Accident Analysis and Prevention 132 (2019) 105252

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

# Spatial analysis of traffic accidents near and between road intersections in a directed linear network

Álvaro Briz-Redón*, Francisco Martínez-Ruiz, Francisco Montes

*Statistics and Operations Research, University of Valencia, C/ Dr. Moliner, 50, 46100 Burjassot, Spain*

ARTICLE INFO

*Keywords:*
Segment and intersection accident analysis
Accident causality
Linear network
Spatial count models
Hotspot detection
Kernel density estimation

ABSTRACT

Although most of the literature on traffic safety analysis has been developed over areal zones, there is a growing interest in using the specific road structure of the region under investigation, which is known as a linear network in the field of spatial statistics. The use of linear networks entails several technical complications, ranging from the accurate location of traffic accidents to the definition of covariates at a spatial micro-level.

Therefore, the primary goal of this study was to display a detailed analysis of a dataset of traffic accidents recorded in Valencia (Spain), which were located into a linear network representing more than 30 km of urban road structure corresponding to one district of the city. A set of traffic-related covariates was constructed at the road segment level for performing the analysis. Several issues and methodological approaches that are inherent to linear networks have been shown and discussed. In particular, the network was defined in a way that allowed the explicit investigation of traffic accidents around road intersections and the consideration of traffic flow directionality.

Zero-inflated negative binomial count models accounting for spatial heterogeneity were used. Traffic safety at road intersections was specifically taken into account in the analysis by considering the higher variability and number of zeros that can be observed at these road entities and the differential contribution of the covariates depending on the proximity of a road intersection. To complement the results obtained from the count models fitted, coldspots and hotspots along the network were also detected, with explanatory objectives.

The models confirmed that spatial heterogeneity, overdispersion and the close presence of road intersections explain the accident counts observed in the road network analyzed. Hotspot detection revealed that several covariates whose contribution was unclear in the modelling approaches may also be affecting accident counts at the road segment level.

## 1. Introduction

Traffic accidents are still a quite frequent cause of death for the European population, especially in the younger age groups. Even though the number of accidents has gradually decreased in the most developed countries of the world during the last decade, many efforts, in terms of prevention and road planning, are still being made to reduce their occurrence and severity. In this regard, studies aimed at analyzing the occurrence and distribution of traffic accidents can be very helpful, and could be broadly classified according to three main objectives: finding road and/or traffic characteristics associated with a higher occurrence of accidents, detecting zones with a high concentration of accidents and discovering the types of accidents that tend to produce more serious consequences for the vehicle passengers or road users involved. In this paper, the modelling of accident counts at the road segment level with explanatory purposes is the main goal, although the detection of microzones of the network that show a singular risk of accident is also carried out. This section starts with a literature review on both topics: modelling traffic accidents outcomes and finding zones of high accident risk. This is followed by a review of the literature on the analysis of traffic accidents occurring in intersection and non-intersection zones. This issue has also been addressed in the analysis contained in this paper.

### 1.1. Review of models and methods

Many important quantitative studies that have focused on factors that may be affecting traffic safety have been carried out through areal units of analysis. For instance, Quddus (2008) modelled traffic accident counts at the census ward level, which made it possible to explain the

* Corresponding author.
*E-mail address:* alvaro.briz@uv.es (Á. Briz-Redón).

https://doi.org/10.1016/j.aap.2019.07.028
Received 17 April 2019; Received in revised form 25 July 2019; Accepted 26 July 2019
Available online 19 August 2019

number of accidents from information related to traffic characteristics (volume and speed), road design and socio-demographic factors. Traffic volume and a proxy for poverty showed a significant positive association with traffic accidents. Similarly, Huang et al. (2010) studied traffic accident frequency at the county level considering traffic-related, demographic and socioeconomic characteristics of the counties being studied. This work focused on distinguishing two types of exposure variables: population and average daily vehicle miles travelled (DVMT) per county. The model using DVMT as the exposure yielded more significant associations with traffic accidents, some of which were positive (traffic intensity, density of principal and minor arterials, and percentage of young population) and others were negative (freeway density and average travel time to work).

In order to favour more accurate investigations, road networks have increasingly been used in traffic safety analysis in the last few years. The use of these structures, composed of links (segments) and vertices (points where two or more links meet), is becoming more popular, in spite of the technical difficulties their use entails. In this regard, it is worth noting that many of the factors that have been proved to generally increase the occurrence of traffic accidents require a road segment level analysis. Indeed, the list of infrastructure characteristics that were determined to be risky for drivers and users in a recent systematic review of published studies by Papadimitriou et al. (2019) included traffic volume, road surface (low friction), low curve radius, number of lanes, absence of paved shoulders, narrow shoulders, different junction types, etc. Although an areal-based analysis may also help to gain knowledge about the association of traffic accidents with any of these road characteristics, a road segment level analysis would be recommended to guarantee an appropriate investigation.

Given the convenience of using road networks for analyzing traffic accident outcomes, this paragraph includes a description of several studies that were performed at the road segment level, which enabled their corresponding authors to properly investigate certain infrastructure characteristics. For example, Aguero-Valverde and Jovanis (2008) found a positive association of traffic volume and certain shoulder widths with traffic accidents. In addition, Guo et al. (2017) developed a measure (integration) which reflects the accessibility of a node in the network, depending on its neighbourhood geometry. It was found that networks with a high integration value, which usually resemble a grid pattern, tend to be associated with more traffic accidents. Finally, Barua et al. (2016) analyzed severe and no-injury traffic accidents at the road segment level, finding that road segment length, average annual daily traffic, density of unsignalized intersections, business land use and the number of lanes showed a significant and positive association with both accident types.

On the other hand, several studies that have incorporated linear networks to treat accident datasets have only focused on detecting zones with a high concentration of accidents (hotspots). Indeed, Huang et al. (2016) suggested that the detection of hotspots at the micro-level is more accurate and useful for revealing risky road configurations than the use of areal macro-zones. For example, Xie and Yan (2013) applied kernel density estimation (KDE) to a linear network structure to evaluate distribution of traffic accidents and to find clusters of roads with a high proportion of accidents. They studied the impact of subdividing the network into shorter spatial units (segments), called lixels (Xie and Yan, 2008), and the variations observed depending on the choice of the kernel bandwidth parameter. A similar approach was taken by Nie et al. (2015) to prove that the application of network KDE improved the performance of local indicators of spatial association (LISA) to better identify accident hotspots.

To finish our literature review, we need to highlight certain aspects that deserve attention every time a statistical modelling of accident counts is performed. First, regardless of the choice of areal units or road segments for conducting the analysis, the consideration of spatial effects has almost become a requirement (Mannering and Bhat, 2014; Mannering et al., 2016). Getting back to some of the studies described above, some of them showed that the use of non-spatial models can lead to either spatially autocorrelated model residuals Quddus (2008), Huang et al. (2010) or to a significantly lower fit to the data (Aguero-Valverde and Jovanis, 2008). Both facts suggest that overlooking spatial effects is inappropriate. Moreover, Xu et al. (2017) tested a modification of the model proposed in Huang et al. (2010) that allowed the effects of the covariates to vary spatially. These authors observed that it is even advisable to include such variations, as otherwise biased estimates of the model's coefficients may arise.

Besides the consideration of spatial heterogeneity, other issue that often arises when performing a road segment level modelling of accident counts is the high presence of zeros (segments where no accident has been recorded). Zeng et al. (2017) used a Tobit model to control for left-censored accident rates that may be the consequence of under-reporting. Speed was associated with higher crash rates, whereas average annual daily traffic displayed a significant negative correlation. Anastasopoulos (2016) compared multivariate Tobit and zero-inflated models for modelling accident counts with a high percentage of zeros. Both strategies showed their own limitations, but each was capable of capturing zero-state heterogeneity across the road network.

### 1.2. *Traffic safety at road intersections*

The high rates of traffic accidents that are usually observed in proximity to road intersections is the reason for the existence of many studies on this topic. Thus, this paragraph includes a literature review (in chronological order) on the topic of modelling the occurrence of traffic accidents around road intersections. For instance, Castro et al. (2012) studied the spatio-temporal incidence of accident counts at urban intersections. It proved to be advisable to consider both the spatial and the temporal effect, and a significant effect was found for roadway configuration, approach roadway typology and traffic flow, among other factors. Xie et al. (2014) also developed several modelling approaches to analyze accident occurrence at intersections. The consideration of a hierarchical spatial model accounting for the effects produced at intersections by contiguous segments (corridor-level) clearly outperformed the rest of the models applied. Huang et al. (2017) analyzed accident counts at road intersections considering types of users (pedestrians, bicycles or motor vehicles) involved in accidents with a multivariate Poisson lognormal regression model. Moreover, Lee et al. (2017) used a mixed effects negative binomial model accounting for macro-level and micro-level factors to study accident counts at road intersections. Several covariates constructed at both levels of spatial resolution were found to be associated with more accidents at intersections. Cai et al. (2018) implemented a grouped random parameters multivariate spatial model at two levels, segments and intersections. Covariates were defined separately over segments, intersections and wider zones (allowing the inclusion of covariates, such as socio-economic characteristics, at a lower spatial resolution). Zhao et al. (2018) used multivariate Poisson log-normal and zero-inflated univariate and multivariate Poisson models to study accident frequency (by severity level) at signalized intersections, consisting of the road segments at 200 ft upstream from the signal controlling the intersection. Lastly, Alarifi et al. (2018) proposed the use of a multivariate hierarchical Poisson lognormal model that accounts for the spatial relationships between road segments and intersections located along the same corridor. Average annual daily traffic variables at roadway segments and intersections, absolute speed limit difference between a major and a minor road meeting at an intersection, and driveway density showed positive associations with the number of traffic accidents.

With regard to the distance threshold of 200 ft chosen by Zhao et al. (2018), it needs to be highlighted that the definition of intersection-related traffic accidents presents a low level of agreement. For instance, Miaou and Lord (2003) considered a distance of 15 m ($\simeq$50 ft) from intersection locations, Ye et al. (2009) 75 m ($\simeq$250 ft), Zhao et al. (2018) 60 m ($\simeq$200 ft) and Das et al. (2008) analyzed the range 0 to 60 m at increments of 15 m. Furthermore, besides the lack of agreement, intersection zones are not clearly defined in many of the papers in the field. Finally, it is also highly remarkable how several of these papers focused entirely on intersection entities alone, avoiding consideration of the road segments surrounding intersections. Although Lee et al. (2017) and Cai et al. (2018) considered zonal effects that were shared by close segments and intersections, only Alarifi et al. (2018) have conducted a unified consideration of road segments and intersections. In this regard, Miaou and Lord (2003) pointed out the advisability of modelling data taking all kind of road entities simultaneously, which according to these authors would include segments, intersections and ramps. The implicit assumption of independence between entities may lead to ignoring many important spatial relationships that are likely to exist between them.

This study presents a statistical modelling of traffic accidents at the road segment level. In order to represent true neighbouring relationships between road segments, a directed road network structure accounting for traffic flow has been used. In summary, we had three methodological objectives: to present and discuss some issues that arise when conducting a spatial analysis of traffic accidents located on a road network, to analyze traffic accidents at road intersections, including a specific strategy that draws together both road intersection and non-intersection zones along the network, and to combine the results produced by the two statistical approaches finally chosen, spatial count models and coldspot/hotspot detection, in order to achieve more complete conclusions regarding the effect of various road characteristics on the occurrence of traffic accidents for the road network of interest.

The rest of the paper is structured as follows. The next section contains a complete description of the data employed for the analysis, including the traffic accident dataset recorded during the period of study and the network structure that represents the underlying space where these accidents occurred. This is followed by a methodological section that provides a description of the procedure followed to include the consideration of intersection zones, the definition of spatial neighbourhoods between road segments of the network, the specification of the spatial count models used to fit the data, the methods employed to assess the performance of such models, the definition of one class of network-constrained kernel density estimation and the procedure applied to locate zones of high and low risk along the network. Finally, there is a discussion of the performance and implications of the methods applied.

## 2. Data

### 2.1. Accident information

A total of 5738 traffic accidents recorded by the Local Police Department of the city of Valencia (Spain) during the years 2005 to 2017 in the Eixample District of the city were used. Each of these accidents was geocoded from the address information recorded by the Police minutes after the accident had occurred. Once the coordinates of each accident were obtained, these were projected onto a linear network representing the traffic streets of the Eixample District of Valencia. This two-stage process was supervised by all the authors in order to ensure a high level of accuracy.

### 2.2. Network structure

A linear network composed of 279 vertices and 444 road segments, representing a total length of 33.57 km, was used for the analysis. The vertices where more than two segments meet correspond to road intersections, which were 227 in the case of this network. Fig. 1a contains a map (Graul, 2016; OpenStreetMap contributors, 2017) that shows the zone of the city of Valencia where the road network of interest is located. Some parts of this network were previously simplified without altering its basic geometrical structure in order to reduce the number of short road segments which could hinder the subsequent modelling of the data. Moreover, network preprocessing included the slight modification of highly complex intersections and the removal of pedestrian streets, which were performed with the SpNetPrep R package (Briz-Redón, 2019).

In addition, for the purpose of improving the analysis, the network was given directionality according to the traffic flow of this district of Valencia as of the end of December 2017 (see Fig. 1b). Some of the road segments of the network were defined as bidirectional, representing two-way streets present in the district where no median strip separates the two flows of vehicles. However, bidirectional road segments were only 5% of the total, a fact that completely justifies the definition of traffic flow directionality along the network. In addition, for road segments divided by a median strip two (parallel) road segments were available in the network at a distance proportional to the width of the strip.

Finally, the possible changes in direction of traffic that could have been made during the period of years considered have not been taken into account due to the difficulty of tracking them. However, as Eixample District is very close to the centre of Valencia and is part of a very well-established area of the city, it can be assumed that changes of traffic direction must have been minimal in the period 2005–2017.

### 2.3. Network-related covariates

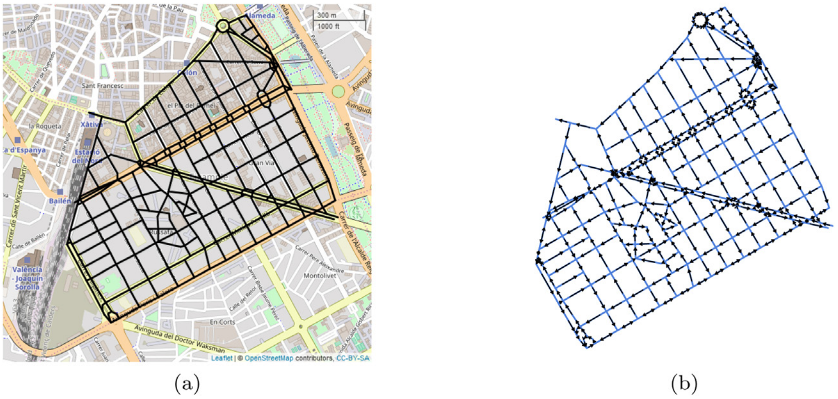
Several factors that could be associated with vehicle collisions are considered at the road segment level. These mainly include the presence of specific public services in the road segment (parking slots, traffic lights and bus stops) and basic characteristics of the roads that the links in the network represent. The latter include the number of lanes in the road, the presence of a bus lane (binary), the type of road (main or not, binary), the number of roads that directly connect to each road segment of the network, distinguishing whether they allow traffic to enter or leave it, a categorical covariate representing average annual daily traffic (AADT) and a categorical covariate assigning a geometric typology to each road segment (this one is described in the Methodology). In this regard, we should note that AADT is not available for every road segment in Valencia, but only for the most travelled avenues and streets. Hence, the data available was used to define a 5-level categorical covariate representing the following ranges for AADT: < 7000 (level 1), 7000–16000 (level 2), 16000–25000 (level 3), 25000–55000 (level 4) and > 55000 (level 5). These ranges represent the least travelled road segments of the city for which scarce data is available (level 1) and the four quartile-based intervals that follow from the available AADT values (levels 2 to 5). It is worth noting that the strategy of categorizing AADT values has already been tested by several authors (Hao and Daniel, 2014; Fan et al., 2015; Yasmin et al., 2016).

Furthermore, numbers of lanes and neighbouring roads (referred to here as neighbours) were truncated and recoded for values higher than 5 and 3, respectively. To obtain the number of neighbours the network that was considered was actually an extension of the final one employed for the analysis, in order to avoid an unrealistic low number of neighbours for the road segments at the edge of the network. Finally, as the network of study represents a fairly small and homogeneous population area, we concluded that the inclusion of demographic or socioeconomic variables was not of interest. Table 1 includes a description and statistical summary of the covariates introduced in this section.

## 3. Methodology

### 3.1. Software

The R programming language (3.4.1 version, R Development Core Team, Vienna, Austria) (R Core Team, 2017) was used to obtain all the results presented in this study. The R packages bayesplot (Gabry and



(a)

(b)

**Fig. 1.** Road structure of study displayed over a map of the city of Valencia (a) and its representation as a linear network made of links and vertices, with arrows indicating traffic flow directionality (b).

Mahr, 2018), brms (Bürkner et al., 2017), ggmap (Kahle and Wickham, 2013), spatstat (Baddeley et al., 2015), spded (Bivand and Piras, 2015) and SpNetPrep (Briz-Redón, 2019) were specifically required for performing the analysis and the data curation process.

### *3.2. Definition of intersection zones*

In order to capture the differential risk between road locations around intersections and road segments between them, the original network structure was modified by creating shorter road segments in the proximity of each road intersection. The insertVertices function of the R package spatstat (Baddeley et al., 2015) was key for performing this task.

Specifically, road segments of 20 meters were inserted around intersection neighbourhoods (so that the furthest point of the segment from the intersection was at a distance of 20 m), which were determined to be intersection analysis zones (IAZs). On the other hand, segments not satisfying this condition, most of which are between two IAZs, were declared as middle analysis zones (MAZs). Thus, the original network of study was divided into 683 IAZs and 292 MAZs, leading to the formation of a new road network (referred to from now on as a split network) made up of 810 vertices and 975 road segments (the original had 279 vertices and 444 road segments). As an illustration, Fig. 2 displays the distribution of IAZ and MAZ along the split network.

Therefore, the definition provided for IAZ and MAZ allowed for the coexistence of street zones subject to different rules and causalities while being represented by a unique geometrical entity: the road segment (note that the sum of the number of IAZ and MAZ coincides with the number of road segments of the final network). This fact led to a unified definition of neighbourhood relationships and covariates for the two types of zone that mainly arise when dealing with traffic accident datasets. Indeed, the term segment is used without distinction for both types of zone throughout the paper, even though in related literature it is only used for what it has been defined as MAZs.

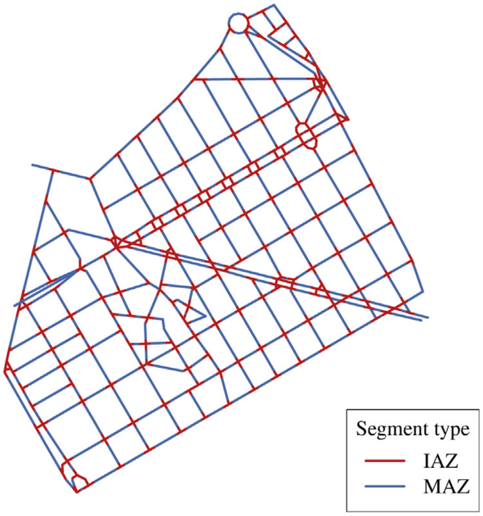


**Fig. 2.** Graphical description of the split network showing the locations of IAZs and MAZs.

The choice of a distance of 20 meters was mainly based on knowledge of the road network of study and on similar distances used in literature (Miaou and Lord, 2003). Indeed, this distance allows a fair representation of IAZs as intersection-approaching or intersection-leaving segments. The selection of a shorter threshold distance was rejected due to the lack of sufficient certainty on the data collection procedure to guarantee the correct location of accidents at such a level of resolution around intersections. Furthermore, the objective was to employ the road segment as the only spatial unit of study, and this would be undermined if a threshold very close to 0 were chosen (as the IAZ would almost become the intersection point itself).

Regarding the definition of the covariates at the level of the new

**Table 1**
Variables description and basic statistics, where SD denotes the standard deviation.

| Variable | Description | Mean | SD |
|---|---|---|---|
| Main road | Main road segment of the city (binary) | 0.626 | 0.484 |
| Parking slots | Existence of public parking slots in the road segment (binary) | 0.613 | 0.671 |
| Traffic light | Presence of a traffic light in the road segment (binary) | 0.617 | 0.487 |
| Bus stops | Existence of public bus stops in the road segment (binary) | 0.110 | 0.314 |
| Bus lane | Presence of a bus lane in the road segment (binary) | 0.572 | 0.495 |
| No. of lanes | Number of traffic lanes in the road segment | 2.176 | 1.349 |
| No. of in-neighbours | Number of neighbouring road segments allowing traffic to enter the road segment | 1.770 | 0.631 |
| No. of out-neighbours | Number of neighbouring road segments allowing traffic to leave the road segment | 1.775 | 0.629 |
| AADT | Average annual daily traffic (5 levels) | 2.286 | 1.539 |

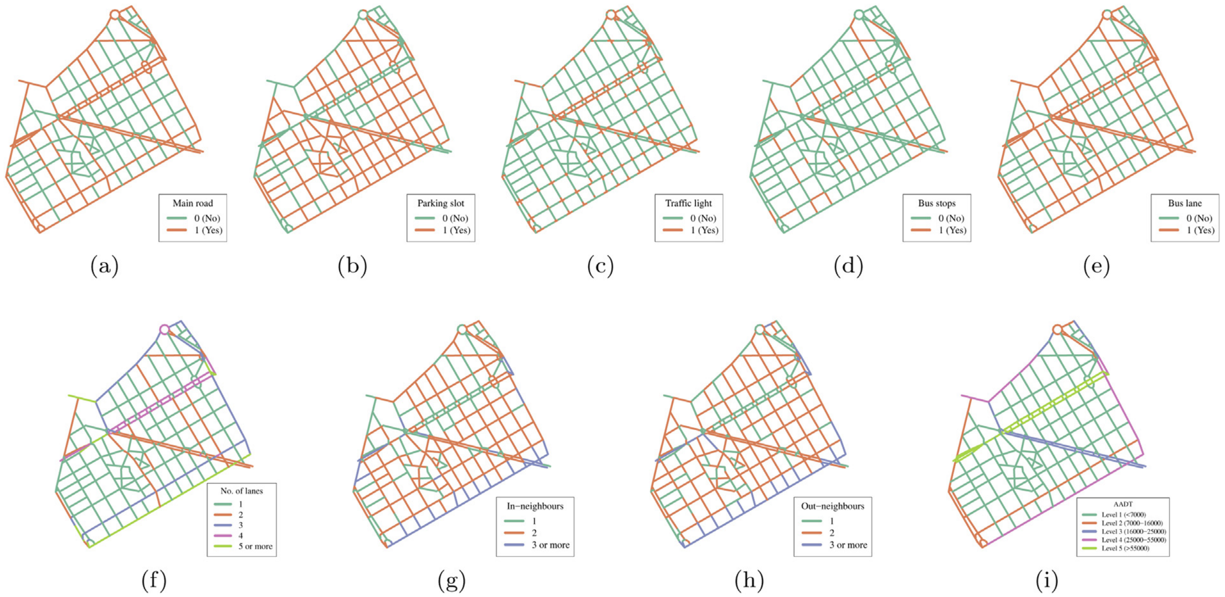


**Fig. 3.** Graphical description at the road segment level of the following network-related variables: (a) main road indicator, (b) parking slot presence, (c) traffic light presence, (d) bus stop presence, (e) bus lane presence, (f) number of lanes, (g) number of in-neighbours (h) number of out-neighbours and (i) AADT.

split road network, these simply follow the values available for the original network. Hence, each IAZ or MAZ of the split road network acquires the value (for a given covariate) of the corresponding whole road segment in the original non-split network. An exception was made with traffic lights, given their frequent location around road intersection zones. For this reason, a value of 1 was assigned to an IAZ or MAZ for the indicator related to traffic light presence if and only if a traffic light was present in the same road segment (before splitting) at a distance lower than 20 m from the middle point of the IAZ/MAZ. As an illustration, Fig. 3 provides a graphical description of every covariate considered for the analysis, which enables us to appreciate the distinction made for traffic lights (Fig. 3c).

### 3.3. Concept of neighbourhoods

The road segments that form the already defined directed network structure constitute the basic spatial units on which to perform the statistical analysis. Given a road segment, $i$, in the directed linear network, its neighbourhood, $N(i)$, can be defined in four different ways depending on whether the traffic flow information available is used. At the simplest level, if this information is not used, two road segments $i$ and $j$ are neighbours if they are connected by a vertex of the network. However, the use of the traffic flow leads to the definition of three other types of neighbourhoods. First, neighbourhood between $i$ and $j$ can be established if it is possible to travel from $i$ to $j$ or from $j$ to $i$, in either direction, without passing through another road segment of the network; this is denoted $N_{dir}(i)$. In addition, if a distinction is made between travelling from $i$ to $j$ or vice versa, it is possible to separate the neighbouring road segments that allow you to reach $i$ ($N_{\text{dir}}^{\text{in}}(i)$) from those that allow you to leave from $i$ to another road segment of the network ($N_{\text{dir}}^{\text{out}}(i)$) (see Fig. 4 for examples of all these types of neighbourhood). From now on these last two types of neighbours are referred to as in-neighbours and out-neighbours, respectively.

The four definitions of neighbourhood structures can lead to the construction of four different adjacency matrices. Thus, a $W_{dir}$ matrix based on $N_{dir}$ neighbourhoods was the only one employed as it was considered the most suitable for the goals established. Regarding this matrix, its entries, $w_{\text{ij}}$, are called weights and it holds that $w_{\text{ij}} = 1/|N(i)|$, if $j \in N(i)$ (row normalization), and 0 otherwise.

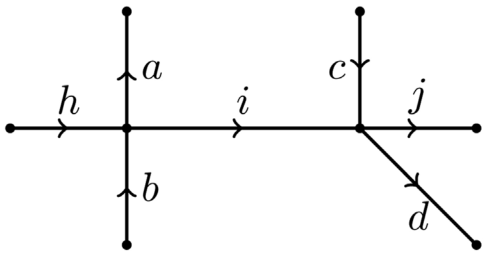


**Fig. 4.** Examples of types of neighbourhood in a directed linear network. The six road segments that are contiguous to road segment $i$ allow the construction of the neighbourhoods $N(i) = \{a, b, c, d, h, j\}$, $N_{dir}(i) = \{b, d, h, j\}$, $N_{\text{dir}}^{\text{in}}(i) = \{b, h\}$ and $N_{\text{dir}}^{\text{out}}(i) = \{d, j\}$.

### 3.4. Road segment neighbourhood geometry

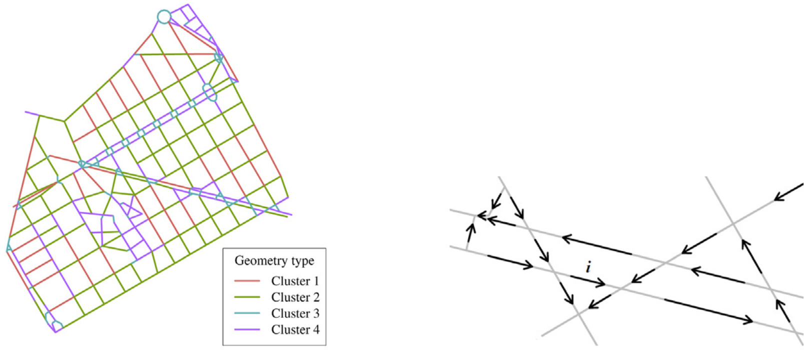
The geometric structure surrounding each road segment of the network was studied, a procedure made possible by the road network structure. Given a road segment of the network, the factors considered for each neighbouring road segment were the neighbourhood type (in or out) and the angles formed between the road segment and its neighbours. Road segment length and the number of in and out neighbours were also included to better discriminate between road segments. As was done with the other covariates previously defined (with the exception of the indicator factor for traffic lights), the geometry is studied from the perspective of the original network. Later, the values obtained are assigned to the road segments of the split network accordingly.

A total of six types of neighbours were defined by combining the angles of the road segments and the direction of the traffic. Angles between road segments were classified (measured in [0°, 180°]) into three groups: straight (]150°, 180°]), right (]60°, 120°[) and sharp ([0°, 60°]$\bigcup$[120°, 150°]). Each of these types of angle was then crossed with the in/out information associated with each neighbouring road segment to create the six possible scenarios.

The same strategy was followed with the lengths of the neighbouring road segments. In this case, the road segments were divided into three groups (short, medium and long) according to the 33.33% and 66.67% quantiles of the road segment length distribution. Again, the three groups created were crossed with the in/out information,

**Table 2**
Mean values of the variables used to perform the clustering of the road segments according to their geometry.

| Cluster | Straight$_{in}$ | Right$_{in}$ | Sharp$_{in}$ | Straight$_{out}$ | Right$_{out}$ | Sharp$_{out}$ | Short$_{in}$ | Medium$_{in}$ | Long$_{in}$ | Short$_{out}$ | Medium$_{out}$ | Long$_{out}$ | Length | $\lvert N_{\text{dir}}^{\text{in}}\rvert$ | $\lvert N_{\text{dir}}^{\text{out}}\rvert$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0.45 | 0.81 | 0.48 | 0.40 | 0.74 | 0.74 | 0.21 | 0.64 | 0.88 | 0.29 | 0.69 | 0.90 | 170.66 | 1.83 | 1.98 |
| 2 | 0.45 | 0.93 | 0.63 | 0.42 | 0.98 | 0.59 | 0.28 | 0.58 | 1.15 | 0.27 | 0.57 | 1.15 | 115.84 | 2.17 | 2.11 |
| 3 | 0.21 | 0.46 | 0.88 | 0.21 | 0.52 | 0.86 | 0.75 | 0.48 | 0.31 | 0.74 | 0.50 | 0.34 | 23.80 | 1.58 | 1.61 |
| 4 | 0.25 | 0.59 | 0.84 | 0.29 | 0.53 | 0.78 | 0.53 | 0.75 | 0.39 | 0.52 | 0.77 | 0.31 | 68.40 | 1.77 | 1.78 |



(a)

(b)

**Fig. 5.** (a) Clustering of the road segments of the spatial network according to their neighbourhood geometry. (b) Detailed neighbourhood of an road segment, $i$, of the directed network. Six road segments share a vertex with $i$, but only four of these allow traffic to flow from $i$ or to $i$. The values of the geometric variables for road segment $i$ are: Straight$_{in}$ = 1, Right$_{in}$ = 0, Sharp$_{in}$ = 1, Straight$_{out}$ = 1, Right$_{out}$ = 0, Sharp$_{out}$ = 1, Short$_{in}$ = 1, Medium$_{in}$ = 1, Long$_{in}$ = 0, Short$_{out}$ = 1, Medium$_{out}$ = 0, Long$_{out}$ = 1, Length = 41.7, $|N_{\text{dir}}^{\text{in}}(i)| = 2$, $|N_{\text{dir}}^{\text{out}}(i)| = 2$.

producing six new classification groups.

The $k$-means algorithm (Hartigan and Wong, 1979) was then applied to a total set of fifteen geometric-related variables for each of the road segments: the number of neighbours belonging to each of the six angle-direction and length-direction combinations, the road segment length and the number of in and out neighbours. A value of $k = 4$ was chosen, since convergence was not reached for higher values of $k$, and this made it possible to form four clusters of 42, 130, 163 and 109 road segments, respectively.

Table 2 summarizes the mean values for the variables employed in the clustering procedure and Fig. 5 includes the graphical representation of the four clusters and an example of construction of the geometric-related variables for a specific road segment. According to these results, Cluster 2 is mainly composed of medium-long road segments with a high average number of neighbouring road segments that form a right angle, which is associated with being part of a crossroads (90-degree intersection). Cluster 3 is formed by very short road segments, a high proportion of which involve acute angles, representing abrupt changes of direction in the directed network. Cluster 1 clearly presents the highest road segment length and an high number of neighbours. Finally, Cluster 4 is made up of short-medium length road segments and quite high connectivity with short-length road segments compared with Clusters 1 and 2.

### 3.5. *Accident count modelling*

A Bayesian spatial model with Zero-Inflated Negative Binomial response (ZINB) was implemented to fit the observed accident counts for the split network structure (composed of 975 road segments). If $Y\sim$ NB $(\mu, \psi)$ (basic negative binomial distribution of mean $\mu$ and shape $\psi$) then it holds that $E(Y) = \mu$, $V(Y) = \mu + \frac{\mu^2}{\psi}$ and $P(Y = x) = (\frac{x+\psi-1}{\psi-1})(\frac{\psi}{\mu+\psi})^{\psi}(\frac{\mu}{\mu+\psi})^{x}$. The zero-inflated version of the NB distribution acts as a double-stage process that makes it possible to increase the probability of value 0. Thus, if $z$ denotes the structural probability of 0 for the ZINB distribution, its probability mass satisfies the next stepwise function:

$$P(Z=0)=\begin{cases} z+(1-z)P(Y=0), & x=0\\ (1-z)P(Y=x), & x>0\end{cases}$$

where $Y\sim$ NB$(\mu, \psi)$ and $Z\sim$ ZINB$(\mu, \psi, z)$.

Then, on the basis of the choice of a ZINB distribution for the response (accident counts) the next spatial model (Model 1) was specified:

$$Y_i \sim \text{ZINB}(\mu_i, \psi, z)$$

$$\log(\mu_i) = \log(\text{Length}_i) + \mathbf{x_i}\boldsymbol{\beta} + \phi_i \quad \text{(Model 1)}$$

where $Y_i$ is the number of accidents observed at road segment $i$, $\mu_i$ and $\psi$ are the mean (for road segment $i$) and overdispersion (shape) values for the ZINB distribution, $z$ is the probability of value 0 for the ZINB distribution, the natural logarithm acts as a link function for the mean risk at segment $i$ ($\mu_i$), the natural logarithm of each segment's length is added as an offset, $\mathbf{x_i}$ is a vector that contains the values for the covariates described in Table 1 corresponding to segment $i$ along with a factor indicating whether the road segment belongs to the IAZ class, $\boldsymbol{\beta}$ is a vector of coefficients to control the effect of these predictors and $\phi_i$ represents a spatial effect for road segment $i$.

The spatial effect was modelled using a conditional autoregressive (CAR) structure (Besag, 1974; Besag et al., 1991):

$$\phi_i|\phi_j, j \neq i \sim N(\alpha\sum_{j=1}^{n} w_{\text{ij}}\phi_j, \tau_i^{-1})$$

where $\alpha \in [0, 1]$ is a spatial dependence parameter that measures the strength of spatial autocorrelation ($\alpha = 0$ reflects the complete absence of such effect), $\tau_i$ is a precision parameter that varies with $i$ and $w_{\text{ij}}$ the entry at the $(i, j)$ position of the neighbourhood matrix $W_{dir}$ ($w_{\text{ii}} = 0, \forall\, i$).

In particular, the joint distribution of $\Phi = (\phi_1, .. ., \phi_n)$ satisfies the Gaussian multivariate probability distribution (Banerjee et al., 2004):

$$\Phi \sim N(0, [\tau(D - \alpha W_{\text{dir}})]^{-1}).$$

where $D$ is a diagonal matrix that contains the number of in- and out-neighbours of each spatial unit.

Moreover, a second model (Model 2) specifically focused on the effect of the covariates on IAZs was implemented with the following linear predictor for $\log(\mu_i)$:

$$\log(\mu_i) = \log(\text{Length}_i) + \mathbf{x_i}\boldsymbol{\beta} + \mathbf{x_i}\boldsymbol{\gamma} I_{\text{IAZ}} + \phi_i \quad \text{(Model 2)}$$

where $I_{\text{IAZ}}$ is an indicator function for IAZ and γ represents the vector of coefficients that measure the effect of the covariates at IAZ. Hence, Model 1 only considers the effect of IAZ as one of the factors being studied, whereas Model 2 allows the determination of the differential effect that each covariate can produce at the road segment level depending on the zone of analysis (IAF or MAF). Furthermore, for these two models, the inflation probability, $z$, was modelled through a logit equation that makes it possible to estimate a different value of $z$ for each zone type:

$$\text{logit}(z) = z_{\text{Intercept}} + z_{\text{Slope}} I_{\text{IAZ}} \longleftrightarrow z = \frac{\exp(z_{\text{Intercept}} + z_{\text{Slope}} I_{\text{IAZ}})}{1 + \exp(z_{\text{Intercept}} + z_{\text{Slope}} I_{\text{IAZ}})} \quad (1)$$

where $I_{\text{IAZ}}$ is again an indicator function for IAZ.

The estimation of the parameters of the two models was performed with the brms R package (Bürkner et al., 2017), which is based on the statistical software Stan (Carpenter et al., 2017).

### 3.6. Model checks

Several techniques were applied in order to check for the propriety of the different models employed for representing the observed accident counts. In this section, the methods used for this task, which included conditional predictive ordinate (CPO), general correlation coefficients and Moran's $I$, are briefly described.

The CPO method (Stern and Cressie, 2000; Marshall and Spiegelhalter, 2003) requires data simulation from the posterior distribution of a fitted Bayesian model. Indeed, if the values for the covariates of the models are left fixed as in the data used to fit the model, the accident counts simulated at each draw behave like replicates of the original counts ($y$) and are denoted by $Y^{rep}$ (Gelman et al., 2013). If a model represents the counts properly, the observed counts should agree with the distribution of a simulated dataset of $Y^{rep}$. Then, a high departure between $y$ and $Y^{rep}$ may indicate a poor performance from the model. In this regard, CPO is a simulation-based tool that has already been used in similar research studies (Yang et al., 2013; Xie et al., 2014) with the main purpose of identifying outliers within the data, which in this case correspond to road segments. For this purpose, the distribution of $Y_i^{\text{rep}}$ for every spatial unit (road segment) $i$ is evaluated from all the original data except $y_i$ itself (in a similar way to the leave-one-out cross-validation procedure). Thus, the goal is to find spatial units whose observed count value is far enough from the simulated distribution of $Y_i^{\text{rep}}|y_{-i}$, where $y_{-i}$ denotes the original data with the exclusion of $y_i$. The determination of a $p$-value that tests this question for unit $i$ is done through a reweighting of the $Y^{rep}$ with the choice of the following weight:

$$\rho_{-i}^{(k)} = \frac{1}{P(y_i|\Lambda^{(k)})}$$

where $k$ is the index for the simulation, $y_i$ is the number of accidents observed for spatial unit $i$, $\Lambda^{(k)}$ represents the parameters sampled for the model at simulation number $k$ (which includes the corresponding values for $\lambda_i$'s, $\psi$, $z$ and $\Phi$) and $P$ represents the probability function of a ZINB distribution that follows the parameters in $\Lambda^{(k)}$. Then, a $p$-value that allows outlier identification is approximated with the next expression (Marshall and Spiegelhalter, 2003):

$$P(Y_i^{\text{rep}} \le y_i|y_{-i}) \approx \sum_{g=0}^{y_i-1} \frac{\sum_{k=1}^{K} P(Y_i^{\text{rep}} = g|\Lambda^{(k)})\rho_{-i}^{(k)}}{\sum_{k=1}^{K} \rho_{-i}^{(k)}} + \frac{1}{2}\frac{\sum_{k=1}^{K} P(Y_i^{\text{rep}} = y_i|\Lambda^{(k)})\rho_{-i}^{(k)}}{\sum_{k=1}^{K} \rho_{-i}^{(k)}} \quad (2)$$

General correlation coefficients are an extension of Pearson's correlation coefficient (Pearson, 1896) making it possible to compare two possibly related numerical vectors of the same length. Specifically, it can be employed to compare the distribution of ranks shown by the observed accident counts and the counts fitted by any statistical model applied. The formula for a general correlation coefficient, $\Gamma$, is:

$$\Gamma = \frac{\sum_{i=1,j=1}^{n} a_{ij} b_{ij}}{\sqrt{\sum_{i=1,j=1}^{n} a_{ij}^2 \sum_{i=1,j=1}^{n} b_{ij}^2}}$$

where the coefficients $a_{ij}$ and $b_{ij}$ must be anti-symmetric ($a_{ij} = -a_{ji}$, $b_{ij} = -b_{ji}$). As two important particular cases, if $r^{\text{obs}}$ and $r^{\text{exp}}$ denote the ranks (in decreasing order) of the observed and fitted accident counts per spatial unit (respectively), the following choices of $a_{ij}$ and $b_{ij}$ correspond to Kendall and Spearman correlation coefficients (Kendall, 1938; Spearman, 1904):

$$a_{ij} = \text{sgn}(r_i^{\text{obs}} - r_j^{\text{obs}}),\ \ b_{ij} = \text{sgn}(r_i^{\text{exp}} - r_j^{\text{exp}})$$

$$a_{ij} = r_i^{\text{obs}} - r_j^{\text{obs}},\ \ b_{ij} = r_i^{\text{exp}} - r_j^{\text{exp}}$$

where $\text{sgn}(x) = x/|x|$ (sign function). A high value of $\Gamma$, regardless of the specific selection of $a_{ij}$ and $b_{ij}$, indicates a high level of agreement between the ranked observed accident counts and the ones predicted by a model. This is a clear sign of a good model fit.

Finally, Moran's $I$ (Moran, 1950a,b) consists in a global estimation of the spatial autocorrelation of a variable indexed in according to a system of spatial units. Its definition is the following:

$$I = \frac{\sum_{i=1}^{n} \sum_{j \in N_{\text{dir}}(i)} \frac{1}{n_i}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

where $x_i$ is a variable indexed by spatial unit and $\bar{x}$ its average. Thus, Moran's $I$ makes use of the predefined neighbourhood structure and behaves as a correlation between the variable of interest and a variable that assigns to each of the spatial units a weighted average of the values of its neighbours. Under the hypothesis of no spatial autocorrelation, it holds that $E(I) = -1/(n-1)$, where $n$ is the number of spatial units (975). Hence, negative Moran's $I$ values for the residuals of the model would be an indicator of a good performance from the fitted Bayesian count models in capturing the spatial effect.

### 3.7. Kernel Density Estimation

Kernel Density Estimation (KDE) is commonly used to obtain the intensity of a point pattern that lies on a space. Particularly, it can be used to estimate the intensity of a point pattern along a linear network, requiring modifications of the classical formulas (valid for areal units) to account for the particularities of this spatial structure (Okabe et al., 2009; Okabe and Sugihara, 2012). In this study, the equal-split continuous kernel density defined by McSwiggan et al. (2017) is computed at the middle point of every road segment $i$ of the linear network following the next equation, by which the $f_\sigma(i)$ values are obtained:

$$f_\sigma(i) = \sum_{x \in A(m_i,\sigma)} k(d_L(x, m_i))a^C(\pi) \quad (3)$$

where $m_i$ is the middle point of the road segment $i$ of the linear network, $\sigma$ is the kernel's bandwidth, $A(m_i, \sigma)$ is the set of points of the network at a distance from $m_i$ up to $\sigma$ where an accident took place, $k(u) = \frac{1}{\sigma\sqrt{\pi}} e^{(\frac{-u}{\sigma})^2}$ is the kernel function (Gaussian), $d_L$ is the distance along the network and $a^C(\pi) = \prod_{j=1}^{m} \frac{2}{\deg(v_j)}$, where $\pi = [v_1, \ldots, v_m]$ denotes the set of vertices of the network that have to be passed through to travel the shortest path that joins $m_i$ with $x$ and deg() represents the degree of a vertex of the network, meaning the number of road segments incident to the vertex. It needs to be remarked that the computation of the distance $d_L$ between any two points of the network makes use of its directed structure, providing a realistic measure of the distance between the two points according to traffic flow.

A Gaussian kernel was selected because it is the most common option, and no other kernel functions were explored because this choice usually has little effect on the results (Silverman, 2018). On the choice of the bandwidth parameter, a value of around $\sigma = 50$ m would be optimal if the non-parametric test proposed by Cronie and Van Lieshout

**Table 3**
Summary of the response (counts at the road segment level) for the original and the split road network.

| | Original network | Split network | |
|---|---|---|---|
| | | IAZs | MAZs |
| Mean | 12.92 | 5.75 | 6.21 |
| Variance | 403.46 | 189.94 | 123.25 |
| Variance/Mean | 31.22 | 33.06 | 19.84 |
| % Zeros | 23.20 | 49.19 | 18.84 |
| Gini Index | 0.67 | 0.80 | 0.66 |
| No. of accidents | 5738 | 3924 | 1814 |
| No. of segments | 444 | 683 | 292 |
| Road length (m) | 33571.09 | 14166.96 | 19404.13 |

(2018) were followed. However, the larger value of $\sigma = 100$ m was applied in agreement with previous studies on road networks that have employed KDE for hotspot detection (Xie and Yan, 2013; Nie et al., 2015).

Finally, edge effects (Okabe and Sugihara, 2012) need to be discussed because the network used for the analysis is to a certain extent artificially bounded, being only a part of the larger road network of Valencia. First, it needs to be remarked that the kernel construction chosen (Equation (3)) alleviates edge effects, as stated by McSwiggan et al. (2017). Second, Eixample District is delimited by pedestrian and secondary roads (to the north and south), a train station (to the west) and a green area (to the east), which to some extent make the district naturally bounded (Fig. 1a enables us to appreciate some of these points). Furthermore, the two roads bordering the network of analysis to the north and south are important avenues of Valencia which account for most of the accidents in the vicinity (these avenues are part of the network analyzed). All these facts allow us to conclude that accident densities estimated along the four roads that form the border of the network are reasonable.

### 3.8. Coldspot/hotspot detection

The use of the count models was supplemented with a search for zones of the network with a particularly low or high incidence of traffic accidents; these are usually known as coldspots and hotspots, respectively. Several approaches to this problem coexist in recent literature on traffic accident data, including some of the studies already mentioned in the Introduction (Xie and Yan, 2013; Nie et al., 2015; Thakali et al., 2015; Harirforoush and Bellalite, 2016). These methods mainly agree in the use of KDE to obtain a smooth representation of the observed point pattern, a process which is commonly followed by the detection of zones of the network whose KDE values present a significant spatial autocorrelation.

Here, KDE was computed with $\sigma = 100$ m at the middle point of each road segment of the network considering the $d_L$ distance along the network that accounts for traffic flow (following Equation (3)). Then, the local version of Moran's $I$ statistic known as LISA (Anselin, 1995) was obtained for each road segment following the next formula:

$$I_i = \frac{x_i - \bar{x}}{\sum_i (x_i - \bar{x})^2/n} \sum_j w_{ij}(x_j - \bar{x})$$

The road segments showing a significant local association (a threshold of 0.1 was used for the $p$-value instead of the usual 0.05 to minimally extend some of the coldspots/hotspots, allowing a wider part of the network to be analyzed) were selected and grouped according to their contiguity. Other inputs such as the accident count per road segment or the accident rates were also considered for computing the LISA values, but KDE was the only one capable of providing a sensible number of zones along the network presenting similar behaviour in terms of dangerousness. Finally, the basic average intensity of the point pattern (number of events per unit length) in each of the zones of interest was compared with the mean intensity in its first-order neighbourhood (the set of all the first-order neighbours of the road segments composing the zone), which made it possible to detect both low-intensity (coldspot) and high-intensity (hotspot) parts of the network showing a differential incidence of traffic accidents in comparison with the road segments in their surrounding areas.

Finally, once the coldspots and hotspots had been located in the network, the values of the covariates of the road segments that formed them were individually analyzed to confirm or put into question the conclusions that could be drawn from the use of the count models.

## 4. Results and discussion

A total of four Monte Carlo Markov chains (MCMC) of length 30000 were run for the two models starting from non-informative priors for the parameters involved. The length of the chains was chosen to be large enough to ensure the convergence of the estimates of all the parameters involved in the models, which was afterwards checked using common validation tools (scale reduction factor close to 1 for all estimates and visual inspection of the chains). The choice of a ZINB model is sensible according to the values in Table 3, which was validated through subsequent predictive checks of a graphical nature. First, accident counts are clearly overdispersed with a variance-to-mean ratio of 31.22 for the original network. Second, 23.2% of the road segments (103 of the 444 road segments that form the original network) have no accidents recorded for the period being considered, which leads to the choice of a zero-inflated response. Furthermore, the inequality observed in the accident counts per road segment leads to a Gini index (Gini, 1912) of 0.67, with more than the 50% of the accidents recorded concentrated in 52 of the segments of the original network (these segments represent only 15% of the length of the network), in agreement with previous studies of a similar nature (although not focused on traffic accidents) referring to the law of crime concentration (Weisburd, 2015). These particularities of the data of study are also present when the network is split into IAZs and MAZs. However, whereas the variance-to-mean ratio is not so heavily affected, the number of zeros is much higher in IAZs. For this reason, the estimation of the zero-inflated probability was made dependent on the zone, as described in the previous methodological section regarding count model specifications.

Table 4 displays the values obtained for the three kinds of validation tools that were applied. Moran's $I$ values were negative for both models, which is a sign of good performance as it indicates the absence of spatial autocorrelation between model residuals. Correlation coefficients ($\Gamma$) derived from the comparison of observed and expected counts at the road segment were very similar and significantly greater than 0 for both models, although a slight improvement can be appreciated for Model 1. Regarding the percentage of potential outliers according to the CPO method (IAZs and MAZs showing a $p$-value lower than 0.05 according to Equation (2)), the results are again very close, but better again for Model 1. In conclusion, Model 1 presents better results than Model 2 by a narrow margin, but both models offer a reasonable basis to allow conclusions to be drawn from them regarding the occurrence of traffic accidents in the network of analysis.

Therefore, let us now concentrate on model parameters and on the effects that the covariates being considered could have had on the

**Table 4**
Values obtained for the statistical tools employed for model comparison.

| | Model 1 | Model 2 |
|---|---|---|
| $\Gamma$ (Kendall) | 0.37 | 0.34 |
| $\Gamma$ (Spearman) | 0.47 | 0.43 |
| % Potential outliers | 13.85 | 15.28 |
| Moran's $I$ | −0.06 | −0.04 |

**Table 5**
Summary of the results obtained with Models 1 and 2. Coefficient estimates ($\beta$ and $\gamma$) in bold represent covariates significant with 90% credibility, whereas Lo and Up denote the lower and upper bounds (respectively) of the 90% credible intervals for such estimates.

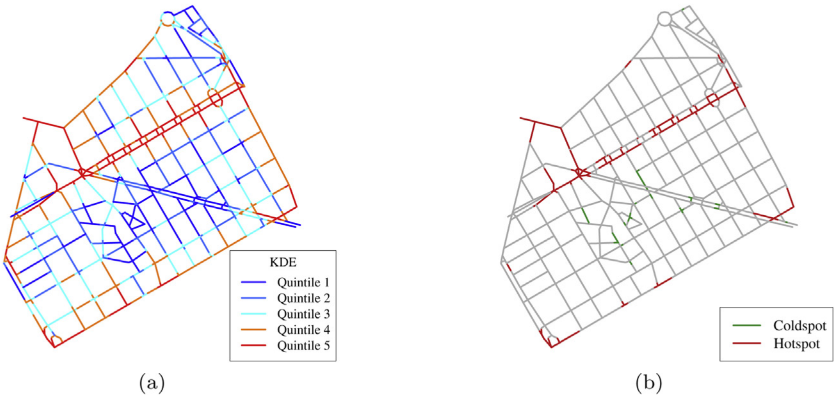
| | Model 1 | | | Model 2 | | |
|---|---|---|---|---|---|---|
| **Covariate** | $\beta$ | Lo | Up | $\beta$ | Lo | Up |
| (Intercept) | **−3.01** | −3.42 | −2.60 | **−3.43** | −3.99 | −2.88 |
| Main road | **0.41** | 0.04 | 0.78 | 0.50 | −0.09 | 1.09 |
| Parking slots | **−0.27** | −0.51 | −0.02 | −0.14 | −0.49 | 0.20 |
| Traffic light | −0.08 | −0.30 | 0.14 | −0.20 | −0.68 | 0.28 |
| Bus stops | 0.02 | −0.25 | 0.30 | 0.21 | −0.19 | 0.60 |
| Bus lane | **0.36** | 0.04 | 0.68 | 0.39 | −0.09 | 0.88 |
| No. of lanes (2) | −0.29 | −0.64 | 0.05 | 0.01 | −0.50 | 0.53 |
| No. of lanes (3) | 0.20 | −0.23 | 0.65 | 0.00 | −0.67 | 0.68 |
| No. of lanes (4) | 0.36 | −0.22 | 0.94 | −0.06 | −1.02 | 0.89 |
| No. of lanes (≥ 5) | −0.04 | −0.64 | 0.56 | 0.01 | −0.87 | 0.90 |
| No. of in-neighbours (2) | 0.05 | −0.16 | 0.27 | 0.03 | −0.31 | 0.36 |
| No. of in-neighbours (≥ 3) | 0.05 | −0.38 | 0.48 | −0.07 | −0.70 | 0.56 |
| No. of out-neighbours (2) | 0.04 | −0.19 | 0.26 | 0.03 | −0.31 | 0.37 |
| No. of out-neighbours (≥ 3) | 0.29 | −0.15 | 0.73 | −0.04 | −0.71 | 0.63 |
| Cluster (2) | 0.06 | −0.23 | 0.35 | 0.04 | −0.35 | 0.44 |
| Cluster (3) | **−0.43** | −0.81 | −0.06 | **−1.24** | −2.25 | −0.25 |
| Cluster (4) | **−0.59** | −0.91 | −0.28 | 0.05 | −0.41 | 0.49 |
| AADT (2) | 0.10 | −0.32 | 0.51 | 0.11 | −0.52 | 0.74 |
| AADT (3) | −0.15 | −0.58 | 0.29 | −0.26 | −0.94 | 0.42 |
| AADT (4) | 0.28 | −0.31 | 0.88 | **1.22** | 0.32 | 2.13 |
| AADT (5) | 0.48 | −0.10 | 1.06 | **1.34** | 0.36 | 2.34 |
| IAZ | **1.58** | 1.35 | 1.79 | **2.43** | 1.67 | 3.18 |
| **Covariate\|IAZ** | $\gamma$ | Lo | Up | $\gamma$ | Lo | Up |
| Main road\|IAZ | - | - | - | −0.27 | −1.01 | 0.46 |
| Parking slots\|IAZ | - | - | - | −0.22 | −0.68 | 0.25 |
| Traffic light\|IAZ | - | - | - | 0.07 | −0.47 | 0.61 |
| Bus stops\|IAZ | - | - | - | −0.30 | −0.83 | 0.23 |
| Bus lane\|IAZ | - | - | - | −0.04 | −0.66 | 0.58 |
| No. of lanes (2)\|IAZ | - | - | - | −0.33 | −1.01 | 0.34 |
| No. of lanes (3)\|IAZ | - | - | - | 0.42 | −0.43 | 1.28 |
| No. of lanes (4)\|IAZ | - | - | - | 1.05 | −0.12 | 2.24 |
| No. of lanes (≥ 5)\|IAZ | - | - | - | −0.07 | −1.21 | 1.08 |
| No. of in-neighbours (2)\|IAZ | - | - | - | 0.09 | −0.33 | 0.51 |
| No. of in-neighbours (≥ 3)\|IAZ | - | - | - | 0.19 | −0.64 | 1.01 |
| No. of out-neighbours (2)\|IAZ | - | - | - | 0.01 | −0.42 | 0.44 |
| No. of out-neighbours (≥ 3)\|IAZ | - | - | - | 0.70 | −0.16 | 1.57 |
| Cluster (2)\|IAZ | - | - | - | 0.00 | −0.54 | 0.55 |
| Cluster (3)\|IAZ | - | - | - | 0.68 | −0.40 | 1.79 |
| Cluster (4)\|IAZ | - | - | - | **−1.19** | −1.79 | −0.59 |
| AADT (2)\|IAZ | - | - | - | 0.03 | −0.76 | 0.82 |
| AADT (3)\|IAZ | - | - | - | 0.04 | −0.81 | 0.89 |
| AADT (4)\|IAZ | - | - | - | **−1.76** | −2.91 | −0.61 |
| AADT (5)\|IAZ | - | - | - | **−1.55** | −2.76 | −0.36 |
| **Parameter** | Est. | Lo | Up | Est. | Lo | Up |
| $\psi$ | **1.71** | 1.06 | 2.65 | **1.56** | 1.07 | 2.33 |
| $z_{\text{Intercept}}$ | **−8.69** | −15.87 | −4.68 | **−8.61** | −15.63 | −4.63 |
| $z_{\text{Slope}}$ | **8.32** | 4.30 | 15.49 | **8.27** | 4.28 | 15.30 |
| $\alpha$ | **0.11** | 0.01 | 0.32 | **0.17** | 0.01 | 0.47 |

accidents that occurred in the network of study during the period 2005–2017. Table 5 shows the results for Model 1 and Model 2, in which the missing levels of any covariate are implicitly present as they are considered the reference levels for the covariate (the other levels are estimated in relation to the missing one). If the estimation for the coefficient corresponding to a covariate ($\beta$'s and $\gamma$'s) or a structural parameter ($\psi$, $z$ and $\alpha$) lies in the 90% credible interval (all derived from the MCMC procedure), then the effect of that covariate or structural parameter is significant with 90% credibility. All structural parameters were found to be significant in all models. Hence, a modelling approach that includes spatial heterogeneity ($\alpha > 0$), overdispersion ($\psi > 0$) and a zero-inflated distribution that depends on the zone (MAZ or IAZ) is justified. Parameters $\psi$ and $z$ driving the ZINB distribution present very similar estimates for the two models. The higher percentage of zeros in IAZs is clear from the estimates obtained for the slope parameter ($z_{\text{Slope}} > 8$), which models $z$ through a logit equation. This particularly means (for instance, for Model 1) that $z = \frac{\exp(-8.69+8.32)}{1+\exp(-8.69+8.32)} \simeq 0.41$ in IAZs (following Equation (1)), a value that is not surprising in view of Table 3. On the other hand, $z = \frac{\exp(-8.69)}{1+\exp(-8.69)} < 2\cdot10^{-4}$ indicates that zero-inflation is not needed for modelling accident counts in MAZs, that is, a non-modified NB distribution would be suitable enough.

With regard to covariate effects, Model 1 indicates that main roads, roads containing a bus lane and approaching-intersection segments (IAZs) are associated with a higher accident count. In contrast, the existence of parking slots in the road and geometries of type 3 and 4 correlate with fewer traffic accidents at the road segment level for the network of study. Regarding these two geometry types, as mentioned previously, they mainly include short and sharp road segments (Cluster 3) and short-medium segments that are highly connected with the latter (Cluster 4). The sign of the coefficient related to Cluster 3 may be considered inconsistent with literature reporting higher crash risks for skewed road intersections (Harwood et al., 2000; Nightingale et al., 2017; Kumfer et al., 2019), which is supported by the fact that skewed intersections cause longer traverse times than 90-degree intersections and poor visibility for drivers, among other things (Gattis and Low, 1998). However, it is worth noting that most of the research regarding skewed intersections is based on high-speed rural intersections. The Eixample network analyzed in this study represents a low-to-moderate-speed urban area. No significant associations are found for the rest of the covariates, including the multilevel categorical ones representing the number of lanes in the road, the number of entrances/exits and the AADT level.

On the other hand, Model 2 provides a more complex depiction of the effect of the covariates being studied, as it considers a differential effect for each of them depending on the zone type (IAZ or MAZ). Among the $\beta$ parameters, which now represent effects within MAZs, only the one representing the effect of Cluster 3 remains significant (with the same sign as in Model 1). Despite not being significant with 90% credibility, the effects of main roads and the presence of a bus lane should not be completely overlooked according to the confidence intervals obtained. In addition, Model 2 points out the different contribution that some factors may make to the risk of traffic accidents in IAZs or MAZs. Road geometry reflected by Cluster 4 now appears as significant only for IAZs, presenting a even more negative coefficient than in the case of Model 1. Moreover, the association of the two highest levels of AADT, 4 and 5, with traffic accidents presents a differential behaviour between MAZs and IAZs. Indeed, both $\beta$ parameters are significant and positive, suggesting an increase in the number of traffic accidents in MAZs, but the two corresponding $\gamma$ parameters are negative, indicating a protective effect (or, at least, a less detrimental effect) against traffic accidents in the most travelled road segments when a road intersection is near. Therefore, the distinction between IAZs and MAZs has allowed us to find a significant association between some AADT levels and traffic accidents that depends on the proximity to road intersections, a result that somehow compensates the surprising (according to previous research) non-significant estimates found for the AADT levels in Model 1.

The computation of network-constrained KDE values with $\sigma = 100$ m leads to the smooth representation of traffic accidents shown in Fig. 6a. This Figure shows that one of the main avenues located in the network (which also has the highest values for AADT) has a very high accident rate along all its length. Similarly, avenues and main roads bordering the network contain some zones of high accident rates. In contrast, the central part of the network shows much lower values than neighbouring locations. Therefore, the KDE values computed at the middle points of the 975 segments forming the split network were used to find (through LISA values) coldspots and hotspots accurately located



(a) (b)

**Fig. 6.** Quintile distribution of the KDE values for $\sigma = 100$ m (a) and coldspots (green) and hotspots (red) detected after the computation of LISA statistics from these KDE values for the split network made of 20 m IAZ. In (a), each road segment of the split linear network is coloured according to the KDE value at its middle point.

**Table 6**
Relative frequencies of the covariates in the road segments that form the coldspots, average zones and hotspots detected. Each frequency is obtained by averaging the values of the covariates for all the road segments in each set, but weighting them according to their corresponding lengths. For the binary variables only the frequencies of presence at the road segment (value of 1) are shown.

| Covariate | Coldspots | | Average | | Hotspots | |
|---|---|---|---|---|---|---|
| | IAZ | MAZ | IAZ | MAZ | IAZ | MAZ |
| Main road (1) | 0.33 | 0.00 | 0.53 | 0.47 | 0.94 | 0.94 |
| Parking slots (1) | 0.63 | 0.65 | 0.65 | 0.77 | 0.23 | 0.34 |
| Traffic light (1) | 0.25 | 0.00 | 0.52 | 0.07 | 0.63 | 0.29 |
| Bus stops (1) | 0.08 | 0.00 | 0.14 | 0.20 | 0.13 | 0.15 |
| Bus lane (1) | 0.33 | 0.00 | 0.52 | 0.50 | 0.82 | 0.98 |
| No. of lanes (1) | 0.71 | 1.00 | 0.56 | 0.59 | 0.17 | 0.07 |
| No. of lanes (2) | 0.25 | 0.00 | 0.21 | 0.21 | 0.09 | 0.11 |
| No. of lanes (3) | 0.04 | 0.00 | 0.13 | 0.12 | 0.12 | 0.23 |
| No. of lanes (4) | 0.00 | 0.00 | 0.07 | 0.04 | 0.40 | 0.21 |
| No. of lanes (≥ 5) | 0.00 | 0.00 | 0.04 | 0.04 | 0.21 | 0.38 |
| No. of in-neighbours (1) | 0.28 | 0.26 | 0.26 | 0.24 | 0.46 | 0.34 |
| No. of in-neighbours (2) | 0.61 | 0.74 | 0.63 | 0.63 | 0.36 | 0.37 |
| No. of in-neighbours (≥ 3) | 0.08 | 0.00 | 0.11 | 0.14 | 0.18 | 0.30 |
| No. of out-neighbours (1) | 0.16 | 0.39 | 0.27 | 0.21 | 0.42 | 0.33 |
| No. of out-neighbours (2) | 0.84 | 0.61 | 0.63 | 0.67 | 0.37 | 0.34 |
| No. of out-neighbours (≥ 3) | 0.00 | 0.00 | 0.10 | 0.12 | 0.21 | 0.33 |
| Cluster (1) | 0.12 | 0.00 | 0.13 | 0.31 | 0.04 | 0.00 |
| Cluster (2) | 0.20 | 0.39 | 0.40 | 0.51 | 0.22 | 0.62 |
| Cluster (3) | 0.09 | 0.00 | 0.20 | 0.02 | 0.49 | 0.00 |
| Cluster (4) | 0.59 | 0.61 | 0.28 | 0.16 | 0.25 | 0.37 |
| AADT (1) | 0.71 | 1.00 | 0.60 | 0.67 | 0.11 | 0.06 |
| AADT (2) | 0.04 | 0.00 | 0.13 | 0.13 | 0.12 | 0.11 |
| AADT (3) | 0.25 | 0.00 | 0.12 | 0.08 | 0.09 | 0.13 |
| AADT (4) | 0.00 | 0.00 | 0.07 | 0.08 | 0.13 | 0.40 |
| AADT (5) | 0.00 | 0.00 | 0.07 | 0.04 | 0.55 | 0.30 |
| Total road length (m) | 506.87 | 144.95 | 10969.22 | 17626.74 | 2682.96 | 1632.44 |
| No. of road segments | 25 | 3 | 538 | 256 | 120 | 33 |
| No. of accidents | 7 | 1 | 1989 | 1066 | 1928 | 747 |

in the network, which are displayed in Fig. 6b. Identifying coldspots and hotspots enables us to compare the values presented by the covariates in the road segments forming them, but also in the rest of the network. Table 6 contains the mean values (weighted by each road segment's length) of the covariates at coldspots, hotspots and average road segments (neither a coldspot nor a hotspot), which enables us to check the high relative frequency of main roads, bus lanes, 4 or more lanes, 3 or more in- and out-neighbours and levels 4 and 5 of AADT in the road segments belonging to hotspots in comparison to those in coldspots or in average microzones. On the latter, it must be remembered that many of these covariates or levels were not yielded as a significant factor by the count models. Finally, the geometries of type 3 (for IAZs), 2 and 4 (for MAZs) are particularly high in hotspots, which may be unexpected according to the results shown by the two models fitted. In this regard, it is worth remarking that one should not expect road characteristics particularly represented in hotspots/coldspots to display a significant association with traffic accidents from a global modelling perspective. Hence, the combination of two statistical methodologies can either strengthen the validity of the conclusions or call them into question.

## 5. Conclusions

Traffic safety analyses set over areal spatial units have been of interest for many years, but the recent development of statistical techniques on linear networks is bringing new advances and challenges for this subject. Specifically, in this study, a linear network has been used to analyze a geocoded dataset of accidents that took place in the city of Valencia (Spain) during the period 2005–2017. In this regard, the proper consideration of road intersections and the combination of several statistical techniques have been emphasized.

Indeed, the study of traffic accidents around road intersections is of special interest given the high percentage of them that occur close to these road entities. Typically, these analyses are done independently of the values observed for road segments between intersections. This strategy can potentially lead us to miss important relationships (mainly of a spatial type) between intersections and segments in between which may detract from the validity of the results. In this article, the definition of IAZs and MAZs along the directed linear network available has provided a unified approach (involving spatial relationships and the definition of covariates) to this issue that does not exclude any type of road entity.

On the other hand, from a modelling perspective, the coexistence of multiple methodologies to treat accidents datasets provides a flexible framework for analyzing many kinds of specific questions of interest, but this fact also leads to great difficulties when trying to decide on a particular approach. In this study, overdispersion of accident counts and the disparate effects that arise at road segments near intersections, producing both a high concentration of traffic accidents and a high presence of zeros, were addressed through a zero-inflated negative binomial distribution. In addition, spatial relationships between road segments were included with a CAR distribution based on a neighbourhood matrix that accounted for traffic flow. Later, model quality was assessed employing several validation tools, including checks based on simulated data that led to outlier detection, but also more classical techniques such as correlation coefficients and Moran's $I$.

Furthermore, this study has combined the use of spatial count models with the detection of coldspots and hotspots. The results derived from each of the approaches have been discussed and compared, providing coherent results even though some differences were noted. This kind of local analysis could be very useful for validating the results from the statistical models and and questioning some of the conclusions yielded by the former, increasing the robustness of the final results. In this regard, the nature of KDE alleviates the existence of geocoding inaccuracies that may arise when conducting a spatial analysis of this kind, especially when it is done at the road segment level. Indeed, the risk of making mistakes as a consequence of bad geocoding are higher for the construction of the response variable representing accident counts at the road segment level. Here, a small inaccuracy can lead to situating a traffic accident in the wrong road segment, altering the counts of two segments. Kernel density estimation, however, produces a smooth representation of the intensity of traffic accidents along the network that can even absorb some of the geocoding inaccuracies that usually occur, as suggested by Harada and Shimada (2006) and Zandbergen (2009).

Overall, the modelling approach revealed that spatial heterogeneity, overdispersion and the effect of road intersections on adjacent road segments (including zero-inflation) must be accounted by analyzing the distribution of accident counts in the Eixample District of Valencia. The generalization of these findings to other urban areas may be risky, because this kind of analysis is always data-dependent, but it should always be reasonable to consider it. In addition, the detection of hotspots and coldspots identified the fact that main roads, the existence of a bus lane in the road, 4 or more lanes and high AADT values are associated with higher accident counts at the road segment level. The effect of other covariates remained unclear or non-significant and may require further analysis. In any case, this study was slightly limited in terms of covariates, which should be addressed in the future with the availability of more complete and accurate geographic information systems.

## Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

## Conflicts of Interest

The authors declare that they have no conflicts of interest.

## Acknowledgements

We thank José Serrano, Chief of Valencia Local Police Department, the *Unitat d'Atestats i Seguretat en el Transport* and the Statistics Office of Valencia for providing the data used in the study.

## References

Aguero-Valverde, J., Jovanis, P., 2008. Analysis of road crash frequency with spatial models. Transport. Res. Rec.: J. Transport. Res. Board (2061), 55–63.

Alarifi, S.A., Abdel-Aty, M., Lee, J., 2018. A Bayesian multivariate hierarchical spatial joint model for predicting crash counts by crash type at intersections and segments along corridors. Acc. Anal. Prevent. 119, 263–273.

Anastasopoulos, P.C., 2016. Random parameters multivariate tobit and zero-inflated count data models: Addressing unobserved and zero-state heterogeneity in accident injury-severity rate and frequency analysis. Anal. Methods Acc. Res. 11, 17–32.

Anselin, L., 1995. Local indicators of spatial association-LISA. Geogr. Anal. 27 (2), 93–115.

Baddeley, A., Rubak, E., Turner, R., 2015. Spatial point patterns: methodology and applications with R. CRC Press.

Banerjee, S., Carlin, B.P., Gelfand, A.E., 2004. Hierarchical modeling and analysis for spatial data. Chapman and Hall/CRC.

Barua, S., El-Basyouny, K., Islam, M.T., 2016. Multivariate random parameters collision count data models with spatial heterogeneity. Anal. Methods Acc. Res. 9, 1–15.

Besag, J., 1974. Spatial interaction and the statistical analysis of lattice systems. J. Roy. Stat. Soc. Ser B 192–236.

Besag, J., York, J., Mollié, A., 1991. Bayesian image restoration, with two applications in spatial statistics. Ann. Inst. Stat. Math. 43 (1), 1–20.

Bivand, R., Piras, G., 2015. Comparing Implementations of Estimation Methods for Spatial Econometrics. J. Stat. Softw. 63 (18), 1–36.

Briz-Redón, Á., 2019. SpNetPrep: An R package using Shiny to facilitate spatial statistics on road networks. Res. Ideas Outcomes 5, e33521.

Bürkner, P.-C., et al., 2017. brms: An R package for Bayesian multilevel models using Stan. J. Stat. Softw. 80 (1), 1–28.

Cai, Q., Abdel-Aty, M., Lee, J., Wang, L., Wang, X., 2018. Developing a grouped random parameters multivariate spatial model to explore zonal effects for segment and intersection crash modeling. Anal. Methods Acc. Res. 19, 1–15.

Carpenter, B., Gelman, A., Hoffman, M.D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., Riddell, A., 2017. Stan: A probabilistic programming language. J. Stat. Softw. 76 (1).

Castro, M., Paleti, R., Bhat, C.R., 2012. A latent variable representation of count data models to accommodate spatial and temporal dependence: Application to predicting crash frequency at intersections. Transport. Res. B Methodol. 46 (1), 253–272.

Cronie, O., Van Lieshout, M.N.M., 2018. A non-model-based approach to bandwidth selection for kernel estimators of spatial intensity functions. Biometrika 105 (2), 455–462.

Das, A., Pande, A., Abdel-Aty, M., Santos, J., 2008. Characteristics of urban arterial crashes relative to proximity to intersections and injury severity. Transport. Res. Rec. J. Tranport. Res. Board (2083), 137–144.

Fan, W., Kane, M.R., Haile, E., 2015. Analyzing severity of vehicle crashes at highway-rail grade crossings: multinomial logit modeling. J. Transport. Res. Forum 54, 39–56.

Gabry, J., Mahr, T., 2018. bayesplot: Plotting for Bayesian Models. R package version 1.6.0.

Gattis, J., Low, S.T., 1998. Intersection angle geometry and the driver's field of view. Transport. Res. Rec. 1612 (1), 10–16.

Gelman, A., Stern, H.S., Carlin, J.B., Dunson, D.B., Vehtari, A., Rubin, D.B., 2013. Bayesian data analysis. Chapman and Hall/CRC.

Gini, C., 1912. Variabilità e mutabilità. Reprinted in Memorie di metodologica statistica (Ed. Pizetti E, Salvemini, T). Libreria Eredi Virgilio Veschi, Rome (1912).

Graul, C., 2016. leafletR: Interactive Web-Maps Based on the Leaflet JavaScript Library. R package version 0.4-0.

Guo, Q., Xu, P., Pei, X., Wong, S., Yao, D., 2017. The effect of road network patterns on pedestrian safety: A zone-based Bayesian spatial modeling approach. Acc. Anal. Prevent. 99, 114–124.

Hao, W., Daniel, J., 2014. Motor vehicle driver injury severity study under various traffic control at highway-rail grade crossings in the united states. J. Saf. Res. 51, 41–48.

Harada, Y., Shimada, T., 2006. Examining the impact of the precision of address geocoding on estimated density of crime locations. Comput. Geosci. 32 (8), 1096–1107.

Harirforoush, H., Bellalite, L., 2016. A new integrated GIS-based analysis to detect hotspots: a case study of the city of Sherbrooke. Accident Analysis & Prevention.

Hartigan, J.A., Wong, M.A., 1979. Algorithm as 136: A k-means clustering algorithm. J. Roy. Stat. Soc. Ser C (Appl. Stat.) 28 (1), 100–108.

Harwood, D.W., Council, F., Hauer, E., Hughes, W., Vogt, A., 2000. Prediction of the expected safety performance of rural two-lane highways. Federal Highway Administration, United States Technical report.

Huang, H., Abdel-Aty, M., Darwiche, A., 2010. County-level crash risk analysis in Florida: Bayesian spatial modeling. Transport. Res. Rec. J. Tranport. Res. Board (2148), 27–37.

Huang, H., Song, B., Xu, P., Zeng, Q., Lee, J., Abdel-Aty, M., 2016. Macro and micro models for zonal crash prediction with application in hot zones identification. J. Transport. Geogr. 54, 248–256.

Huang, H., Zhou, H., Wang, J., Chang, F., Ma, M., 2017. A multivariate spatial model of crash frequency by transportation modes for urban intersections. Anal. Methods Acc. Res. 14, 10–21.

Kahle, D., Wickham, H., 2013. ggmap: Spatial visualization with ggplot2. R J. 5 (1), 144–161.

Kendall, M.G., 1938. A new measure of rank correlation. Biometrika 30 (1/2), 81–93.

Kumfer, W., Harkey, D., Lan, B., Srinivasan, R., Carter, D., Patel Nujjetty, A., Eigen, A.M., Tan, C., 2019. Identification of Critical Intersection Angle through Crash Modification Functions. Transport. Res. Rec page 0361198119828682.

Lee, J., Abdel-Aty, M., Cai, Q., 2017. Intersection crash prediction modeling with macro-level data from various geographic units. Acc. Anal. Prevent. 102, 213–226.

Mannering, F.L., Bhat, C.R., 2014. Analytic methods in accident research: Methodological frontier and future directions. Anal. Methods Acc. Res. 1, 1–22.

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

Marshall, E., Spiegelhalter, D., 2003. Approximate cross-validatory predictive checks in disease mapping models. Stat. Med. 22 (10), 1649–1660.

McSwiggan, G., Baddeley, A., Nair, G., 2017. Kernel density estimation on a linear network. Scand. J. Stat. 44 (2), 324–345.

Miaou, S.-P., Lord, D., 2003. Modeling traffic crash-flow relationships for intersections: dispersion parameter, functional form, and Bayes versus empirical Bayes methods. Transport. Res. Rec. J. Tranport. Res. Board (1840), 31–40.

Moran, P.A., 1950a. Notes on continuous stochastic phenomena. Biometrika 37 (1/2), 17–23.

Moran, P.A., 1950b. A test for the serial independence of residuals. Biometrika 37 (1/2), 178–181.

Nie, K., Wang, Z., Du, Q., Ren, F., Tian, Q., 2015. A network-constrained integrated method for detecting spatial cluster and risk location of traffic crash: A case study from Wuhan, China. Sustainability 7 (3), 2662–2677.

Nightingale, E., Parvin, N., Seiberlich, C., Savolainen, P.T., Pawlovich, M., 2017. Investigation of Skew Angle and Other Factors Influencing Crash Frequency at High-Speed Rural Intersections. Transport. Res. Rec. 2636 (1), 9–14.

Okabe, A., Satoh, T., Sugihara, K., 2009. A kernel density estimation method for networks, its computational method and a GIS-based tool. Int. J. Geogr. Inform. Sci. 23 (1), 7–32.

Okabe, A., Sugihara, K., 2012. Spatial analysis along networks: statistical and computational methods. John Wiley & Sons.

OpenStreetMap contributors, 2017. Planet dump. retrieved from https://planet.osm.org. https://www.openstreetmap.org.

Papadimitriou, E., Filtness, A., Theofilatos, A., Ziakopoulos, A., Quigley, C., Yannis, G., 2019. Review and ranking of crash risk factors related to the road infrastructure. Acc. Anal. Prevent. 125, 85–97.

Pearson, K., 1896. Mathematical contributions to the theory of evolution. III. Regression, heredity, and panmixia. Philos. Trans. Roy. Soc. London Ser A 187, 253–318.

Quddus, M.A., 2008. Modelling area-wide count outcomes with spatial correlation and heterogeneity: an analysis of London crash data. Acc. Anal. Prevent. 40 (4), 1486–1497.

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

Silverman, B.W., 2018. Density estimation for statistics and data analysis. Routledge.

Spearman, C., 1904. The proof and measurement of association between two things. Am. J. Psychol. 15 (1), 72–101.

Stern, H.S., Cressie, N., 2000. Posterior predictive model checks for disease mapping models. Stat. Med. 19 (17-18), 2377–2397.

Thakali, L., Kwon, T.J., Fu, L., 2015. Identification of crash hotspots using kernel density estimation and kriging methods: a comparison. J. Modern Tranport. 23 (2), 93–106.

Weisburd, D., 2015. The law of crime concentration and the criminology of place. Criminology 53 (2), 133–157.

Xie, K., Wang, X., Ozbay, K., Yang, H., 2014. Crash frequency modeling for signalized intersections in a high-density urban road network. Anal. Methods Acc. Res. 2, 39–51.

Xie, Z., Yan, J., 2008. Kernel density estimation of traffic accidents in a network space. Comput. Environ. Urban Syst. 32 (5), 396–406.

Xie, Z., Yan, J., 2013. Detecting traffic accident clusters with network kernel density estimation and local spatial statistics: an integrated approach. J. Transport. Geogr. 31, 64–71.

Xu, P., Huang, H., Dong, N., Wong, S., 2017. Revisiting crash spatial heterogeneity: a Bayesian spatially varying coefficients approach. Acc. Anal. Prevent. 98, 330–337.

Yang, H., Ozbay, K., Ozturk, O., Yildirimoglu, M., 2013. Modeling work zone crash frequency by quantifying measurement errors in work zone length. Acc. Anal. Prevent. 55, 192–201.

Yasmin, S., Eluru, N., Lee, J., Abdel-Aty, M., 2016. Ordered fractional split approach for aggregate injury severity modeling. Transport. Res. Rec. 2583 (1), 119–126.

Ye, X., Pendyala, R.M., Washington, S.P., Konduri, K., Oh, J., 2009. A simultaneous equations model of crash frequency by collision type for rural intersections. Saf. Sci. 47 (3), 443–452.

Zandbergen, P.A., 2009. Geocoding quality and implications for spatial analysis. Geogr. Compass 3 (2), 647–680.

Zeng, Q., Wen, H., Huang, H., Abdel-Aty, M., 2017. A Bayesian spatial random parameters Tobit model for analyzing crash rates on roadway segments. Acc. Anal. Prevent. 100, 37–43.

Zhao, M., Liu, C., Li, W., Sharma, A., 2018. Multivariate Poisson-lognormal model for analysis of crashes on urban signalized intersections approach. J. Tranport. Saf. Security 10 (3), 251–265.