatment costs as well as societal costs after person injury is covered by those patients seeking treatment at the hospital. In general police records only cover 15–18% as seen since the mid-1980s after traffic accidents based on direct coupling at person level of police and hospital records (see www.ouh.dk/uag). This suggests that hospital records give a fuller picture of the health care-related consequences of traffic injuries.

The purpose of this article was to develop a simple yet sufficiently flexible statistical method suited to our dataset for the identification of black spots, the latter being locations with higher accident rates than expected given characteristics of the location and its neighbourhood.

A wide variety of statistical distributions and methods has been proposed for analysing traffic accident count data. Common distributions include Poisson and negative binomial distributions, Poisson-lognormal distributions as well as zero-inflated Poisson and negative binomial distributions, which have been adopted by, e.g. Jovanis and Li Chang (1986), Joshua and Garber (1990), Miaou and Lum (1993), Miaou (1994,1994), Maycock and Hall (1984), Turner and Nicholson (1998), Amoros et al. (2003), Cafiso et al. (2010), Miaou et al. (2005), Lord and Miranda-Moreno (2008), Aguero-Valverde and Jovanis (2008), Lord et al. (2005) and Lord et al. (2007). Random effects can be used to take correlations among observations as well as unobserved heterogeneity into account, see, e.g. Shankar et al. (1998), Miaou et al. (2003), El-Basyouny and Sayed (2009), Venkataraman et al. (2013) and Barua et al. (2015). Further modern modeling strategies, which have

[*] Corresponding author.
E-mail address: uhalekoh@health.sdu.dk (U. Halekoh).

https://doi.org/10.1016/j.aap.2017.11.021
Received 10 April 2017; Received in revised form 20 September 2017; Accepted 16 November 2017
Available online 01 December 2017

been applied to accident data, include latent-class (finite mixture) models (e.g. Park and Lord, 2009; Buddhavarapu et al., 2016; Heydari et al., 2017), Markov switching count models (e.g. Malyshkina and Mannering, 2009, 2010), hierarchical models (e.g. Jones and Jørgensen, 2003; Kim et al., 2007; Dupont et al., 2013), multivariate models (e.g. Miaou and Lord, 2003; Depaire et al., 2008; Dong et al., 2014; Heydari et al., 2017), Bayesian methods (e.g. Li et al., 2007; Elvik, 2008; Pei et al., 2011) and neural networks (Zeng et al., 2016). For a more complete overview of models used in accident research and application studies we refer the reader to Lord and Mannering (2010) and Mannering et al. (2016).

In this article, we develop a spatial autoregressive model for accident counts aggregated to squares of size 1 km^2. We use the family of extended Poisson–Tweedie distributions (Bonat et al., 2017), which provides a flexible class of models to deal with under-, equi- and overdispersed count data as well as highly skewed count data with excessive zeros as usual in traffic accidents applications. Poisson–Tweedie distributions include the Neyman Type A, Pólya-Aeppli, negative binomial and Poisson inverse-Gaussian distributions as special cases.

The dataset is presented in more detail in Section 2, Section 3 introduces Poisson–Tweedie distributions, the statistical model and elaborates on our definition of black spots. Results are given in Section 4 followed by a discussion in Section 5. In Appendix A we give a detailed description of our simulations and in Appendix B a computer code for fitting our proposed model is given.

2. Description of data

The data were collected by the Accident Analysis Group (Hansen and Lauritsen, 2008) at hospitals located on Funen, Denmark, in the period from 2002 to 2007 (Fig. 1).

Each patient reporting at a hospital as having been involved in a traffic accident was asked several questions regarding the accident location and other relevant information. For the analysis we used only accidents for which a location could be related to a house number or an intersection and we confined us to traffic accidents which occurred on public roads.

We covered Funen with a grid of 1 km^2 squares defined by the UTM coordinates (UTM zone 31N, WGS84). The injury data was quality assured and aggregated to the grid as described in Hansen and Lauritsen (2008). The quality assurance excluded hospital contacts with uncertain location (e.g. not on a regular road), imprecise geocoding (e.g. "somewhere on a 20 km long road") or at locations only occurring partially over the years. Among a total of 27,957 verified traffic accidents on public roads 13,924 (50%) could be located with a precision which allowed allocation to a given 1x1 km square and therefore inclusion in the analysis. The traffic accidents in the following analysis are the sum of the accidents in each square for each year. Using data essentially accumulated in grid cells allowed to circumvent the difficult task to relate single accident locations to specific intersections. For other aspects related to the use of grids we refer to Xie et al. (2017).

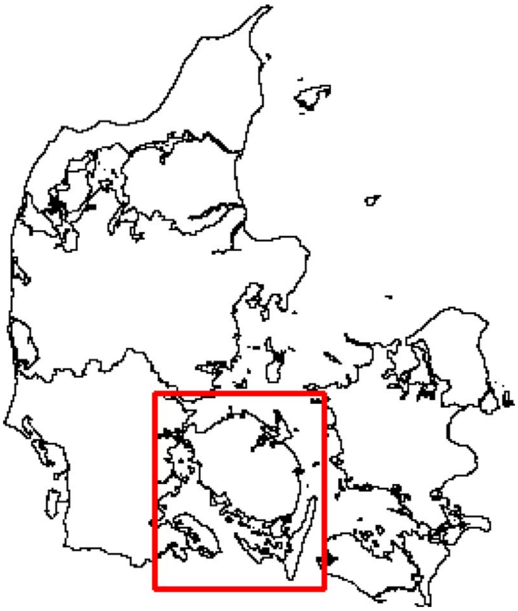

Fig. 1. Denmark with the island Funen.

Fig. 2 shows the number of accidents for the first year (2002) and the average of the year totals over the six years 2002–2007. In 2002 there was a total of 2145 traffic accidents and the average of the year totals was 2321. In 2002 in 3335 (85.3%) of the 3911 squares no accident was reported and over all six years 2427 (62.3%) locations had no reported accident.

The number of intersections and the street-length in a square are used as risk indicators for traffic accidents at locations. These values are shown in Fig. 3 and are assumed to be constant over the six years. Our data did not contain more detailed exposure data, such as accurate traffic intensity records, nor more precise information about local risk factors such as the geometry and capacity of intersections.

3. Statistical model and definition of black spots

Accident counts are known for exhibiting overdispersion and zero-inflation relative to the Poisson distribution (Lord and Mannering, 2010). However, given the wealth of discrete distributions, it is difficult to commit oneself to a single distributional model as being the most appropriate one. Therefore, and in order to take the mentioned features into account in a more unified manner, we consider the broader class of Poisson–Tweedie mixture distributions. A distribution from this family of discrete distributions (see Jørgensen and Kokonendji, 2016 for a formal definition) is specified by three parameters μ, τ and p. Here, $\mu > 0$ denotes the mean, $\tau > 0$ the dispersion and $p \geq 1$ the shape/power parameter. The variance is given by $\mu + \tau \cdot \mu^p$ and τ larger than zero indicates overdispersion. The family of Poisson–Tweedie distributions allows for zero-inflation and can further be extended to incorporate underdispersed count data with nonnegative dispersion τ, see Bonat et al. (2017) and Bonat (2016, 2017). For $p = 1$, $p = 1.5$, $p = 2$ and $p = 3$ the Poisson–Tweedie distribution respectively corresponds to Neyman type A, Pólya-Aeppli, negative binomial and Poisson-inverse Gaussian/Sichel distribution, see Kokonendji et al. (2004), all of which are well-known distributions in accident modelling (Kemp, 1967; Minkova and Balakrishnan, 2014; Özel and İnal, 2010; Lord and Mannering, 2010; Zha et al., 2016). Since the class of extended Poisson–Tweedie distributions comprises major families of distributions used for traffic modelling and is additionally richer than each single of these families alone, we consider it well-suited for our purposes.

In the sequel we denote by Y_{it}, $i = 1, \ldots, 3911$, $t = 2002, \ldots, 2007$, the number of accident counts at location i in year t and consider the following auto-regressive model containing the number of accidents from neighbouring locations, the calendar year, the street length and number of intersections as covariates:

Y_{it} is Poisson–Tweedie distributed with dispersion τ and power p.
Its mean value μ_{it} is given by

$$\log(\mu_{\text{it}}) = m_t + \sum_{d=1}^{D} a_d \log(\overline{Y}_{i,t}^{(d)} + 0.02) + b \log(S_i + 0.5) + c \log(L_i), \tag{1}$$

where S_i and L_i are the number of intersections and street length in location i, $\overline{Y}_{i,t}^{(d)}$ denotes the average accident count at time t over all neighbouring locations of cell i at distance d, and $D \in \{1, 2, \ldots\}$ is the maximum distance considered. As distance measure we use the supremum norm between the square centres. The set of all neighbouring

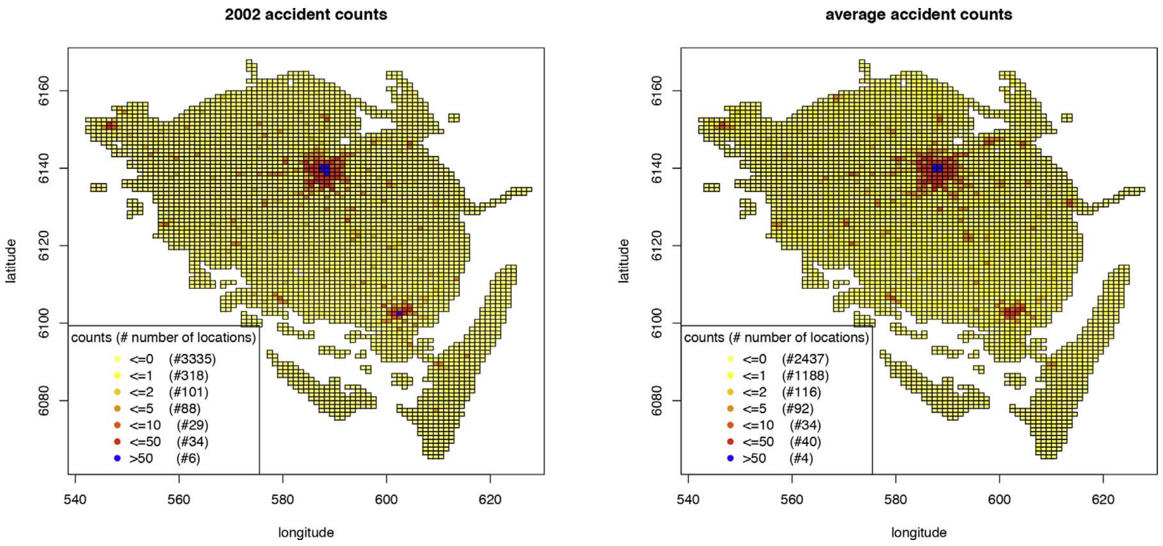

Fig. 2. Number of traffic accidents for each location in 2002 and the average of counts in the years from 2002 to 2007.

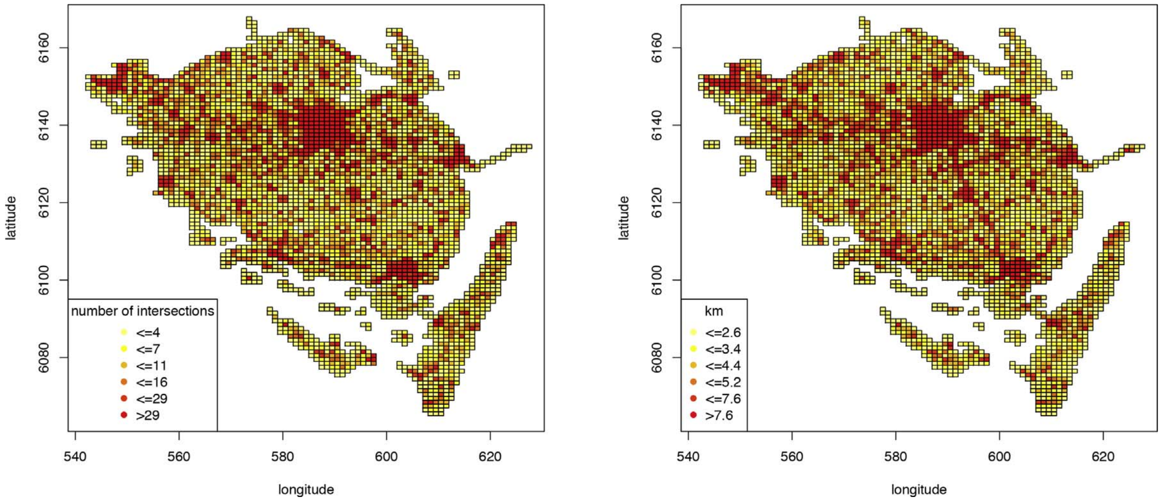

Fig. 3. Left: Number of intersections. Right: Street lengths. These are constant over all six years.

cells at a given distance will be called a layer. For example, at a supremum distance of one all neighbouring squares have an edge or vertex in common with the central square and the corresponding layer at distance one contains at most eight cells. The log-transformation turned out to be necessary to avoid model predictions that were far off the observed values. The 0.02 and 0.5 were half the minimum over all positive observations of the respective variables.

The autoregressive nature of the model takes potential unobserved effects which might be shared by nearby locations into account. The model can further be embedded into the class of multivariate covariance generalized linear models for which recently efficient estimation via estimation equations have been developed (Bonat and Jørgensen, 2016).

We fitted the Poisson–Tweedie model using the estimating function approach proposed in Bonat et al. (2017) and implemented in the mcglm package (Bonat, 2017) for the statistical software R (R Core Team, 2017). The number D of layers included in the model was determined using the pseudo-AIC criterion (Carey and Wang, 2011; Bonat, 2017).

Our concept of black spots is based on estimates of the above model given by (1) and covers locations, where the observed number of counts is unexpectedly high, given the surroundings and characteristics of the location. More formally, black spots are introduced as follows:

Definition 1. Assume that the parameters of model (1) are estimated by $\hat{m}_t$, $\hat{a}_d$, $d = 1, \ldots, D$, $\hat{b}$, $\hat{c}$, $\hat{\tau}$ and $\hat{p}$, and the estimated distribution of Y_{it} is Poisson–Tweedie with power $\hat{p}$, dispersion $\hat{\tau}$ and mean $\hat{\mu}_{\text{it}}$, where $\hat{\mu}_{\text{it}}$ at a location i is specified according to (1) (based on calendar year, surrounding locations and the general characteristics street length and number of intersections) and with the parameters substituted by its estimates.

(A) We call a location i in year t a **potential black spot at level** α, if the observed accident count Y_{it} exceeds the $(1 - \alpha)$-quantile of the Poisson–Tweedie distribution with power $\hat{p}$, mean $\hat{\mu}_{\text{it}}$ and dispersion $\hat{\tau}$.

(B) If a location i additionally exhibits property (A) in all years t, we call it a **consistent black spot at level** α.

Note that our black spot definitions reflect the requirements stated in Elvik (2007, Section 2.5) defining black spots to be any location with a higher expected number of accidents than other similar locations as a result of local risk factors.

Our approach is also similar to, e.g. Nguyen et al. (2016) who identified potential black spots as those with observed counts significantly exceeding the locations expected frequency. As in our approach, naturally occurring random variation is taken into account by significance tests by these authors.

Underlying the above definition of potential black spots is a family of statistical tests for the hypotheses $\mathscr{H}_{\text{it}}$, $i = 1, \ldots, 3911$, $t = 2002, \ldots, 2007$ (one per cell and year), where $\mathscr{H}_{\text{it}}$ describes the null-hypothesis that a cell i's expected number of accidents in year t does not systematically exceed the number of accidents expected in similar cells. Finding a potential black spot at level α in cell i and year t corresponds to rejecting the corresponding null-hypothesis at significance level α. Consequently, cells can be attached p-values reflecting the minimal α-level at which the null-hypothesis is rejected and a cell is declared a potential black spot.

Repeated testing of several hypotheses inflates the error to falsely reject at least one of these. Therefore, for a given significance level α, we also calculated the family-wise error rate (FWER), which corresponds to the probability to (falsely) declare at least one location a potential/consistent black spot provided there are no black spots on Funen. This was estimated using the Markov Chain Monte Carlo method described in Appendix A. In simple terms, data without any real black spot was simulated based on the model given by (1) together with the parameter estimates obtained from our real data. After initializing the number of accidents for each location with 0, more realistic values were obtained by iteratively drawing from the distribution specified through (1) using the values from the previous iteration (or initial values) to calculate the current iterations mean. After a sufficient number of iterations, the simulated values follow the desired distribution and represent a realization of traffic accidents in the different locations when black spots are absent. These realizations were then used to estimate the FWER.

4. Results

We chose an appropriate neighbourhood-size using a pseudo-AIC. Fig. 4 shows the pseudo-AIC against the number D of neighbouring layers. Thereby, pseudo-AIC's drop rapidly from $D = 1$ to $D = 2$ followed by the same level until $D = 6$ where after it drops again. In order to keep a parsimonious model we chose to use a model with $D = 2$.

The estimates of the parameters of the model are given in Table 1. Although none of the parameters m_t, $t \geq 2003$, (here denoted as contrasts for the years compared to 2002), appears to be significant at the 5% level, a Wald test rejects the joint hypothesis of no systematic difference between any of the years with a p-value below 0.001, hence the expected number of accidents changes systematically with time. The other parameters of model (1) corresponding to accidents in neighbouring cells, number of intersections and street length (a_d, b and c) have p-values less than 0.001. The estimated dispersion $\hat{\tau} = 0.944$ ($SE = 0.084$) indicates the presence of overdispersion and together with the power parameter estimate $\hat{p} \approx 1.6$ ($SE = 0.062$) suggests that the Pólya-Aeppli distribution (where $p = 1.5$) would reflect the given data reasonably well.

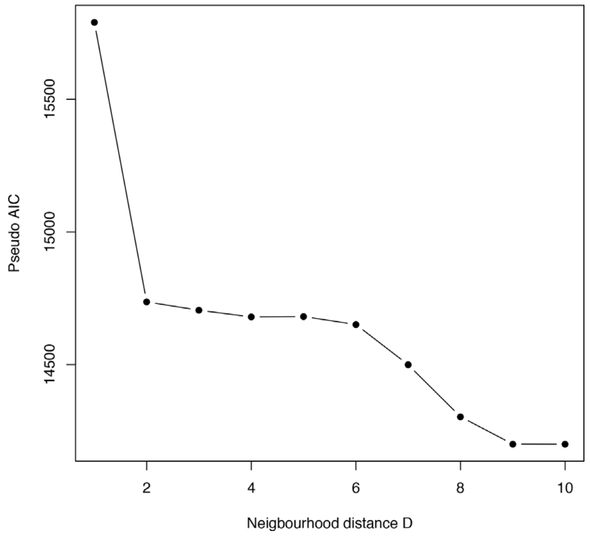

Fig. 4. Pseudo-AIC in dependence of the number D of neighbouring layers.

Table 1
The parameter estimates of the model in Eq. (1).

Parameter	Estimate	Standard-error	p-value
$\hat{m}$	−11.141	0.456	< 0.001
$\hat{m}_{2003}$	0.062	0.053	0.2347
$\hat{m}_{2004}$	0.077	0.052	0.1384
$\hat{m}_{2005}$	0.053	0.053	0.3133
$\hat{m}_{2006}$	−0.008	0.054	0.8851
$\hat{m}_{2007}$	−0.037	0.054	0.4898
$\hat{a}_1$	0.287	0.015	< 0.001
$\hat{a}_2$	0.127	0.014	< 0.001
$\hat{b}$	0.349	0.037	< 0.001
$\hat{c}$	1.104	0.062	< 0.001
$\hat{p}$	1.600	0.062	
$\hat{\tau}$	0.944	0.084	

The fitted model can be used for black spot identification, cp. Definition 1, and Fig. 5 presents the thirteen potential black spots with α below 0.0005. If corrected for multiplicity in testing the adjusted family-wise error rates α_{FWER} are all above 0.256, see (Hochberg and Tamhane, 2008, Ch. 1) and cp. equation (2) in the Appendix A. This means, in the absence of real black spots, there would still be an approx. 26% chance, to find at least one location with level 0.0005, which then would be falsely declared a black spot.

Note however, that none of the potential black spots shown in Fig. 5 stands out consistently, i.e. attains p-values below 0.0005 in all years. More specifically, the locations with the lowest p-values vary from year to year, and none of the sites occurred twice, i.e. in two different years among the potential black spots at level 0.0005.

For less stringent levels, consistent black spots can be found and are given in Table 2 together with the probability for a single location to be flagged as a consistent black spot under the assumption that no black spot exists (that is, assuming model (1) holds with parameter estimates given in Table 1 for all locations and years). Additionally, Table 2 contains analogous probabilities but based on a Pólya-Aeppli distribution with unit variance function (i.e. $p = 1.5$ and $\tau = 1$), revealing only negligible differences.

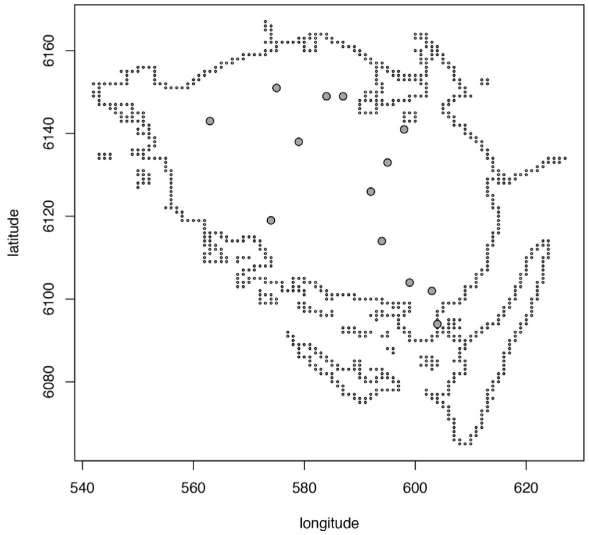

Fig. 5. The 13 potential black spots with levels α below 0.0005. Adjusted family-wise error rates α_{FWER} ranged between 0.256 and 0.999.

Table 2
Observed number of consistent black spots and probability for these to be chance findings based on 1000 simulations and the model specified in Table 1. (values marked by * are instead based on a Pólya-Aeppli model with unit variance function, i.e. using the model from Table 1 but setting $p = 1.5$ and $\tau = 1$).

Level α	Nr. of consistent black spots at level α	Simulated probability to find at least one consistent black spot by chance
0.5	43	0.948 (0.930)*
0.4	28	0.568 (0.539)*
0.3	14	0.159 (0.151)*
0.25	9	0.064 (0.055)*
0.24	9	0.051 (0.045)*
0.23	9	0.046 (0.032)*
0.22	9	0.041 (0.027)*
0.21	9	0.030 (0.025)*
0.2	8	0.026 (0.017)*
0.15	7	0.002 (0.005)*
0.1	7	(0) (0)*
0.05	3	(0) (0)*
0.01	0	(0) (0)*

Fig. 6 displays the nine consistent black spots detected using different levels below $\alpha = 0.24$ corresponding to a probability for at least one consistent black spot to be detected by chance below 0.051. Table 3 contains the characteristics of the nine consistent black spots.

Amongst the consistent black spots found are locations with high as well as low traffic intensity. This reflects our definition of black spots, which is not solely focussing on a high number of injuries, but rather observed accident counts significantly exceeding what is expected for similar locations. Finally note that although our nine consistent black spots are at the same time potential black spots at the same level ($\alpha = 0.23$), it is not to be expected to find all of these already at lower α levels.

5. Discussion

We studied a dataset of traffic accidents based on hospital admissions on Funen, Denmark. To model the expected number of accidents in squares of 1 km^2, we considered a spatial autoregressive model taking neighbouring accident counts, calendar year, street length and number of intersections into account. As a flexible distributional model we used the Poisson–Tweedie mixture distributions which are able to handle overdispersion as well as zero-inflation, both typical for accident data. As a result, we have identified nine consistent black spot quadrants, which can be further studied locally. This leads us to assume that this modelling strategy could play a role in future priority settings of specific targeted area interventions.

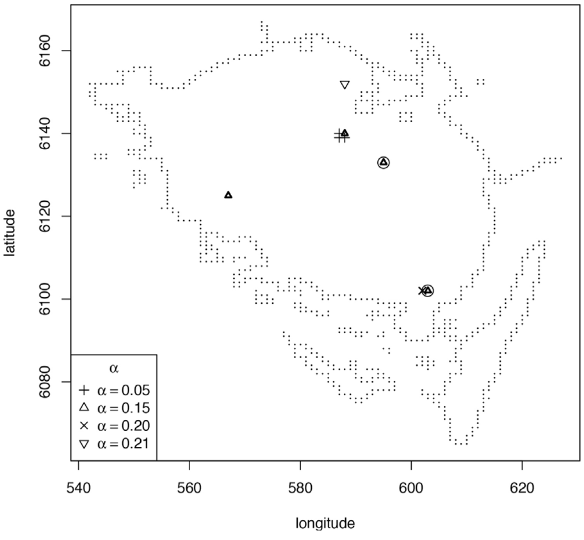

Fig. 6. Consistent black spots over six years with levels α smaller than 0.21. Locations with a lower level were also consistent black spots with a higher level. The two locations with a circle were also potential black spots in Fig. 5.

Table 3
Characterization of the nine consistent black spots shown in Fig. 6. Longitude and latitude are UTM zone 31N coordinates in km. The column ‘count’ shows the average counts over the six years, the column ‘fit’ the corresponding average across the fitted values. Ranges are taken over corresponding values from the different calendar years.

α	Count (range)	Fit (range)	Intersections	Street length	Longitude	Latitude
0.05	81.2 (65,96)	32.7 (29,36)	135	31	587	6140
0.05	89.8 (68,118)	35.6 (33,39)	135	32	588	6139
0.05	97.7 (76,124)	40.8 (37,45)	201	32	587	6139
0.15	1.3 (1,3)	0.1 (0, 0)	4	3	567	6125
0.15	3.0 (2,5)	0.3 (0, 1)	26	6	595	6133
0.15	13.7 (7,19)	2.8 (2, 3)	39	9	603	6102
0.15	73.3 (64,85)	38.4 (35,43)	162	34	588	6140
0.20	36.0 (19,52)	15.1 (12,18)	158	27	602	6102
0.21	8.7 (5,13)	3.1 (2, 4)	99	18	588	6152

One of the features of our data was the lack of reliable exposure data such as detailed and accurate traffic intensity records, as well as the lack of reliable information about local risk factors such as the geometry and capacity of intersections and road segments, which many previous studies have used (see, e.g. Elvik, 2007; Li et al., 2007). Although these and other omitted variables could potentially lead to biased parameter estimates and erroneous inferences, we do believe, that our model is reasonably realistic and street length and intersections serve as good proxies for traffic intensities.

The estimated power parameter of $\hat{p} = 1.6$ indicates that the distribution of accidents is close to a Pólya-Aeppli distribution also seen by the similarity of the probabilities shown in Table 2 for detection of a consistent black spot by chance. The Pólya-Aeppli distribution is a compound Poisson distribution with geometrically distributed summands, and in connection with traffic models, a possible underlying mechanisms could be a Poisson distributed number of accidents with a geometrically distributed number of injured persons per accident, as suggested by Özel and İnal (2010). To investigate to which extent the identified blackspots depend on the chosen distribution, we fitted the following alternative distributions (all of which members of the Poisson–Tweedie family): Neyman A, negative binomial and Sichel distribution. We found that only for the Neyman A and Sichel distribution the pseudo AICs were appreciable different from the one corresponding to our fitted model. All distributions but Neyman A led to the same nine leading black spots, the results for the Neyman A distribution being different in one location.

A further challenge in black spot identification is the multiplicity in testing when searching through individual cells in several years. Although we estimated the effective number of underlying independent hypotheses to be 13,401, cp. Appendix A, which is considerably lower than 3911 · 6, the multiple testing burden is still present. As expected, many potential black spots at low α levels in a specific year cannot be found to be black spots in other years. Our concept of consistent black spots therefore aims at locations, which stand out in all investigated years. Comparable approaches searching consistently noticeable locations are well known in traffic accident research, cp.(Elvik, 2007, Section 2.6). Our data cover six years, which can be debated. The analysis shows, that one year incidents are not consistently identified as “black spots”, but one could then argue why exactly six years. We have no precise answer to this question, but from a practical point of view, we are convinced that with extended length in time we would also need data on road changes. Further studies should look more into the question of stability of locations versus length of period.

The fact that the observed number of potential (after correcting for multiplicity) and consistent black spots on Funen is rather low, could be partially inherent to our approach: A black spot is a location with (observed) accident counts, that are unexpectedly high given the location's characteristics. However, what is expected for this and similar locations is predicted by our model estimated using the whole data, that is, including the black spots themselves. By doing so, it is anticipated, that predictions overestimate reality, at the same time hindering the identification of the underlying black spots. It seems however (simulations not shown) that our approach is reasonably stable in that a minor number of black spots (given the total of 3911 locations) does not considerably corrupt the estimated model parameters.

Analyses of the injuries in the current paper were previously included in an attempt to find black spots based on specific locations of road crossings (Hansen and Lauritsen, 2011), but these were only based on raw counts not on regression modelling. A major drawback to the current paper was the lack of precise geocoding for about half of the accidents. Since very few hospitals can actually provide such data we do find the current study an important addition to black spot definitions based on modern statistical modelling principles. Further studies of management of the imprecision issue and resulting implications for actual black spot identification are welcomed. Our data will be made available for such studies in general. From the municipalities there has been high interest in getting further elaborated results. For priority setting, we have therefore seen a possible strategy to first assess in the current paper whether we could develop a focus area principle (black spots in quadrants based on modelling), and if so the perspective would be that routine implementation of such analysis could give municipal road authorities a map of potential areas for further detailed scrutiny and then detailed observations. This would mediate the reality that background data like road structures or traffic volumes are known within the municipalities partially, but not in a structured sense for the whole area.

Acknowledgements

The general ideas of this paper owe much to the insight of the late Bent Jørgensen.

Funding was attained from Trygfonden (ID: 101703). Without meticulous work by the secretaries in the hospital receptions and patients to inform on location for occurrence data would not have existed.

Appendix A. Simulations

In order to correct for multiplicity in testing when searching for black spots, we consider the probability for these to occur by chance, i.e. in a hypothetical situation where the dataset does not contain any real black spot. As exact formulas are unknown, we estimate these probabilities by means of simulations. Thereby, to simulate black spot free accident counts Y_{1t}, …, Y_{3911t} which follow the distribution given by (1), whereby parameters are substituted by corresponding estimates, we use the following Markov Chain Monte Carlo method known as asynchronous Gibbs sampling (Terenin et al., 2015), which generates the iterates $Y_{\text{it}}^{(k)}$, where k refers to the iteration number:

- Initialization: Set $Y_{\text{it}}^{(0)} = 0$ for $i = 1, \ldots, 3911$ and $t = 2002, \ldots, 2007$.
- Iteration step: For $t = 2002, \ldots, 2007$ and $i = 1, \ldots, 3911$ do:
  Simulate $Y_{\text{it}}^{(k)}$ according to Eq. (1) but using the counts $Y_{\text{jt}}^{(k-1)}$, $j = 1, \ldots, 3911$, from the previous step in the calculation of a cell's expected value $\mu_{\text{it}}^{(k)}$.

Specifically, for each k the simulated values $Y_{\text{it}}^{(k)}$ represent accident counts for all six years for each cell on Funen. The described procedure belongs to the class of parametric bootstrap methods in that simulations are based on a parametric model with parameters given by a fitted model.

In the sequel, we simulated 1000 independent chains, discarding the first 20 iterations as burn-in, and only using one iterate per chain. (This number of burn-in iterations has empirically proven to be sufficient.) These simulations will be denoted by $Y_{\text{it}}^{\text{sim},1}$, $Y_{\text{it}}^{\text{sim},2}$, …, $Y_{\text{it}}^{\text{sim},1000}$.

For potential black spots, defined based on a location's value in a specific year, we then apply the concept of family-wise error rates (Hochberg and Tamhane, 2008) in order to take multiple testing into account. Thereby, testing a family of null hypotheses is said to have a family-wise error rate of α_{FWER} if the probability to reject at least one hypothesis is α_{FWER} given that all null hypotheses are true. Using the simulated data, the family-wise error rate is estimated as follows:

- For each simulation $j = 1, \ldots, 1000$ separately, we estimated the parameters of model (1) and calculated cell- and year-wise p-values $\alpha_{\text{it}}^{\text{sim},j}$, $i = 1, \ldots, 3911$, $t = 2002, \ldots, 2007$, given by

  $$\alpha_{\text{it}}^{\text{sim},j} = \min\{\alpha | Y_{\text{it}}^{\text{sim},j} \geq \text{PTw}_{\hat{p}^{\text{sim},j}}(\hat{\mu}_{\text{it}}^{\text{sim},j}, \hat{\tau}^{\text{sim},j}; 1 - \alpha)\}.$$

  Hereby, $\text{PTw}_{\hat{p}^{\text{sim},j}}(\hat{\mu}_{\text{it}}^{\text{sim},j}, \hat{\tau}^{\text{sim},j}; 1 - \alpha)$ denotes the empirical $(1 - \alpha)$-quantile of a Poisson–Tweedie mixture distribution with parameters substituted by their corresponding estimates, i.e. mean $\hat{\mu}_{\text{it}}$, power $\hat{p}$ and dispersion $\hat{\tau}$.
- For each simulation j, let $\alpha_{\text{min}}^{\text{sim},j} = \min_{i,t} \alpha_{\text{it}}^{\text{sim},j}$ denote the minimum over all cells $i = 1, \ldots, 3911$ and years $t = 2002, \ldots, 2007$.
- The α level for a potential black spot corresponding to a family-wise error rate of α_{FWER} was estimated by the α_{FWER}-quantile $F_{\text{min};\alpha_{\text{FWER}}}^{\text{sim}}$ of the sample of simulation-wise minima $\alpha_{\text{min}}^{\text{sim},1}$, $\alpha_{\text{min}}^{\text{sim},2}$, …, $\alpha_{\text{min}}^{\text{sim},1000}$.
  Especially, for $\alpha_{FWER} = 0.05$ resp. 0.1, we obtained levels of $\alpha = 3.02\text{e}-6$ resp. $\alpha = 6.5\text{e}-6$.

We further calculated an estimate for the effective number n_{eff} of independent tests by

$$\text{median}_{\alpha_{\text{FWER}} \in \{0.01, 0.02, \ldots, 0.99\}} \frac{\log(1 - \alpha_{\text{FWER}})}{\log(1 - F_{\text{min};\alpha_{\text{FWER}}}^{\text{sim}})} \approx 13401,$$

when searching for potential black spots among all cells in all years performed. Hence, testing all locations in all years corresponds to 13401 independent tests, which is roughly half of the actual number of tests performed. Using n_{eff}, family-wise error rates α_{FWER} can now be directly translated into α-levels (and vice versa) through

$$\alpha \approx 1 - (1 - \alpha_{\text{FWER}})^{1/n_{\text{eff}}}. \tag{2}$$

For consistent black spots, we used the simulated data to estimate the probability to find at least one black spot by chance for varying levels of α.

Appendix B. Code for fitting the model in Eq. (1)

```
#Installing the mcglm package from CRAN (if not yet installed)
install.packages("mcglm")
#Loading mcglm package
library(mcglm)
#Loading the Matrix package
library(Matrix)
#Reading data
load("Darticle.Rdata")
# count: accidents counts for each location and each year
# yearf: factor variable containing the years
# logintersect: the logarithm of (the number of intersection + 0.5)
# logstreet: logarithm of the street lengths
# X1: the logarithm of (average counts in the sumpremum 1 neighbourhood + 0.02)
# X2: the logarithm of (average counts in the sumpremum 2 neighbourhood + 0.02)

#Setting up the formula
formglm < - formula("count ~ yearf + logintersect + logstreet + X1 + X2")
#Simple fit of a generalized linear model to get initial parameter estimates
modelStart < - glm(formglm, family = quasipoisson, data = Darticle)

#Setting initial values
Z0 < -Diagonal(nrow(Darticle), 1)
list_initial = list()
list_initial$regression=list(coef(modelStart))
list_initial$power < - list(1.5)
list_initial$tau < - list(0.5)
list_initial$rho = 0

#Fitting the mcglm-model
model < - mcglm(linear_pred = c(formglm),
matrix_pred = list(list(Z0)),
link = "log", variance = "poisson_tweedie",
data = Darticle,
control_initial = list_initial,
power_fixed=FALSE)

# Print model parameters
summary(model)
```

References

Aguero-Valverde, J., Jovanis, P., 2008. Analysis of road crash frequency with spatial models. Transp. Res. Rec.: J. Transp. Res. Board 55–63.

Amoros, E., Martin, J.L., Laumon, B., 2003. Comparison of road crashes incidence and severity between some French counties. Accid. Anal. Prev. 35, 537–547.

Barua, S., El-Basyouny, K., Islam, M.T., 2015. Effects of spatial correlation in random parameters collision count-data models. Anal. Methods Accid. Res. 5, 28–42.

Bonat, W.H., 2016. mcglm: Multivariate Covariance Generalized Linear Models. R package version 0.3.0. https://CRAN.R-project.org/package=mcglm.

Bonat, W.H., 2017. Multiple response variables regression models in R: the mcglm package. J. Stat. Softw (in press).

Bonat, W.H., Jørgensen, B., 2016. Multivariate covariance generalized linear models. J. R. Stat. Soc. Appl. Stat. C 65, 649–675.

Bonat, W.H., Jørgensen, B., Kokonendji, C.C., Hinde, J., Demétrio, C.G.B., 2017. Extended Poisson–Tweedie: properties and regression models for count data. Stat. Modell. 26. http://dx.doi.org/10.1177/1471082X17715718. (published online August 2017).

Buddhavarapu, P., Scott, J.G., Prozzi, J.A., 2016. Modeling unobserved heterogeneity using finite mixture random parameters for spatially correlated discrete count data. Transp. Res. Part B: Methodol. 91, 492–510.

Cafiso, S., Di Graziano, A., Di Silvestro, G., La Cava, G., Persaud, B., 2010. Development of comprehensive accident models for two-lane rural highways using exposure, geometry, consistency and context variables. Accid. Anal. Prev. 42, 1072–1079.

Carey, V.J.C., Wang, Y., 2011. Working covariance model selection for generalized estimating equations. Stat. Med. 30, 3117–3124.

De Pauw, E., Daniels, S., Brijs, T., Hermans, E., Wets, G., 2014. Safety effects of an extensive black spot treatment programme in Flanders-Belgium. Accid. Black Spot Detect. Traff. Accid. Anal. Prev. 66, 72–79.

Depaire, B., Wets, G., Vanhoof, K., 2008. Traffic accident segmentation by means of latent class clustering. Accid. Anal. Prev. 40, 1257–1266.

Dong, C., Clarke, D.B., Yan, X., Khattak, A., Huang, B., 2014. Multivariate random-parameters zero-inflated negative binomial regression model: an application to estimate crash frequencies at intersections. Accid. Anal. Prev. 70, 320–329.

Dupont, E., Papadimitriou, E., Martensen, H., Yannis, G., 2013. Multilevel analysis in road safety research. Accid. Anal. Prev. 60, 402–411.

El-Basyouny, K., Sayed, T., 2009. Accident prediction models with random corridor parameters. Accid. Anal. Prev. 41, 1118–1123.

Elvik, R., 2007. State-of-the-Art Approaches to Road Accident Black Spot Management and Safety Analysis of Road Networks. Technical Report. Institute of Transport and Economic, Norwegian Centre for Transport Research, Norway.

Elvik, R., 2008. The predictive validity of empirical Bayes estimates of road safety. Accid. Anal. Prev. 40, 1964–1996.

Hansen, D., Lauritsen, J.M., 2008. Localization and reporting of data from emergency departments from tables to GIS (in Danish). Geoforum Perspektiv 13, 25–31.

Hansen, D., Lauritsen, J.M., 2011. Identification of black spots for traffic injury in road intersections dependence of injury definition. Inj. Prev. 16, 261.

Heydari, S., Fu, L., Miranda-Moreno, L.F., Joseph, L., 2017. Using a flexible multivariate latent class approach to model correlated outcomes: a joint analysis of pedestrian and cyclist injuries. Anal. Methods Accid. Res. 13, 16–27.

Hochberg, Y., Tamhane, A.C., 2008. Introduction. In: Multiple Comparison Procedures. John Wiley & Sons, Inc.. pp. 1–16.

Jones, A.P., Jørgensen, S.H., 2003. The use of multilevel models for the prediction of road accident outcomes. Accid. Anal. Prev. 35, 59–69.

Jørgensen, B., Kokonendji, C.C., 2016. Discrete dispersion models and their Tweedie asymptotics. AStA Adv. Stat. Anal. 100, 43–78.

Joshua, S.C., Garber, N.J., 1990. Estimating truck accident rate and involvements using linear and Poisson regression models. Transp. Plann. Technol. 15, 41–58. http://dx.doi.org/10.1080/03081069008717439.

Jovanis, P.P., Li Chang, H., 1986. Modeling the Relationship of Accidents to Miles Traveled. Transportation Research Record, No. 1068. TRB, National Research Council, pp. 42–51.

Kemp, C.D., 1967. On a contagious distribution suggested for accident data. Biometrics 23, 241–255.

Kim, D.G., Lee, Y., Washington, S., Choi, K., 2007. Modeling crash outcome probabilities at rural intersections: application of hierarchical binomial logistic models. Accid. Anal. Prev. 39, 125–134.

Kokonendji, C.C., Demétrio, C.G.B., Dossou-Gbété, S., 2004. Overdispersion and Poisson–Tweedie exponential dispersion models. Monografías del Seminario Matemático García de Galdeano 31, 365–374.

Lauritsen, J.M., Kidholm, K., Skov, O., Nørgård, L., 2002. Average costs and an proportions of total costs related to hospital related injuries (in Danish). Ugeskrift for Laeger 164, 5107–5112.

Li, L., Zhu, L., Sui, D., 2007. A GIS-based Bayesian approach for analyzing spatial–temporal patterns of intra-city motor vehicle crashes. J. Transp. Geogr. 15, 274–285.

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

Lord, D., Miranda-Moreno, L.F., 2008. Effects of low sample mean values and small sample size on the estimation of the fixed dispersion parameter of Poisson-gamma models for modeling motor vehicle crashes: a Bayesian perspective. Saf. Sci. 46, 751–770.

Lord, D., Washington, S., Ivan, J.N., 2007. Further notes on the application of zero-inflated models in highway safety. Accid. Anal. Prev. 39, 53–57.

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

Malyshkina, N.V., Mannering, F.L., 2009. Markov switching multinomial logit model: an application to accident-injury severities. Accid. Anal. Prev. 41, 829–838.

Malyshkina, N.V., Mannering, F.L., 2010. Zero-state Markov switching count-data models: an empirical assessment. Accid. Anal. Prev. 42, 122–130.

Mannering, F.L., Shankar, V., Bhat, C.R., 2016. Unobserved heterogeneity and the statistical analysis of highway accident data. Anal. Methods Accid. Res. 11, 1–16.

Maycock, G., Hall, R., 1984. Accidents at 4-Arm Roundabouts.

Miaou, S.P., 1994. The relationship between truck accidents and geometric design of road sections: Poisson versus negative binomial regressions. Accid. Anal. Prev. 26, 471–482.

Miaou, S.P., Bligh, R., Lord, D., 2005. Part 1: Roadside safety design: developing guidelines for median barrier installation: benefit–cost analysis with Texas data. Transp. Res. Rec.: J. Transp. Res. Board 2–19.

Miaou, S.P., Lord, D., 2003. Modeling traffic crash-flow relationships for intersections: dispersion parameter, functional form, and Bayes versus empirical Bayes methods. Transp. Res. Rec.: J. Transp. Res. Board 31–40.

Miaou, S.P., Lum, H., 1993. Modeling vehicle accidents and highway geometric design relationships. Accid. Anal. Prev. 25, 689–709.

Miaou, S.P., Song, J.J., Mallick, B.K., 2003. Roadway traffic crash mapping: a space-time modeling approach. J. Transp. Stat. 6, 33–58.

Minkova, L.D., Balakrishnan, N., 2014. On a bivariate Pólya-Aeppli distribution. Commun. Stat. Theory Methods 43, 5026–5038.

Nguyen, H.H., Taneerananon, P., Luathep, P., 2016. Approach to identifying black spots based on potential saving in accident costs. Eng. J. 20, 109–122.

Özel, G., İnal, C., 2010. The probability function of a geometric Poisson. J. Stat. Comput. Simul. 80, 479–487.

Park, B.J., Lord, D., 2009. Application of finite mixture models for vehicle crash data analysis. Accid. Anal. Prev. 41, 683–691.

Pei, X., Wong, S., Sze, N., 2011. A joint-probability approach to crash prediction models. Accid. Anal. Prev. 43, 1160–1166.

R Core Team, 2017. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Shankar, V., Albin, R., Milton, J., Mannering, F., 1998. Evaluating median crossover likelihoods with clustered accident counts: an empirical inquiry using the random effects negative binomial model. Transp. Res. Rec.: J. Transp. Res. Board 44–48.

Terenin, A., Simpson, D., Draper, D., 2015. Asynchronous Gibbs Sampling. arXiv:1509.08999.

Thomas, B., DeRobertis, M., 2013. The safety of urban cycle tracks: a review of the literature. Accid. Anal. Prev. 52, 2019–2227.

Turner, S., Nicholson, A., 1998. Using accident prediction models in area wide crash reduction studies. In: Road Engineering Association of Asia and Australasia (REAAA), Conference, 9th. 1998, Wellington, New Zealand, vol. 1. pp. 255–260.

Vandenbulcke, G., Thomas, I., Panis, L., 2014. Predicting cycling accident risk in Brussels: a spatial case–control approach. Accid. Anal. Prev. 62, 341–357.

Venkataraman, N., Ulfarsson, G.F., Shankar, V.N., 2013. Random parameter models of interstate crash frequencies by severity, number of vehicles involved, collision and location type. Accid. Anal. Prev. 59, 309–318.

WHO, 2013. Global Status Report on Road Safety 2013: Supporting a Decade of Action. Report. World Heath Organisation, Geneva.

Xie, K., Ozbay, K., Kurkcu, A., Yang, H., 2017. Analysis of traffic crashes involving pedestrians using big data: investigation of contributing factors and identification of hotspots. Risk Anal. 37, 1459–1476.

Zeng, Q., Huang, H., Pei, X., Wong, S., 2016. Modeling nonlinear relationship between crash frequency by severity and contributing factors by neural networks. Anal. Methods Accid. Res. 10, 12–25.

Zha, L., Lord, D., Zou, Y., 2016. The Poisson inverse Gaussian (PIG) generalized linear regression model for analyzing motor vehicle crash data. J. Transp. Saf. Secur. 8, 18–35.