Annals of Operations Research
https://doi.org/10.1007/s10479-024-06049-7

ORIGINAL RESEARCH

Hierarchical spatial network models for road accident risk assessment

Gian Paolo Clemente[1] · **Francesco Della Corte**[1] · **Diego Zappa**[2]

Received: 19 December 2023 / Accepted: 7 May 2024

Abstract
This paper addresses the critical issue of road safety and accident prevention by integrating road features, network theory, and advanced statistical models. It emphasises the importance of understanding the relationship between road infrastructure and accident risk, which impacts on various administrative stakeholders and on citizens' safety. While existing literature focuses on road features and engineering solutions, this paper highlights the need to consider implicit spatial constraints as well. Our study builds on prior research by proposing a novel approach that merges conditional autoregressive modelling with a two-stage mixed Geographically weighted Poisson regression. This integrated methodology allows us to consider both the effect of risk factors at a global level and at a local road level. By leveraging the strengths of these two methods, we aim to capture both overarching trends and local variations of risk factors, thereby offering a comprehensive understanding of accident risk factors. Using data from the Open Street Map database, which covers the wide province of Milan in Italy, our models identify influential street characteristics, providing valuable insights for informed decision-making regarding road safety measures. Our method can be applied to any region in the world. The paper describes the models used, the dataset employed, and presents a detailed numerical analysis demonstrating the effectiveness of the approach in identifying and understanding accident risk factors within road networks. This information can help guide investments for the benefit of society.

Keywords Accident risk · Claim counts · Spatial dependence · Conditional autoregressive modelling · Geographically weighted Poisson regression

✉ Diego Zappa
diego.zappa@unicatt.it

Gian Paolo Clemente
gianpaolo.clemente@unicatt.it

Francesco Della Corte
francesco.dellacorte1@unicatt.it

[1] Department of Mathematics for Economics, Financial and Actuarial Sciences, Università Cattolica del Sacro Cuore, Largo Gemelli, 1, 20123 Milan, Italy

[2] Department of Statistical Sciences, Università Cattolica del Sacro Cuore, Largo Gemelli, 1, 20123 Milan, Italy

Published online: 06 June 2024

1 Introduction

Road crashes are a significant cause of death, grief and loss worldwide and require thorough investigation and concerted efforts. According to data from the World Health Organization (WHO), road crashes cause over one million deaths annually, placing them among the top ten causes of death in all age groups. Greater compliance with traffic signals and the respect of driving rules is likely to reduce the number of accidents but exploring ways to improve road safety could also be beneficial.

Understanding the relationship between accident risk and road infrastructure has benefits for a range of stakeholders, including urban planners, policymakers, and transportation authorities. However, it also directly affects citizens' experiences when using public spaces. A deeper understanding of how road infrastructure affects accident probabilities enables individuals to make informed decisions about routes, modes of transportation, and overall risk exposure. This knowledge allows individuals to proactively seek safer routes and exercise greater caution in high-risk areas, thereby enhancing personal safety and well-being.

To help policymakers and administration to address their decision to increase road safety, a wide literature has been published to highlight the weakest features of roads or network of roads. From an engineering point of view, pioneering researches by Davies (1944) and Smeed (1949) delved into optimal road composition to mitigate accidents, while recent studies like Christensen et al. (2022) propose the use of Energy Absorbing Structures for crash mitigation. In Greibe (2003) and Mackay (1994) the challenges in preventing road accidents, as road and vehicle improvements, are discussed. However, road network structures often carry implicit constraints within their spatial domains. Tang et al. (2023) focus on urban traffic accidents in China, scrutinising both moving-vehicle and fixed-object crashes using multiscale geographically weighted regression. The research unveils similar clustering patterns for both crash types, uncovering overlapping accident-prone areas. They focus not only on technical aspects but also on selecting road characteristics and their contextual factors influencing accident occurrences.

In this framework, spatial road safety analysis has been developed in order to examine the geographical distribution and patterns of road accidents, enabling targeted interventions and precise adjustments to road infrastructure aimed at reducing risks and enhancing overall safety within specific areas (see Ziakopoulos and Yannis (2020) for a review of spatial approaches). Specifically, models operating on road networks have been extensively used in recent years for analysing streets at a very detailed level. In particular, point processes have been employed to assess the occurrence of events (such as accidents), by considering spatial and topological road features (see, e.g., Baddeley et al. (2021) for a review on these topics). Yet, due to constraints in available datasets, count models often stand as a pragmatic adaptation of point processes within this context (see, e.g., Tang et al., 2023; McSwiggan, 2019; Baddeley et al., 2020).

Within the domain of count models, various approaches exist in the literature, each offering unique insights. In our proposal, we present a merge of two components: a modification of the conditional autoregressive modelling (see Boulieri et al., 2016; Gilardi et al., 2022) incorporating spatial lagged effects, to estimate efficiently the risk of accidents at the road level, and a two-stage mixed geographically weighted Poisson regression (see Murakami et al., 2023; Briz-Redón et al., 2019; Gomes et al., 2017) to unveil local heterogeneity. Two basic versions (semi-parametric geographically weighted Poisson regression and conditional autoregressive prior) of these models have been indeed compared in Xu and Huang (2015), highlighting advantages and drawbacks. We show in this paper how a fusion between them

helps to capture both network-wide trends and local nuances, providing a comprehensive understanding of accident risk factors. This combination is relevant because it combines the strengths of both approaches: the modified conditional autoregressive modelling captures trends in accident risk across the whole area, while the mixed geographically weighted Poisson regression reveals specific behaviour at the road level. By integrating these two methodologies, our approach offers a more comprehensive understanding of accident risk factors, ensuring that both network-wide patterns and local aspects are adequately addressed.

To demonstrate the effectiveness of our approach, we applied it to the province of Milan in Italy. We obtained road features from the OpenStreetMap contributors (2017) database and accident data from the ISTAT - Italian National Institute of Statistics (2021) for the period between 2016 and 2020. By using both models, we identified the key street characteristics that influence accident risk and revealed the distribution of relevant covariates at a road level. This offers valuable insights for informed decision-making in road safety measures.

The paper is structured as follows: Sect. 2 focuses on models specifically designed for spatial data. In particular, Sect. 2.1 provides a rationale for employing count models in our data analysis. Section 2.2 underscores the significance of incorporating spatial dependence in accident risk assessment models and outlines the general framework. Our approach is summarised in Sect. 3. Additionally, Sect. 3.1 details the application of proper conditional autoregressive models with spatial lags, while Sect. 3.2 introduces the two-stage mixed Geographically Weighted Poisson regression method. The description of the dataset and its structure are highlighted in Sect. 4. The subsequent Sect. 5, presents the numerical analysis and it is dedicated to summarise our findings. The discussion focuses on the results derived from the conditional autoregressive models (Sect. 5.1) and the geographically weighted Poisson regression (Sect. 5.2). Conclusions are drawn in the final section.

2 Models for spatial data for road safety: why do we use count model

Due to the absence of modern technology and appropriate algorithms for analysing street-level crash events, the initial papers on spatial road safety analysis were developed using the areal approach (Miaou et al. 2003; Aguero-Valverde and Jovanis 2006; Boulieri et al. 2016). As deeply reported in McSwiggan (2019) problems with aggregation of data are related to the so called " ecological fallacy". Freedman (1999) states: " the ecological fallacy consists in thinking that relationships observed for groups necessarily hold for individuals"(this issue is also known as the modifiable areal unit problem (MAUP), see Gilardi et al., 2022). In our context, groups are the areal clusters used for the analysis and individuals are the accidents in the road segment. For example, we may state that speed is related to the crash intensity in a region but roads have different speed limits and some of them (or many of them) are pedestrian zones. Also thanks to the increasing solutions of computational burdens in the analysis of network of lines, the number of papers estimating the risk at the road level is increasing. Some recent applications are in Borgoni et al. (2021) and Gilardi et al. (2022). In this section, we motivate the use of count models, then we focus on count models with spatial dependence and finally we introduce proper conditional autoregressive models. We highlight some limitations of the existing models in our context in order to justify the approach proposed in Sect. 3.

2.1 Spatial point patterns vs count models

Models based on a network of lines have recently gained popularity due to the availability of open-source spatial databases (e.g. OpenStreetMap contributors, 2017) for analysing street networks at a very detailed level (see Barua et al., 2014; Barrington-Leigh & Millard-Ball, 2017; Mooney & Minghini, 2017; Briz-Redon et al., 2019, Gilardi et al., 2022). In this context, many approaches have employed spatial smoothing to simplify the estimation process. However, in this paper, we adopt count models on a network of lines because they offer high flexibility, allowing us to investigate crash intensity at the street segment level. This approach can provide more informative insights for social and policy monitoring. Alternative strategies and interpretations of road networks are extensively described in Lord and Mannering (2010), Savolainen and Mannering (2011), and Ziakopoulos and Yannis (2020).

To explain the reasons for using count models, we revisit the fundamental principles of a linear network (see Baddeley & Nair, 2012). It is defined as the union $L = \bigcup_{i=1}^{m} s_i$ of a finite number m of line segments $s_1, \ldots, s_m$ in the plane, where $s_i = [\zeta_i, \nu_i] = \{w : w = t \cdot \zeta_i + (1-t) \cdot \nu_i, 0 \leq t \leq 1\}$ is the line segment with endpoints ζ_i, ν_i, belonging to the two-dimensional space. Crashes located on a network correspond to a point pattern $\mathbf{x}$ on L. A point pattern is a finite set $\mathbf{x} = \{x_1, \ldots, x_n\}$ of distinct points $x_i \in L$, where $n \geq 0$. For any set $B \subset L$, let $N_x(B) = N(\mathbf{x} \cap B)$ be the number of points of $\mathbf{x}$ lying in B (Rakshit et al., 2021).

To assess the impact of covariates on occurrences, many models employed in spatial point processes use a density of the form $\frac{g(u,\boldsymbol{\theta})}{G(\boldsymbol{\theta})}$ where g represents an explicitly defined function expressed in terms of interaction functions with the data, $\boldsymbol{\theta}$ is a vector of parameters, and $G(\boldsymbol{\theta})$ is a normalizing constant that cannot be explicitly evaluated. Since $G(\boldsymbol{\theta})$ is unknown, conducting standard likelihood estimation becomes challenging (Jensen & Møller, 1991).

In modelling car crashes, let us assume that point locations are distinct and possess an intensity function $\lambda(u), u \in L$, enabling the computation of the average accidents in L, defined as it follows:

$$\mathbb{E}[N_X(B)] = \Lambda(B) = \int_B \lambda(u)\, \mathrm{d}_1 u, \tag{1}$$

where $\mathrm{d}_1 u$ denotes integration with respect to arc length (Ang et al., 2012; Baddeley et al., 2017; Rakshit et al., 2021). In the case of non-homogeneous Poisson point process, which is the most commonly used process for modelling crashes on a linear network, a significant implication is that the points in $\mathbf{x} \cap B$ are assumed to be independent and identically distributed (i.i.d.) with a probability density function $f(u) = \frac{\lambda(u)}{\Lambda(B)}$. Consequently, the log-likelihood can be expressed as:

$$l = \sum_{i=1}^{n} \log \lambda(x_i) - \int_L \lambda(u)\, \mathrm{d}_1 u. \tag{2}$$

This expression is numerically intractable and necessitates approximation, which can be achieved using the Berman–Turner device specifically designed for networks (Berman & Turner, 1993; McSwiggan, 2019).

A crucial technical consideration in this context is that a spatial covariate V on L is assumed to be a real- or vector-valued function $V(u), u \in L$. It is furthermore assumed that the values $V(u)$ are fixed and known (in principle) for all locations $u \in L$, in order to model, within a Generalised Linear Model (GLM) framework, links of the form $\lambda(u) = \exp(\boldsymbol{\beta}^T V(u))$, where $\boldsymbol{\beta}$ represents the standard vector of parameters. In practical scenarios, these values may only be available at specified sample locations. and it is relevant to possess information about $V(u)$ at locations u beyond those confined to the point pattern. An alternative to GLM

is the estimation of spatially-varying event densities using kernel density estimators. The goal of density estimation is to statistically infer the spatially-varying density from an observed point pattern $\mathbf{x}$ with only minimal assumptions about the underlying point process. However, it is important to note that this topic falls outside the scope of this paper. We simply mention that kernel estimation on a linear network is not a standard application of kernels and can be computationally complex; various different techniques have been proposed (see Baddeley et al., 2021 for a review on this topic).

In general, the framework depicted above suggests that, in principle, any point process model applied to a linear network should not assume that a covariate remains constant along any edge of the network. This is achievable, for instance, when information about the distance of a crash from an intersection can be accurately computed or when the variational risk (such as the length of road visible to a driver at their location) along a curve can be determined. However, it is worth noting that in cases where extremely high map resolution and precise location data are not always available, some authors Briz-Redon et al. (2019) propose a workaround by capturing the differential risk, for example, between road locations around intersections and road segments. This is done by dividing the original network structure into shorter road segments in the proximity of each road intersection for a more refined analysis.

Similar to many other applications described in the literature, the dataset available to us does not allow the fit of a point process. This is primarily due to the fact that the location of accidents often lacks the corresponding street number or the exact kilometer of the event occurrence, especially in cases of accidents along highways. Consequently, we will investigate count models with spatial components.

If we assume that the linear network is partitioned into J disjoint subsets $l_1, \ldots, l_J$ (for instance, the edges of the network) and if the spatial covariate functions are assumed to be invariant on each subset, then

$$\mathbf{V}(u) = \mathbf{v}_j \text{ for } u \in l_j \tag{3}$$

where $\mathbf{v}_j = (v_{j1}, \ldots, v_{jp})$ is the vector of p features of l_j. This allows us to aggregate the point processes over these subsets, resulting in observable random variables $N_j = N(\mathbf{x} \cap l_j)$, representing the counts of points falling into each subset, with $\mu_j = \mathbb{E}[N_j]$ that denotes the expected number of counts in l_j. These assumptions hold significance in our context, aligning with the customary practice in most of road accident research where the subsets l_j correspond to the original segments which defined the network, and the available covariates remain constant along each segment.

In this scenario, the non-homogeneous Poisson process with a log-linear intensity exhibits a constant intensity within each subset. This implies that N_j can be used to fit a Poisson log-linear regression with

$$N_j \sim \text{Pois}(\mu_j)\ \forall j \text{ i.i.d. r.v. with } \mu_j = \int_{l_j} \lambda(u)\, \mathrm{d}_1 u. \tag{4}$$

Consequently, the log-likelihood of a point process corresponds to a Poisson count model

$$l = \sum_j \left(n_j \log \mu_j - \mu_j\right) \tag{5}$$

2.2 Spatial dependence

In order to model spatial data (see, for example, Glaser, 2017; Gschlossl & Czado, 2007), we must introduce the concept of spatial dependence, which necessitates the definition of a

proximity matrix. Let $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)^T$ be the vector of random variables observed in n different regions (in our case, edges). Consider the $n \times n$ proximity matrix or neighbourhood matrix, $\mathbf{W}$. In an area partitioned into n subareas, the element w_{ij} may represent the weight, the connectivity degree, or the spatial proximity intensity between l_i and l_j, with $w_{i,i} = 0 \; \forall i \in \{1, ..., n\}$. The other elements can be selected rather arbitrarily. One of the most common choices to define the non-null weights w_{ij} is based upon neighbourhood considerations:

$$w_{ij} = \begin{cases} 1, & \text{if } j \in Q(i) \\ 0, & \text{if } j \notin Q(i) \end{cases} \tag{6}$$

where the set $Q(i)$ denotes the neighbours of variable Y_i. Specifically, w_{ij} equals 1 when areas i and j share boundaries, and 0 otherwise.

The next ingredient is the model. The most common model is the conditionally autoregressive (CAR) model by Besag (1974). Its general form is:

$$Y_i = \mu_i + \phi_i + e_i \tag{7}$$

where, ϕ_i and e_i are the spatially structured random effect and the unstructured effect, respectively, i.e., a location-specific component. To incorporate spatial covariates, a mean specification dependent on covariates can be introduced by setting:

$$\mu_i = \mathbf{x}_i^T \boldsymbol{\beta} \tag{8}$$

where $\mathbf{x}_i^T = (1, x_{i1}, \ldots, x_{ik})^T$ is the transposed vector of k covariates observed in the region/link i, and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)^T$ is a vector of parameters. For identification purposes, this model is typically addressed within a Bayesian framework to capture local and spatial uncertainty due to factors not measurable only through the datasets, also because of the dynamics of crash occurrences. A prior distribution for each ϕ_i is assumed to be:

$$\phi_i \mid \phi_j, j \neq i \sim N\left(\sum_j w_{ij}\phi_j, \sigma^2\right) \tag{9}$$

i.e., it is assumed to be distributed according to a normal random variable with an average equal to the sum of the weighted values of its neighbours and an unknown variance. The joint distribution of $\boldsymbol{\phi}$ is a multivariate normal random variable $\boldsymbol{\phi} \sim N(\mathbf{0}, \mathbf{Q}^{-1})$ with $\mathbf{Q} = [\tau\mathbf{D}(I - \alpha\mathbf{B})]$ where $\mathbf{D}$ is a diagonal matrix whose entries represent the number of neighbours, τ is the precision parameter, α controls the degree of spatial correlation ($\alpha = 0$ implies spatial independence and $\alpha = 1$ implies complete spatial correlation). Finally, $\mathbf{B} = \mathbf{D}^{-1}\mathbf{W}$ is the scaled adjacency matrix.

Due to computational complexity (see Cantaluppi et al., 2023), the Intrinsic Conditional Auto-Regressive (ICAR) models are often employed, setting $\alpha = 1$. In this case, $\mathbf{Q}$ simplifies to $\tau(\mathbf{D} - \mathbf{W})$. It is possible to prove that the ICAR prior may face rank-deficiency issues. In general, a set of sum-to-zero constraints on the vector $\boldsymbol{\phi}$ is necessary for each group of connected segments (Hodges et al., 2003). In our application, we deal with a fully connected road network, so we only need to impose one set of constraints (Gilardi et al., 2022).

Applying this method to data of a network of lines has several limitations, as it does not consider the true distance between two locations in a directed graph. Additionally, it must be applied to a whole area, making it challenging to interpret the impact on the estimates of different domain characteristics.

Therefore, in the following section, we provide a different proposal that allows us to solve the issues listed above.

3 What model for crash intensity?

We propose an extension of existing models to fit car crashes. Specifically, our proposal consists of two main steps. The first one enhances CAR by incorporating a spatial lag of $\mathbf{X}$ (SLX), a concept borrowed from the econometric literature. This step is relatively fast, reliable and flexible. It allows to model spatial heterogeneity and to estimate in a compact way the risk of accidents of the edges of the area under consideration. The second step introduces a novel two-stage Graphically Weighted Poisson Regression (GWPR). This step proves invaluable for a more in-depth exploration of the characteristics that influence a specific segment or edge of the road network, revealing the key components that are likely to contribute significantly to the expected local accident rate. The two solutions will be detailed in Sects. 3.1 and 3.2, respectively.

3.1 CAR with SLX

With the purpose of extending CAR, we need to consider the distance along a directed network, taking into account the real constraints of street navigation from one point to another. To include the intuitive notion that links too far apart are conceptually not dependent, we employ the bi-square kernel function (see Nakaya et al., 2005), where the weights w_{ij} with $i \neq j$ of the proximity matrix are defined as follows:

$$w_{ij} = \begin{cases} \left[1 - \left(\frac{d_{ij}}{h}\right)^2\right]^2, & \text{if } d_{ij} < h \\ 0, & \text{if } d_{ij} \geq h \end{cases} \tag{10}$$

and d_{ij} is the distance between i and j, h is the bandwidth and it will be selected using a cross-validation procedure. It is worth pointing out that the weight decreases as the distance increases, and the selection of h allows to add sparsity to the matrix, resulting in a significant reduction of the computational time.

Additionally, to account for the constraints imposed by the road infrastructure, we do not define d_{ij} as the Euclidean distance but as the weighted shortest path distance. Specifically, a directed path in a graph denotes the shortest route from one node (or vertex) to another, where links have the same specific direction. The length of this path, in the case of a weighted graph, is obtained by summing the weights of each link involved in the path.

Given the distances outlined above, we propose a hybrid of two spatial autoregressive models (SLM and SLX, for detailed information and further references see Vega & Elhorst, 2015). In general, the average count of a spatially dependent variable in the i-th road is modelled according to:

$$Y_i | \{y_j, j \in N(i)\} \sim \text{Pois}(E_i \mu_i) \tag{11}$$

where the dependent variable, conditioned on its neighbours, follows a Poisson distribution. For practical applications in the insurance context, see, for instance, (Gschlossl & Czado, 2007; Tufvesson et al., 2019).

Notice that E_i represents the exposure parameter, also known as offset. This component is added to the model to consider the occurrences of the response variable in relation to a unit measure. In this context, it is common to use metrics such as the length of the road, traffic volume, or the vehicle miles travelled (VMT), which is the product of the road length and the number of cars over a specified time unit. Specifically, VMT serves as an indicator of the total distance travelled (in miles or kilometers) by all vehicles within a specific area/road

and time period. It aids in identifying regions that experience higher travel frequency. For our purposes, we will use the VMT.

In general, traffic information is available at some cost. Most common providers (as Google, HERE, Bing, TomTom, etc.) partially allow the download of traffic flows for free, but only for limited areas or periods. Alternative indicators can be obviously used. Boulieri et al. (2016) use annual average daily traffic, which is the number of vehicles travelling along a given point on a highway on an average day in the year. Authors in Gilardi et al. (2023) propose measurement error models to filter the uncertainty in measuring traffic. For our purposes, we saved traffic data for the Milan province.

Considering μ_i, typically it is assumed that

$$\log(\mu_i) = \sum_{j=1}^{k} x_{ij} \cdot \beta_j + \phi_i + e_i \tag{12}$$

The fixed and the spatial components on the right, which encompass the link characteristics, are not indexed with respect to time because they refer to the structure of the network, which is substantially assumed to be constant over the reference time horizon. It models the average occurrences on a link in the time unit. Due to the lack of consistent data at sufficiently short intervals, for the response variable we consider the time component only at the yearly level using the approach proposed in Korn (2021).

To address the aforementioned details, we propose combining the CAR model and the SLX model to limit unobservable components in formula (7) to measure clearly the impact of features on risk. We will also explicitly incorporate the spatial dependence of features and accident occurrences. This approach will ensure a logical flow of information to consider the influence of neighbouring network structure.[1] We set:

$$\boldsymbol{\mu} = \rho \mathbf{W}\mathbf{y} + \mathbf{X}\boldsymbol{\beta} + \mathbf{W}\mathbf{X}_{(-1)}\boldsymbol{\eta} \tag{13}$$

with $w_{ij} = 0$ for $i = j$ and $\|w_{i.}\| = 1$ ensuring row-wise normalization of weights to 1. The parameter ρ modulates the spatial dependence among occurrences, making it a valuable component for identifying areas more susceptible to accidents. $\boldsymbol{\eta}$ is a vector of parameters that measures the marginal impact of the features of explanatory variables associated with neighbouring links. The notation $\mathbf{X}_{(-1)}$ indicates the $\mathbf{X}$ matrix without the column associated to the intercept.

The average of the spatial Poisson model can be expressed in a reduced form as:

$$\log(\lambda) = (\mathbf{I}_n - \rho \mathbf{W})^{-1} \left(\mathbf{X}\boldsymbol{\beta} + \mathbf{W}\mathbf{X}_{(-1)}\boldsymbol{\eta}\right) + \text{offset} \tag{14}$$

To facilitate rapid estimation of ρ, $\boldsymbol{\beta}$, $\boldsymbol{\eta}$, and their associated uncertainty, we employ the software INLA[2] (Gómez-Rubio et al., 2020, 2021; Lindgren & Rue, 2015), utilising the homonymous R package (R Core Team, 2022; Bivand et al., 2015). A hierarchical structure is assumed. We have set $0 < \rho < 1$ with $\theta \sim \text{logitbeta}(1, 1)$[3] where $\theta = \log\left(\frac{\rho}{1-\rho}\right)$. For each parameter in $\boldsymbol{\beta}$, $\boldsymbol{\eta}$, Gaussian vague priors, i.e. $N(0, 1000)$, are employed.

To reduce the overall complexity of formula (14) we cannot perform standard variable selection methods. There are no p-values in INLA. Importance or significance of variables

[1] This model is also known under the name of Spatial Durbin Model (SDM)

[2] Detailed documentation available at https://www.r-inla.org/documentation.

[3] If $\theta = logit(\rho)$ and $\rho \sim Beta(a, b)$, then θ is distributed as a Logit-Beta with density algebraically written as: $f(\rho) = \dfrac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \rho^{a-1}(1-\rho)^{b-1}$.

requires to examine if their (e.g.) 2.5% and 97.5% posterior estimates overlap zero. This involves removing covariates and seeing how this might change model fit according to the model's Deviance Information Criterion (DIC). We executed this step iteratively, reducing progressively the number of covariates until when the removal of any additional covariate resulted in a DIC increase exceeding 10%. In addition a hierarchical selection has been applied: if a variable in $\mathbf{X}$ is removed then also the corresponding variable in $\mathbf{WX}_{(-1)}$ is removed. The vice versa was not allowed. To consistently account for the year-to-year effect, the year component has never been removed. The step is somehow computational demanding but it helped to improve significantly the interpretation of the final models.

3.2 GWPR: a hierearchical approach

To integrate the detailed analysis of a Poisson point process with the approximations of a Poisson count model and to leverage the flexibility offered by variable selection techniques in identifying subsets of covariates that can capture local heterogeneity, we propose a two-stage mixed geographically weighted Poisson regression (GWPR).

Geographically weighted regression (GWR) (see Brunsdon et al., 1998) performs regression analyses using local samples within a specified bandwidth distance to explore spatially varying relationships between explanatory and dependent variables. Specifically, we focus on the GWPR version, which assumes at the generic position u:

$$Y_j \sim \text{Pois}(\mu_j E_j) \tag{15}$$

Y_j are i.i.d. and $\mu_j = \exp(\mathbf{x}_j \boldsymbol{\beta}(u))$ and E_j is the offset, while $\boldsymbol{\beta}(u)$ is the vector of varying coefficients. These coefficients are calibrated using a kernel regression methodology, in which we estimate smoothed geographical variations of parameters with a spatial weighting kernel. The estimates of the parameters are calibrated in a point-wise way (see Brunsdon et al., 2005). The local log-likelihood at position u is:

$$l(u) = \sum_{j=1}^{n} (\mu_j E_j + y_j \log(\mu_j E_j)) \cdot w_{ij}(\|u - u_j\|) \tag{16}$$

where w_{ij} is the geographical weight of the j-th observation at the i-th regression point. The weighting kernel used is the bi-square kernel function provided in formula (10), with d_{ij} set to the shortest path distance.

It is noteworthy that the parameters depend on u. This implies an interaction between geographical location and the functional relationships within the linear predictor. The model potentially encompasses different coefficients for each u. What is particularly intriguing is that while conventional kernel regression modelling aims at estimating a regression function f by approximating it with polynomials centred on specific values of $\mathbf{x}_i$, in GWPR, the likelihood is geographically weighted with the weights being determined by a kernel function centred on u. The key distinction lies in the fact that in GWR, the kernel is defined in a geographical space while the regression model pertains to predictor-variable space.

Therefore, this approach enables us to map the variation in the regression coefficients, providing insights into the spatial patterns between the predictor and response variables. An example of the application of standard GWR for car crashes is Pirdavani et al. (2014). What the literature on crash occurrences has not yet thoroughly examined is the potential geographical dependency of certain covariates, while others may not exhibit this characteristic. For instance, traffic lights in urban areas may demonstrate spatial dependence, but in rural

areas, that may not hold true. Hence, for μ_j in (15) we propose the following structure:

$$\mu_j = \exp\left(\sum_{j=1}^{k} x_{ij}\beta_j(u_i) + \sum_{m=1}^{q} x_{im}\gamma_m\right) \tag{17}$$

where γ_m are the coefficients of the variables that do not show geographical dependence. The estimation process involves an iterative method (see (Nakaya et al., 2005)). However, it appears crucial how to effectively partition the explanatory variables into the two subsets.

To this purpose, we propose a hierarchical GWPR algorithm:

1. Setting the bandwidth
   - Start with an initial value for the bandwidth, denoted as h.
   - Define $\mathbf{z}_{(-i)}$ as the vector $\mathbf{z}$ without the i-th element or the statistics obtained by removing the i-th row of the data matrix.
   - Solve the following optimization problem:

$$\hat{h} : \min \sum_{i=1}^{n} \left[y_i - \hat{y}_{(-i)}(h)\right]^2 \tag{18}$$

   where

$$\hat{y}_{(-i)}(h) = E_j \exp(\mathbf{x}_j \hat{\boldsymbol{\beta}}(u)) \tag{19}$$

   and

$$\hat{\boldsymbol{\beta}}(u) : \max_{\boldsymbol{\beta}(u)} \left\{l_{(-i)}(u)\right\} \tag{20}$$

   Note: The mean squared prediction error is used, but it can be replaced with other appropriate cost functions like deviance, Akaike Information Criteria, etc.

2. Penalised GWPR
   - Formulate the penalised GWPR problem as follows:

$$\max_{\boldsymbol{\beta}(u)} \left\{ l(u) - \gamma \left(\sum_{l=1}^{p} \left[(1-\alpha)\beta_l^2(u) + \alpha|\beta_l(u)|\right] \right) \,\middle|\, \gamma, \alpha, \hat{h} \right\} \tag{21}$$

   where $\gamma > 0$ is the regularisation parameter and $0 \leq \alpha \leq 1$ moderates the elastic net penalty (Zou & Hastie, 2005).

3. Splitting the explanatory variables
   - Divide the explanatory variables into $\mathbf{X}_{(-\mathbf{z})}$ and $\mathbf{Z}$, where $\mathbf{Z}$ contains the not significant variables. In case of $\alpha = 1$ (i.e., the LASSO framework), $\mathbf{Z}$ will include variables with coefficients equal to 0. $\mathbf{X}_{(-z)}$ contains all the variables but those in $\mathbf{Z}$.
   - $\mathbf{Z}$ might contain variables that are either truly irrelevant to the problem or are local, i.e., with negligible to null geographical effect.

4. Orthogonal subsets
   - Using the vector of residuals at step 3, apply a penalised (not geographical) elastic net method using only the variables in $\mathbf{Z}$. Let $\mathbf{Q}$ be the variables selected by the method. $\mathbf{Q}$ should contain the subset of not geographical dependent variables.

5. Mixed GWPR

Table 1 Comparison of CAR-SLX and GWPR approaches

Approach	Pros	Cons
CAR-SLX (Spatial Durbin model)	Flexible weighting scheme Estimates of the risks of a local network in one step Extension to hierarchical setup	Domain Partition is necessary Difficult interpretation of estimates related to lagged components Estimates of risk factors fixed for a given area
GWPR (Geographically weighted Poisson regression)	Possibility to identify road characteristics that specifically contribute to the risk Straightforward interpretation of estimates Possibility to study the distribution of local risk factors	long computational time Local estimates may be affected by clustering Pre-calibration is compulsory

- In case the set $\mathbf{Q}$ is not empty, consider potential covariate dependence not addressed in step 3. Fit a mixed-GWPR (see Nakaya et al., 2005) with:

$$\log(\mu(u)) = \mathbf{X}_{(-\mathbf{z})}\boldsymbol{\beta}_{(-\mathbf{z})}(u) + \mathbf{Q}\boldsymbol{\delta} \tag{22}$$

with geographical weights applied only to $\mathbf{X}_{(-\mathbf{z})}$ and penalization given by:

$$\max_{\boldsymbol{\theta}} \left\{ l(u) - \gamma \sum_{l=1}^{p} [(1-\alpha)\theta_l + \alpha|\theta_l|] \,\middle|\, \gamma, \alpha, \hat{h} \right\} \tag{23}$$

where $\boldsymbol{\theta} = \begin{bmatrix} \boldsymbol{\beta}_{(-\mathbf{z})}(u) \\ \boldsymbol{\delta} \end{bmatrix}$.

To our knowledge, the attempt to blend elastic net methodology and GWPR is novel in the literature. Although the computational time may be non-negligible, advancements in technology have somewhat mitigated this concern. The main aim of our proposal is to gain evidence regarding the local impact of variables. While car crashes are undoubtedly affected by general factors (e.g. traffic, density of population, speed,...), addressing specific/local components (e.g. lack of traffic lights, pedestrian crossings,...) is essential. Therefore, this approach allows for more effective implementation of political and administrative decisions without affecting overly large domains. An alternative approach would involve fitting a penalised CAR-SLX Poisson model to identify which components in the lagged $\mathbf{WX}$ variables are not selected. The main limitation of CAR-SLX lies in interpreting the estimates (see (Golgher & Voss, 2016) for a comprehensive overview of parameters interpretation in spatial econometric models) while GWPR offers interpretation consistent with generalised linear modelling. In our opinion both methods provide important information to investigate the risk of accidents. In Table 1 we list some pros and cons of both.

4 Datasets for car crash modelling: description and issues

In this section, we provide details about the datasets used for assessing the risk associated with car crashes. Our focus is on presenting the key characteristics and challenges we have dealt with.

Table 2 Characteristics retrieved for road networks

Features	Features
Pedestrian crossings	Congestion hazard
Traffic lights	Priority over oncoming traffic
Number of crossings*	Crossing with priority
Start/end of no overtaking	Type of lanes
Sharp curves	Speed limit categories
Stop signs	Urban roads
School zones	Roundabouts
Steep hill	Functional class*/type of roads
Animal crossing	Direction travel
Slippery road	Population density*
Curvature degree*	Building density*

Variables with "*" can be added to the database implementing a specific procedure necessary for their computation

- *Road Network and its Characteristics* In our applications, we extensively used the Open Street Map (OSM). This data can be obtained through various methods, such as using the R packages `osmdata` Padgham and Rudis (2017) or via the Overpass Application Programming Interface (API). This API provides custom-selected portions of the OSM map data for download. Numerous details are available for each segment of the network. A comprehensive list of covariates used to model spatial point objects is reported in Table 2.
- *Specific Considerations for Variables* "*Number of Crossings*" *and* "*Curvature Degree*" The native OSM database does not naturally provide information on the number of road crossings for each link. To include this key risk factor, we estimated the number of crossings by computing the number of links that intersect with other links and, at the same time, culminate in a stop sign. This joint condition is necessary because typically a road can be split into subsequent segments, where the vertices do not always represent road crossings. Similarly, we have also estimated the curvature degree of each road. This metric indicates the minimum number of shape points required to maintain a curve, within 2 ms of a road's centreline. Both of these features have frequently shown significance in predicting accident occurrences.
- *Socio-demographic Features* Many studies (see, e.g., ISTAT - Italian National Institute of Statistics (2021); Choudhari and Maji (2019)), have reported the correlation between local socio-demographic factors (such as population density, family concentrations, housing etc.) and road casualties. This information is typically not available at a street level. On Italian official websites, this classification is reported at a census level, which represents the highest degree of territorial subdivision in Italy. To integrate socio-demographic data with specific road segments, a specific procedure has been implemented to match for each road the corresponding socio-demographic value. For applications these features have been grouped into 7 classes with the same frequency (we have used the empirical quantiles to fix the extremes of the classes).
- *Types of Accidents and Accident Locations* The list of accidents we have used refers to crashes provided by the Italian National Office of Statistics (ISTAT) that includes only accidents reported to the police. Thus it mostly lists crashes related to fatalities or injuries. In Appendix A, we report references for obtaining this kind of data in Italy

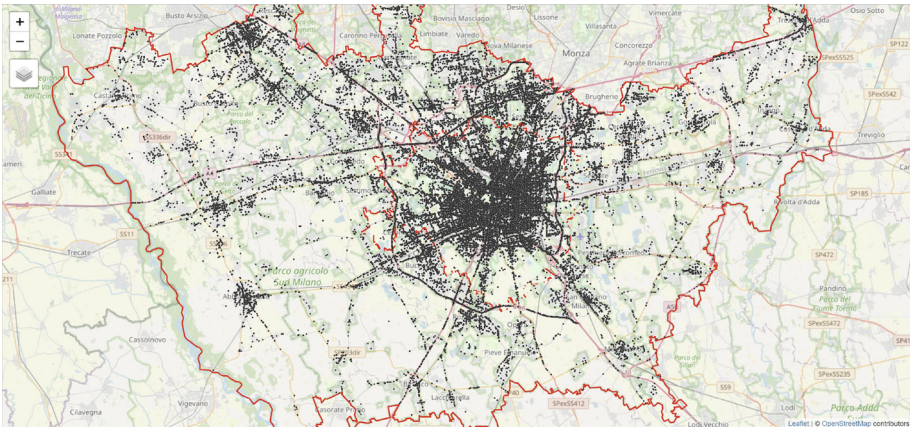

Fig. 1 Map of Milan Province displaying accident locations marked with black dots. The inner red line delineates the boundary of the city of Milan. (Color figure online)

and in other countries. For the Italian dataset, most of accident locations lack details such as house numbers or precise positions along primary roads or highways. In such cases, the accidents have been geographically centred at the middle of the street. As detailed in the previous section, in the count model framework this approximation has a negligible impact. The list of accidents' locations lacked georeferencing, necessitating reverse geocoding to map them onto the road network. It is important to notice that coordinates returned by the reverse geocoding procedure (in our case based on the R package `hereR` Unterfinger and Possenriede (2023)) may not perfectly match the road network provided by a different source, such as OSM. Some degrees of approximation is possible (Kılıç et al. (2023)). Consequently, it is possible that the coordinates are just in close proximity to road segments. To overcome this issue, we employed the strategy to project accidents orthogonally onto the nearest roads selecting the link with the shortest distance from them. To account for the potential misassignment of an accident to a wrong road through this strategy, we implemented a code to randomly assign those accidents to one of the roads within a convex hull built near the coordinates that do not align with the road network. The impact on the estimates is negligible and we do not report these results in this paper. The map illustrating accident locations is depicted in Fig. 1. In Fig. 2, we present an example from a specific area in Milan, where certain misalignments were observed and subsequently addressed by employing the random assignment method described above. From Fig. 1 it is clear the high concentration of events in Milan city.

- *Short Segments and Structural Zero Accidents* Short segments and structural zero accidents pose a challenge in our network, composed of thousands of short segments. Indeed, some segments may be associated with zero accidents not only because no events have occurred in the past, but also due to the extremely short length of the segments, making it improbable for an event to be observed there. This is not a problem that can be resolved with a zero inflated model alone. Such an issue, also noted in Gilardi et al. (2022), may artificially inflate the number of segments with zero crashes, potentially biasing the results of models. To mitigate this, each model was fitted by systematically filtering out edges with lengths ranging from 0 to 500 ms in intervals of 25 ms. The best filter has been set at the level that allowed the best prediction in terms of minimizing the deviance of residuals on the set of segments removed by the process. The procedure

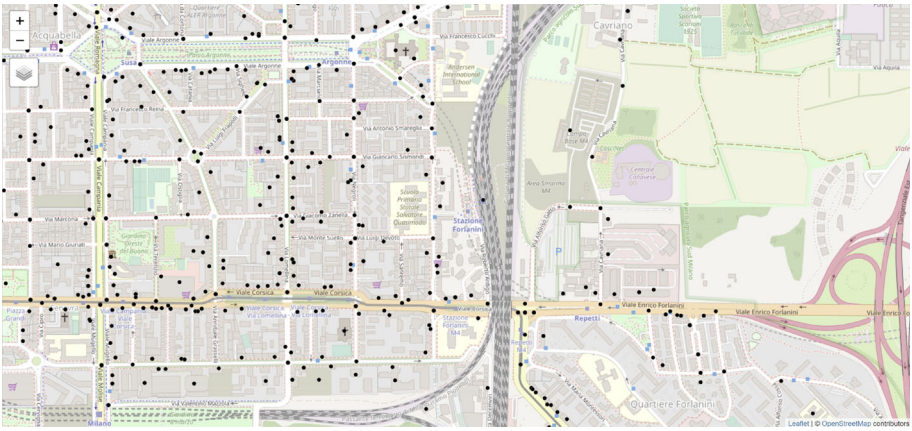

Fig. 2 An excerpt from the map of Milan Province showing accident locations and highlighting specific reverse geocoding misalignments, which were subsequently resolved using the random assignment method described in Sect. 4

significantly enhanced the quality of the fit while minimizing the impact of roads with potential structural zero accidents.

5 Case study

The proposed approach has been subjected to testing employing both CAR-SLX and GWPR methods across the extensive areas of Milan city and its province. This assessment aims to evaluate the method's robustness when applied to different network structures. The dataset covers the period from 2016 to 2020. It comprises 168,965 links totalling 49,377 claims related to personal injury occurred along a total of 11,672.88 km network length. By splitting the information between the city of Milan and the province, we display details in Table 3, where we report some statistics for most relevant road features.

With the only exception of Milan the area has been split into the 132 municipalities. For Milan we have used the corresponding 38 postal code areas (ZIP code). Different tassellation methods can be adopted. Our choice is addressed by the possibility of easy interpretation of results for decision makers. The weighing scheme in model (13) and, natively, in the GWPR model mitigate the subjective choice of any tassellation scheme. To consider the local network complexity the overall length of the network of each area has been used to weight the averages of the road feature of each sub-area. The weighted average of accidents (per year) equals 25.44 for the province versus an average of 171.51 accidents for Milan, thus showing the much higher risk of accidents in the Milan urban area. The empirical quantiles of order 5 and 95 compared with the average show a high asymmetry for most of the variables. As expected pedestrian crossing and traffic lights are much more concentrated in Milan city than in the province.

Although we have not reported the details in Table 3, depending on the areas under investigation, we observed sometimes a strong unbalance in some dichotomous explanatory variables. It may cause quasi-collinearity problems. To address this issue, our program systematically checks the numerical stability of results. In the rare instances where the problem was observed, we have implemented a function to investigate the causes of the failure. Cate-

Table 3 Some descriptive statistics of selected road features of the province of Milan and of Milan city

	Province of Milan (city excluded)				City of Milan			
	Total	Average	Quantile 5%	Quantile 95%	Total	Average	Quantile 5%	Quantile 95%
# areas	132	–	–	–	38	–	–	–
# crashes per year	3358	25.44	1.20	109.91	6517	171.51	92.74	282.14
# edges / links	134,172	1016.45	261.70	2429.85	34,793	915.61	519.62	1482.25
% edges retained	96.18%	96.76%	89.10%	99.80%	99.13%	99.09%	98.07%	99.84%
# edges removed	5129	32.94	–	–	302	8.31	–	–
Length of the network (in km)	9416.96	100.41	20.43	152.71	2255.92	71	30.63	108.90
# pedestrian crossings	8240	91.24	10.55	162.90	4019	121.55	40.25	182.15
# traffic lights	355	3.91	0.00	7.45	186	5.12	0.00	13.15
# vertical signals of danger or warning	9038	100.46	10.00	182.95	1572	48.52	10.55	72.15
# signals of no overtaking	1792	16.52	1.00	33.45	131	4.73	0.00	20.05
# vertical signals of stop	4139	46.66	0.55	150.00	1405	37.07	19.400	64.90

gories found to be collinear with others are removed, and the code restarts from the beginning to ensure robustness of results. To limit the impact of short segments, 5431 segments have been filtered out as per Sect. 4. For each sub-areas on average the percentage of edges of the original network we have used to fit our models is 96.18% and 99.13% for the province and the city of Milan, respectively. The high percentages of valid edges we have used allowed to map almost completely the whole network.

5.1 CAR-SLX model fit

We start focusing on the fit of the model in formula (14). Statistics for each of the 170 subareas used to divide the province of Milan cannot be reported. Therefore, we present some statistics of risk coefficient estimates for the main covariates used in formula (13) in aggregated mode (see Table 4). It is worth noting that, for the province of Milan, the quantile of order 95% of $\hat{\rho}$ equals 0.438 and the median is 0. Considering the city of Milan, the same statistics are 0.734 and 0.661, respectively. As almost expected, crash occurrences show a stronger and more persistent spatial dependence in areas with high road density compared to less concentrated road networks.

The column titled "# estimates >0" indicates the number of times the average estimate of an area is greater than zero, thus increasing the risk of accidents. The number of road crossings, pedestrian crossings, and traffic lights is generally positive both inside and outside Milan city centre, and this is persistent in most areas. Priority roads appear to have a significant impact on the risk of accidents in the province of Milan. In the city, roads with a fixed speed limit of 50 km/h are more exposed to the risk of accidents. Additionally, the higher the building density, the higher the risk coefficient. The situation is different for the population data. One possible explanation is the lower availability of this data in the province of Milan compared to the city. The column labelled "# feature N/A" indicates the number of times that a particular feature was not available for the specific subset. In some areas, at least one population or building category could not be used. Conversely, the number of times that the posterior interval of estimates overlaps zero for the population and building variables lagged in formula (13) is moderately limited and often positive. This means that the contribution to the risk of accidents at a specific location also depends on how many people live in the neighbourhood of that location. The year effect shows a decrease in risk relative to 2016, primarily due to a slight reduction of accidents, except notably in 2018, which experienced a moderate peak in accidents. This is reflected in the decrease in values within the column labelled "# estimates >0". It is interesting to notice that the posterior interval of the year's estimates never overlaps zero, highlighting the significant contribution of this component. Further details can be found in (ISTAT - Italian National Institute of Statistics, 2021).

The last rows in Table 4 report some descriptive statistics of the crash risk estimate, i.e. $\hat{\lambda}$ from formula (14), at a street/edge level for year. Just for interpretation, the column labelled "crash risk estimate (quantile 1%)" contains the vector of the 1st quantiles of the risk estimates of each areas. The column "average" reports the corresponding weighted average. Moving from the province of Milan to the city of Milan it is clear the estimates and, correspondingly the risk, increase. The annual average probability to record one accident outside Milan equals 0.6% versus 4.59% in Milan city.

Figure 3 shows into details the distribution of the risk on the province of Milan. To facilitate the view of the map we have grouped the vector of $\hat{\lambda}$ of each areas into classes of deciles and used colours to plot each of them. Areas where it is expected a high concentration of traffic or of population seem to be coherently more exposed to the risk of accidents.

Table 4 CAR-SLX estimates

	Province of Milan (City excluded) (# areas = 132)					
	Average	Quantile 5%	Quantile 95%	# estimates > 0	# feature N/A	# posterior int overleaps 0
(Intercept)	−6.743	−10.935	−4.321	—	—	—
ρ	0.029	0 (median=0.0)	0.438 (max=0.7613)	–	–	–
vertical signals of danger or warning	−0.206	−2.648	1.231	44	5	22
Road_class_2	−0.149	−2.283	1.220	17	83	12
Road_class_3	−0.239	−2.306	1.559	33	39	20
Road_class_4	−0.507	−2.282	0.592	18	10	28
Road_class_5	−0.738	−3.068	0.027	7	5	51
# crosses	0.169	−0.082	0.539	104	5	9
# pedestrian crossings	0.274	−1.940	2.098	70	5	26
# traffic lights	0.689	−2.116	2.310	92	9	8
Urban area	0.363	−0.820	2.087	88	5	15
Shape_points	−0.795	−2.652	0.359	18	5	3
Lane_Cat_2	−0.235	−2.683	0.710	34	10	31
Lane_Cat_3	−0.043	−2.491	4.627	11	90	12
Roundabout	−0.224	−1.896	1.362	25	5	52
Priority_Road	0.921	−0.025	3.075	110	5	10
No_Overtaking	−0.466	−2.635	1.731	36	5	15
Stop_Sign	−0.334	−2.786	1.880	43	11	15
spd_lim_cat(30,50]	0.014	−1.060	1.001	37	5	46
spd_lim_cat(50,90]	−0.134	−1.635	1.209	33	5	41
spd_lim_cat(90,130]	−0.070	−2.340	2.745	30	64	16
Sharp_Curve	−0.513	−2.722	1.742	28	6	25
pop_class_2	0.127	−1.525	1.305	65	5	31

Table 4 continued

	Province of Milan (City excluded) (# areas = 132)					
	Average	Quantile 5%	Quantile 95%	# estimates > 0	# feature N/A	# posterior int overleaps 0
pop_class_3	0.063	−1.088	1.234	52	7	43
pop_class_4	−0.020	−1.794	1.285	44	10	34
pop_class_5	0.086	−1.862	1.702	37	44	26
bldg_class_2	0.048	−1.576	1.285	52	5	32
bldg_class_3	0.076	−1.602	1.661	52	7	33
bldg_class_4	0.108	−1.087	1.217	59	6	29
bldg_class_5	0.249	−0.738	1.872	61	12	25
pop_class_2_lag	−0.050	−2.093	2.418	47	5	31
pop_class_3_lag	−0.240	−2.174	2.346	29	7	50
pop_class_4_lag	0.358	−2.428	3.700	61	10	34
pop_class_5_lag	−0.024	−2.812	4.674	53	44	26
bldg_class_2_lag	0.186	−1.964	3.274	57	5	32
bldg_class_3_lag	−0.038	−2.491	2.082	46	7	38
bldg_class_4_lag	0.143	−2.057	1.805	52	6	45
bldg_class_5_lag	0.224	−2.194	2.804	57	12	27
Year_2017	−0.014	−1.026	0.804	62	0	0
Year_2018	0.228	−0.624	1.500	85	0	0
Year_2019	−0.224	−1.127	0.514	45	0	0
Year_2020	−0.402	−1.360	0.923	32	0	0
crash risk estimate (quantile 1%)	0.0004	8.8E-08	0.0012	–	–	–
crash risk estimate (quantile 50%)	0.0032	0.0002	0.0085	–	–	–
crash risk estimate (average)	0.0062	0.0012	0.0140	–	–	–

Table 4 continued

	Province of Milan (City excluded) (# areas = 132)					
	Average	Quantile 5%	Quantile 95%	# estimates > 0	# feature N/A	# posterior int overleaps 0
crash risk estimate (quantile 95)	0.0189	0.0043	0.0396	–	–	–
crash risk estimate (quantile 99%)	0.0507	0.0116	0.1206	–	–	–
	City of Milan (# areas = 38)					
	Average	Quantile 5%	Quantile 95%	# estimates > 0	# feature N/A	# posterior int overleaps 0
(Intercept)	−4.238	−5.731	−2.735	–	–	–
ρ	0.085	0 (median=0.6607)	0.734 (max=0.7705)	–	–	–
vertical signals of danger or warning	0.027	−0.604	0.796	17	0	4
Road_class_2	−0.142	−1.617	0.021	1	26	3
Road_class_3	0.090	−0.766	1.212	15	6	4
Road_class_4	0.115	−0.191	0.851	16	4	10
Road_class_5	−0.789	−1.262	0	1	0	2
# crosses	0.139	0.044	0.287	37	0	0
# pedestrian crossings	0.138	−1.161	1.248	20	0	6
# traffic lights	0.498	0.171	0.914	37	0	1
Urban area	0.231	0.014	0.942	20	17	1
Shape_points	−0.698	−1.133	−0.315	0	0	0
Lane_Cat_2	0.026	−0.330	0.281	19	0	7
Lane_Cat_3	−0.437	−2.024	1.519	6	9	3
Roundabout	0.022	−0.462	0.677	14	0	12
Priority_Road	0.040	−0.494	0.739	18	0	6
No_Overtaking	−0.014	−0.293	0.140	1	22	11
Stop_Sign	−0.214	−0.978	0.411	7	0	5

Table 4 continued

	City of Milan (# areas = 38)					
	Average	Quantile 5%	Quantile 95%	# estimates > 0	# feature N/A	# posterior int overleaps 0
spd_lim_cat(30,50]	0.196	−0.130	0.558	27	0	7
spd_lim_cat(50,90]	−0.406	−2.150	0.100	2	10	8
spd_lim_cat(90,130]	−0.153	−2.389	0.827	3	29	2
Sharp_Curve	−0.227	−1.318	0.740	3	8	11
pop_class_2	0.039	−0.328	0.392	15	0	11
pop_class_3	0.009	−0.276	0.260	13	0	12
pop_class_4	0.027	−0.084	0.155	13	0	16
pop_class_5	0.067	−0.008	0.207	25	0	11
bldg_class_2	−0.059	−0.280	0.302	8	0	13
bldg_class_3	0.010	−0.375	0.340	18	0	7
bldg_class_4	0.076	−0.031	0.313	20	0	15
bldg_class_5	0.079	−0.056	0.354	21	0	10
pop_class_2_lag	0.020	−1.008	0.807	13	0	15
pop_class_3_lag	−0.202	−1.652	1.155	12	0	12
pop_class_4_lag	0.242	−0.116	1.079	20	0	16
pop_class_5_lag	0.333	−0.148	1.035	24	0	11
bldg_class_2_lag	−0.050	−0.686	0.746	7	0	16
bldg_class_3_lag	−0.137	−1.107	0.396	11	0	12
bldg_class_4_lag	0.229	−0.674	1.321	19	0	15
bldg_class_5_lag	0.238	−0.261	0.810	22	0	12

Table 4 continued

	City of Milan (# areas = 38)					
	Average	Quantile 5%	Quantile 95%	# estimates > 0	# feature N/A	# posterior int overleaps 0
Year_2017	0.027	− 0.119	0.146	28	0	0
Year_2018	0.098	− 0.001	0.175	36	0	0
Year_2019	− 0.095	− 0.254	0.019	4	0	0
Year_2020	− 0.719	− 0.862	− 0.600	0	0	0
Crash risk estimate (quantile 1%)	0.0042	0.0006	0.0111	–	–	–
Crash risk estimate (quantile 50%)	0.0291	0.0108	0.0547	–	–	–
Crash risk estimate (average)	0.0482	0.0204	0.0868	–	–	–
Crash risk estimate (quantile 95)	0.1419	0.0637	0.2427	–	-	–
Crash risk estimate (quantile 99%)	0.3321	0.1667	0.6772	–	–	–

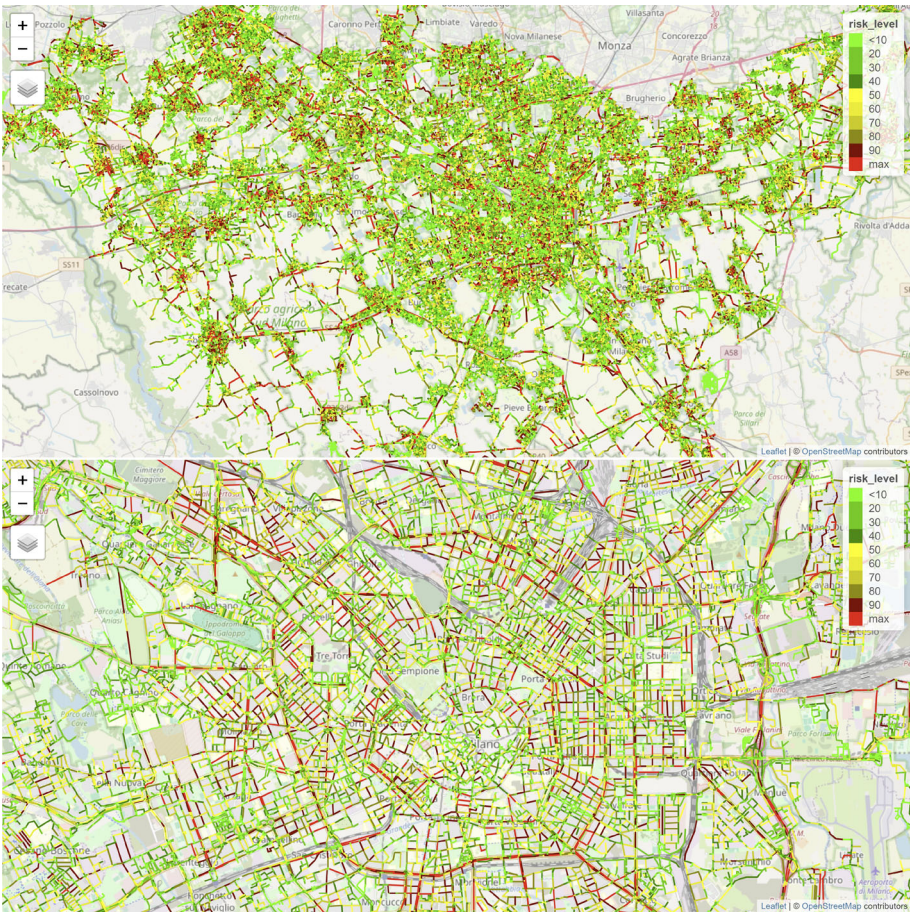

Fig. 3 (top) Map of the risk of accidents for the whole province of Milan; (bottom) Zoom on Milan city centre. In both figures, roads are coloured from the min (green) to the max (red) local risk, clustering roads in deciles. (Color figure online)

From now on we present the results at the city level. Specifically, we compare the averages of the fitted values with the observed values for each city. For Milan, we further break down the results for each ZIP code. Figure 4 shows that, on average, the model provides reliable estimates of the risk with respect to the observed claims.

Additionally, it can be observed how Milan is characterised by a higher risk expressed in terms of accidents with respect to the province. Cities typically experience higher rates of car accidents due to increased traffic density, diverse road users, varying driving speeds, and intersections, all of which contribute to a higher likelihood of collisions compared to less populated neighbourhoods.

The emphasis on these trends becomes more pronounced when examining the fitted values, as displayed in Fig. 5. Notably, all ZIP codes within Milan show a consistently higher level of risk. Moreover, cities situated in close proximity (such as Cinisello Balsamo, Cormano, Sesto San Giovanni) or located along significant thoroughfares linking to neighbouring provinces (like on the connection between Milan and Monza) exhibit relevant levels of risk. This

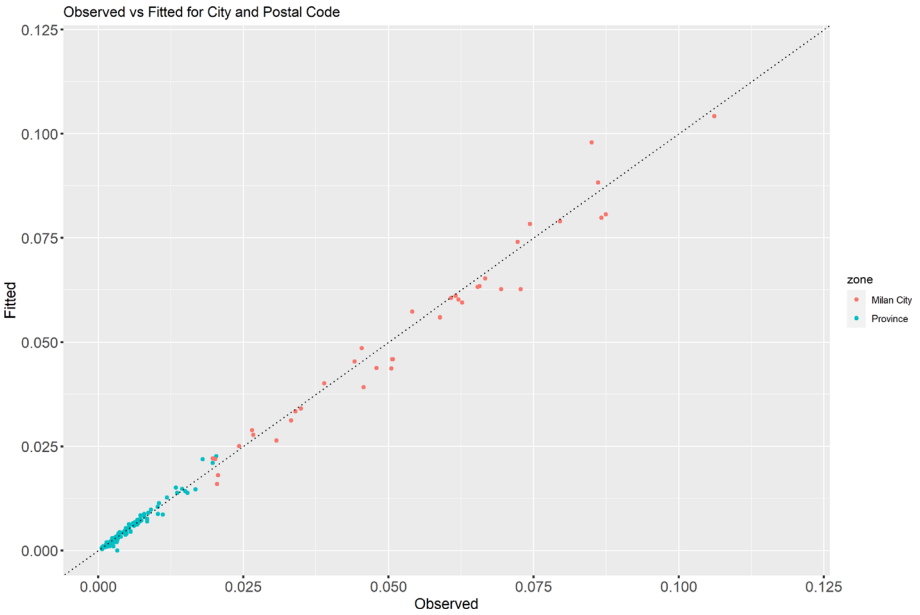

Fig. 4 Comparison between average observed and fitted values for cities. For Milan, we further break down the results for each ZIP code

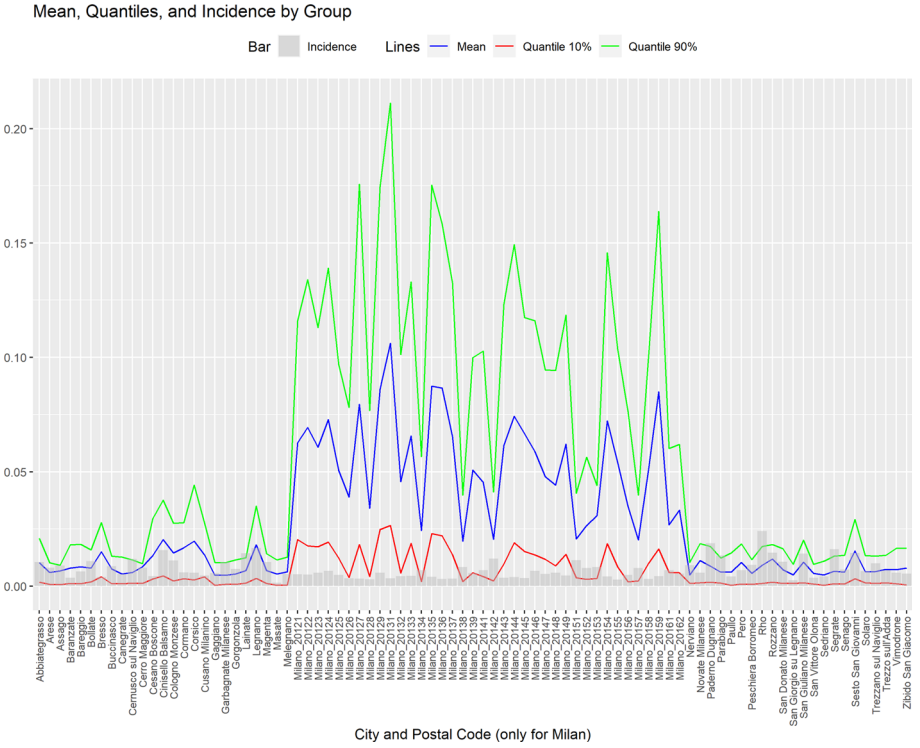

Fig. 5 Mean, quantiles at 10% and 90% and incidence in terms of number of links

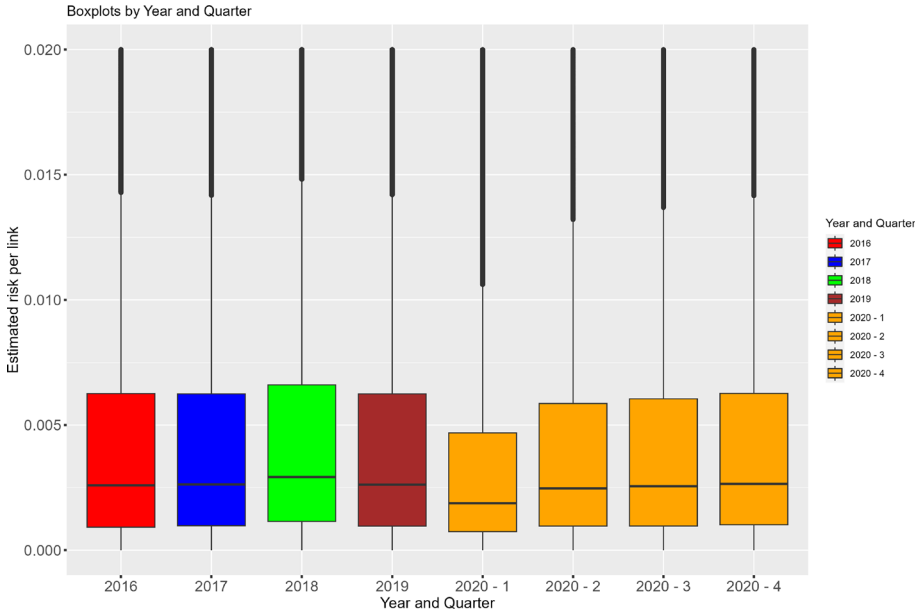

Fig. 6 Boxplot of fitted values according to year and quarter

accentuates the influence of immediate proximity to larger cities or key transportation routes on the heightened risk profiles observed in these areas.

Figure 6 analyses the yearly behaviour of the estimates. Specifically, the data for the year 2020 is dissected into quarters, providing a detailed view of the trends. It is clear that there is a significant decrease in car accidents during the first quarter, which coincided with the nationwide lockdown imposed by the Italian government on March 9, 2020. This stringent measure restricted population movement, permitting only essential outings for specific work, health, and urgent needs as a response to the escalating COVID-19 pandemic in the country. While the second quarter of 2020 showcased a marginal uptick in accidents, it remained notably lower. This period witnessed the relaxation of certain restrictions, yet factors like shifts in commuting patterns due to widespread remote work practices and a heightened emphasis on health and safety measures continued to influence the accident rates, contributing to this observed decrease despite the partial easing of lockdown measures.

Moving our attention to the intricacies and significance of street infrastructure, our analysis in Fig. 7 displays road behaviour based on their allowance for traffic movement. Meanwhile, Fig. 8 focuses on the prevalence of crossroads. These visualizations show a well defined trend: a discernible increase in accident occurrences aligning with the escalation of street complexity. Specifically, as depicted in the data, there exists a direct relationship between the complexity of roads and accident frequency. As the number of crossroads surges, so does the incidence of accidents. This phenomenon is rooted in the increased potential for collisions at these junctures, as intersections inherently pose complexity and potential hazards within traffic flow dynamics. The increased number of intersecting points elevates the complexity of navigation, thereby contributing significantly to the risk of accidents within such environments.

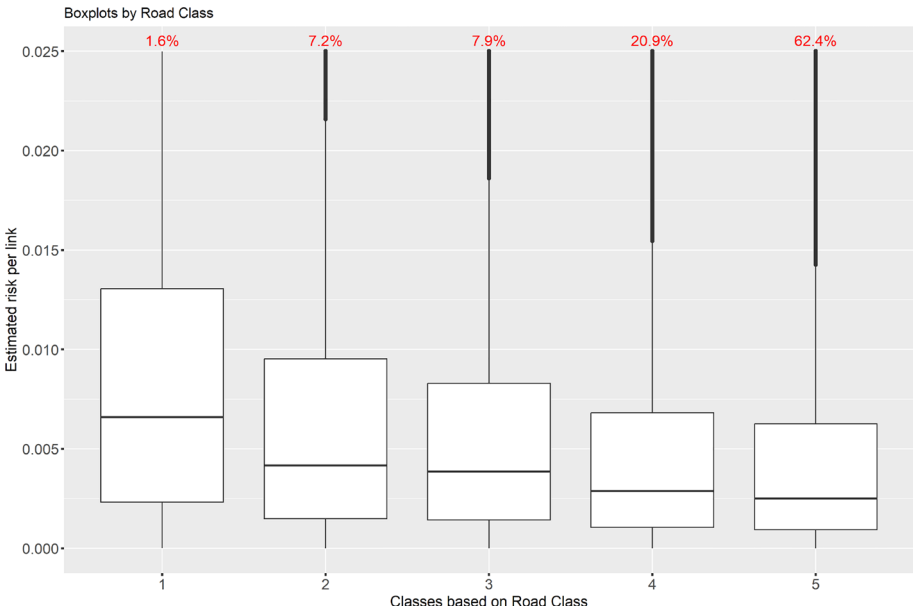

Fig. 7 Boxplot of fitted values according to the complexity/importance of the road. In red the incidence for each class (i.e. number of links over the total). Class 1 includes roads that allow for high volume, maximum speed traffic movement between and through major metropolitan areas. Class 5 is applied to roads whose volume and traffic movement are below the level of any other class. In addition, walkways, truck only roads, bus only roads, and emergency vehicle only roads receive. (Color figure online)

We focus in Fig. 9 on the effect of several dichotomous variables that have been considered in the model. We observe that the presence of specific traffic control mechanisms significantly influences the risk profile associated with different street segments. Our analysis reveals distinct risk differentials linked to the presence or absence of roundabouts and traffic lights. Roundabouts emerge as a notable factor in mitigating accident risks. Streets equipped with roundabouts exhibit an average estimate risk around 1% lower than those streets lacking this traffic control feature. The design and functionality of roundabouts, promoting continuous traffic flow and reduced collision points, contribute to this lower risk quotient. Their ability to enforce reduced speeds and encourage cautious manoeuvring diminishes the probability of severe accidents, thereby positively impacting overall safety on such road segments. Conversely, the presence of traffic lights significantly elevates the risk profile. Our estimates shows a noteworthy average risk higher than 7% in streets where traffic lights are present. Traffic lights, while crucial for regulating traffic flow and pedestrian safety, can inadvertently heighten risks due to factors such as sudden changes in signal phases, potential red-light violations, and the potential for high-speed collisions at intersections. This elevation in risk underscores the complexities and challenges associated with managing safety at signal-controlled intersections, necessitating a closer examination of strategies to mitigate these heightened risks within such environments.

Particularly noteworthy is also the impact of the " No overtaking" signal compared to the " Stop" signal on accident risk. Streets marked with the " No overtaking" signal show a risk reduction trend in comparison to those featuring the " Stop" signal. The imposition of restrictions on passing opportunities fosters safer driving conditions by limiting overtaking actions, thereby diminishing the probability of accidents along these segments. Conversely,

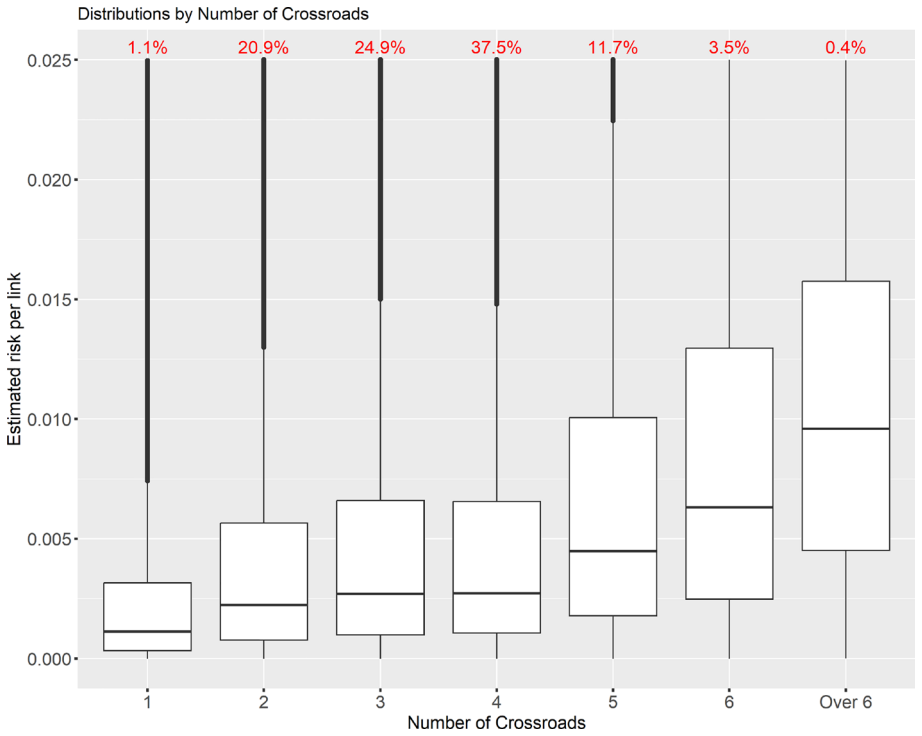

Fig. 8 Boxplot of fitted values according to number of crossroads. In red the incidence for each class (i.e. number of links over the total). (Color figure online)

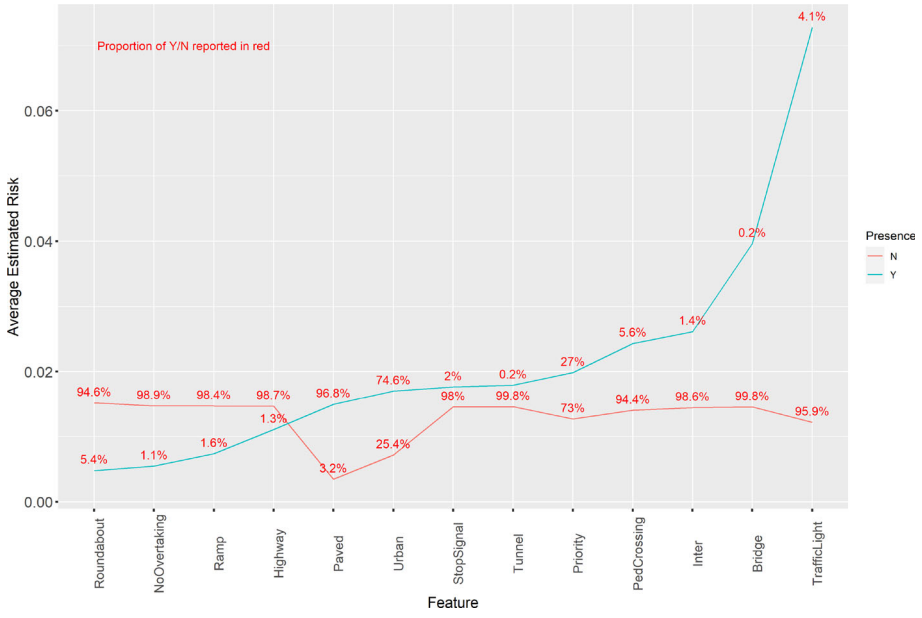

Fig. 9 Incidence and risk for dichotomous features used in the model

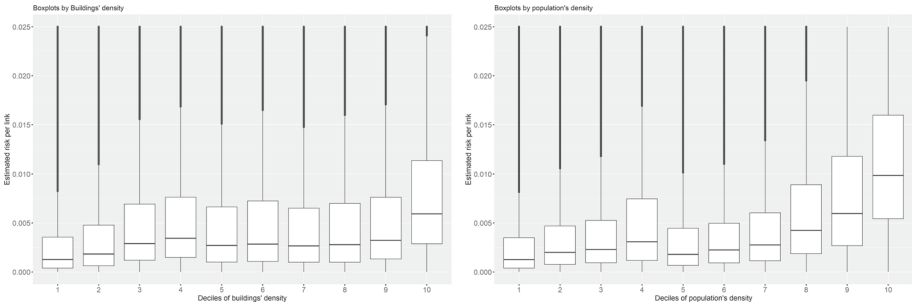

Fig. 10 Boxplot of fitted values according to building and population densities. Building and population classes are based on deciles. In red the incidence for each class (i.e. number of links over the total). (Color figure online)

streets regulated by " Stop" signals portray a comparatively higher risk profile. The imposition of these signals, while essential for controlling intersections and managing right-of-way, appears to contribute to greater risk levels. Factors such as sudden stops, intersection complexities, and potential non-compliance with stop indications may collectively elevate the risk quotient for streets governed by these signals.

The presence of pedestrian crossings seems also correlated with increased risk on roadways. This is probably due to several contributing factors, as increased interaction points, reduced visibility and concentration of vulnerable users. Indeed, pedestrian crossings introduce interaction points between vehicles and pedestrians. These locations become critical areas where differing speeds and modes of transportation intersect, heightening the risk of accidents. Drivers must constantly be vigilant for pedestrians crossing, potentially leading to sudden stops. In some cases, pedestrian crossings, especially those without proper visibility aids or in poorly lit areas, can decrease the visibility of pedestrians for drivers, increasing the likelihood of accidents, particularly during low-light conditions or adverse weather. Additionally, they often concentrate vulnerable road users in specific areas, such as children, elderly individuals, or individuals with disabilities. This concentration increases the potential severity of accidents if a collision occurs.

Also bridges, integral parts of roadways, often pose heightened risk factors contributing to potential accidents. Challenges emerge from restricted visibility around bends or inclines, hindering anticipatory actions. These narrower lanes necessitate cautious navigation, potentially leading to abrupt lane changes or limiting flexibility in maneuvering. Transitioning onto bridges can induce speed variations, impacting traffic flow and increasing collision risks. Additionally, bridges are weather-sensitive, prone to hazardous conditions like icy surfaces or high winds, further elevating accident probabilities. Complex traffic dynamics often accompany bridges, with merges, exits, or intersections nearby, fostering congestion and abrupt traffic changes. Collectively, these aspects accentuate the complexity of bridge-related driving, warranting heightened driver vigilance and careful navigation to navigate these inherent risks effectively.

We also notice in Fig. 10 how higher densities in buildings and population tend to amplify the risk of accidents due to several interrelated factors. In particular, higher densities in buildings and population contribute to a more complex and congested road environment, characterised by increased traffic volume, pedestrian activity, limited space, and intricate intersection dynamics. These factors collectively elevate the risk of accidents, necessitating heightened awareness, patience, and caution from drivers navigating such areas.

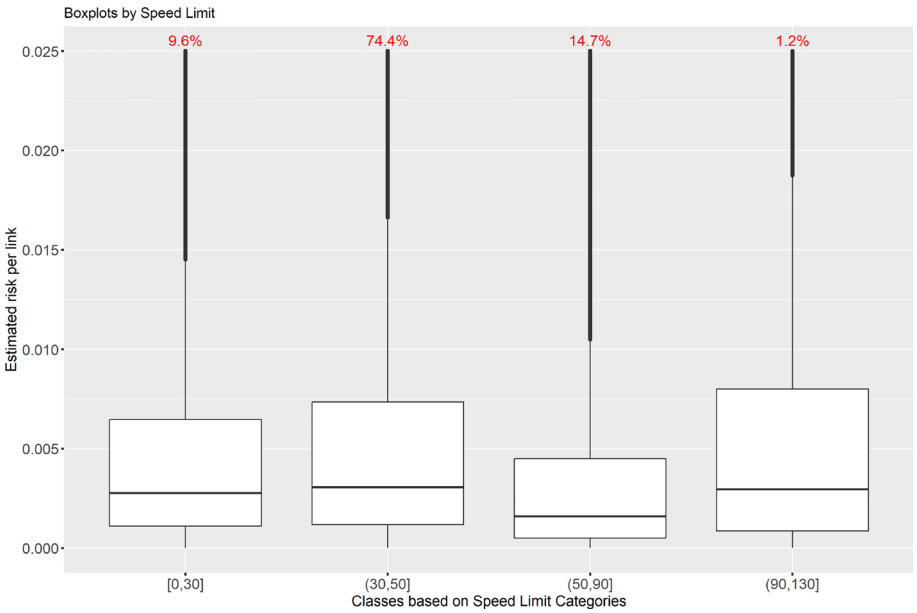

Fig. 11 Boxplot of fitted values according to speed limit categories. Speed categories are based on the maximum speed allowed in the street. In red the incidence for each class (i.e. number of links over the total). (Color figure online)

Lastly, we conclude by examining the relation between risk and limit speed categories. We observe that urban roads (i.e. class (30,50]) often exhibit higher accident risks due to increased traffic density, diverse road users (pedestrians, cyclists, motorists), frequent intersections, and varying driving speeds. The close proximity of vehicles and the complexity of navigating through city streets increase the chances of collisions, particularly at intersections or during congested periods. On the other hand, we observe that roads with very low speed limits, often at or below 30 km/h, reduce accident risks by minimizing speed differences between vehicles, providing more reaction time for drivers to respond to unexpected events, enhancing visibility and control, ensuring safety for pedestrians and cyclists, calming traffic flow, and aligning with community safety objectives in residential areas. These factors collectively create a safer driving environment and lower the probability of accidents on these roads (Fig. 11).

Highways and belt ways typically have higher speeds and fewer intersections, leading to a different set of risks. Accidents on highways often result from high-speed collisions, lane changes without proper signalling, driver fatigue, and sudden braking due to congestion or road hazards. However, highways generally have fewer points of conflict compared to urban roads, which can reduce certain types of accidents, especially those related to intersections.

5.2 GWPR results

In the previous section, we highlighted the CAR-SLX results, which provide insights into relevant features for evaluating accident risk at the area level. In this section, we concentrate on GWPR results, which allow us to separate the effects at the link level, taking into account the heterogeneity across different streets. The majority of the estimates' distributions of formula

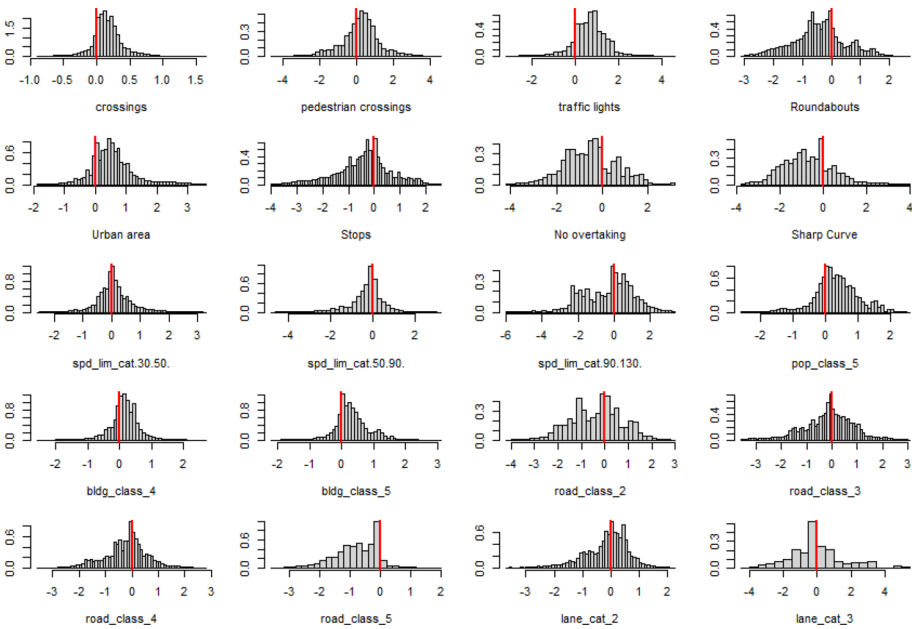

Fig. 12 Distribution of estimates at the road level for various features. Included are distributions solely for features demonstrating a non-zero modal value or displaying considerable skewness, highlighting their discernible patterns within the dataset

(13) at the road level are symmetric with respect to zero. To reduce the number of plots, Fig. 12 displays the distributions of features that have a non-zero modal value or significant skewness. At the street level, it is confirmed that the number of intersections, pedestrian crossings, and traffic lights have a positive mode. Conversely, roundabouts, stops, roads with no overtaking signals, sharp curves, and roads with lower traffic volume and movement compared to other classes are mostly in the negative domain. To explore the dependence between them, the corresponding Pearson correlation matrix is plotted in Fig. 13. Hierarchical clustering was used to group variables that naturally show clustering, and combinations with a significance value greater than 10% are hidden by a black square. Research has shown that traffic lights and pedestrian crossings can increase the risk of accidents when used together. However, in areas where speed limits are lower, such as city centres, the complexity of signals can help to reduce the risk. On the other hand, the number of crossings and roads with speeds exceeding 90 km/h can increase the risk of accidents.

The availability of estimates at a street level enables the creation of a choropleth map, providing a visual representation of the locations or areas where a feature significantly contributes to the risk. An example of this can be seen in Fig. 14. To facilitate comparison, each vector of estimates has been classified using the same classes as those reported in the legend of the figure. In the right-hand column, we provide a comparison of the spatial distribution of pedestrian crossings and traffic lights at the area level from top to bottom. From the colour intensity, it is clear that areas with traffic lights are darker than the corresponding areas with pedestrian crossings. By following the same procedure as Fig. 3, it is possible to determine where a specific feature has the most impact on an edge. As shown in Fig. 14 on the right, we can delve deeper. To account for local spatial dependence we have plot estimates at the

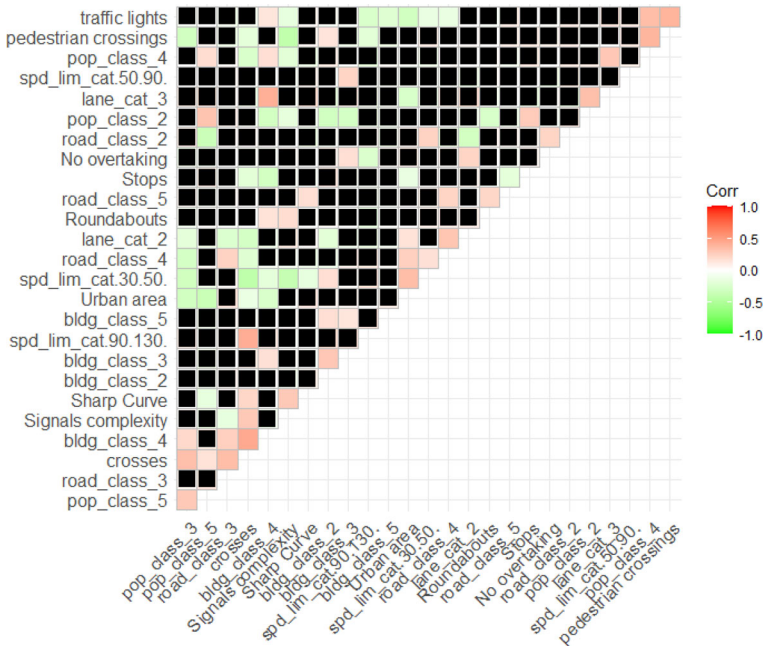

Fig. 13 Pearson correlation matrix between the vector of estimates obtained at road level for features. We have masked combinations with p-values $\geq 10\%$ using a black square to focus on the most relevant correlations

street level with a focus on the centre of Milan city. The figures reveal that certain areas on the outskirts of Milan are vulnerable to both traffic lights and pedestrian crossings.

Another interesting result concerns the analysis of the variables selected to compose the matrix $\mathbf{Q}$ of formula (22). Table 5 shows the number of times each variable has been used as a local feature. The list is sorted according to the overall total number of occurrences. The majority of links does not depend on specific local variables (referred to as "no local variables"). However, it is worth noting that the variables at the top of the list represent specific characteristics of the road, while those at the bottom of the list mostly depend on the local road network architecture. In those cases it is suggested that the risk of accidents cannot be mitigated by acting on just one component. Instead, the overall local infrastructure must be carefully considered.

6 Conclusion

This paper aims to provide a comprehensive understanding of the accident risk associated with road infrastructure, beyond statistical analysis. The objective is to furnish policy makers with necessary information to make informed decisions to reduce the societal impact of crashes. The analysis of road accident data using spatial models emphasises the need to address road safety concerns.

Drawing upon a wide literature, Conditional Autoregressive modelling and geographically weighted Poisson regression have been merged to create an innovative approach. This fusion

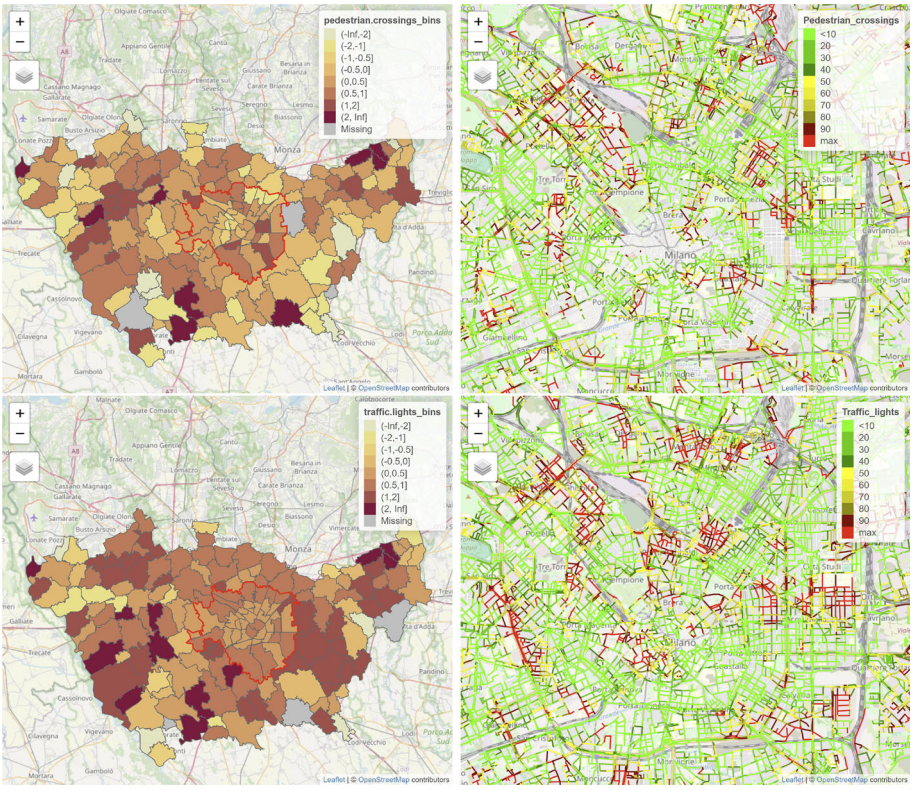

Fig. 14 Choropleth maps depicting the varying contributions of two distinct features to the risk assessment across different areas. The top panel shows the influence of pedestrian crossings, with the left side illustrating data for the province of Milan and the right side for the city of Milan. Meanwhile, the bottom panel exhibits the impact of traffic lights, again divided between the province of Milan on the left and the city of Milan on the right

was designed not only to capture overarching trends within road networks but also to unveil local nuances, allowing for a detailed understanding of accident risk factors.

The application of these models to real-world data from the city of Milan (Italy) and its province between 2016 and 2020 has yielded insightful results. Valuable insights for informed decision-making in road safety measures have been provided through the identification of key street characteristics that influence accident risk and the spatial distribution of covariates at a road level. This approach not only offers a practical solution but also sets a precedent for leveraging open data for crucial societal issues. The case study focuses on a specific area, but this does not limit the proposal's applicability. Scalability is feasible for any region with a relevant accident location database, as computational burdens are no longer a challenge. The CAR-SLX model was successfully fitted for the city of Milan, covering over 2000 km, in approximately 31 min using a standard laptop.

Therefore, the proposed approach, combining spatial modelling techniques, offers a promising way to understand the root causes of road crashes and how to mitigate their occurrences. Its effectiveness in identifying critical risk factors at a detailed level underlines its potential to guide targeted interventions and policy decisions aimed at improving road safety. The comprehensive nature of this approach, encompassing both network-wide trends

Table 5 Number of times each variable has been used as of a local feature

	Feature	Province of Milan	City of Milan	Total
1	(no local variables)	117,897	28,922	146,819
2	spd_lim_cat(90,130]	22,625	2877	25,502
3	spd_lim_cat(50,90]	9149	9970	19,119
4	road_class_5	14,036	4122	18, 158
5	road_class_3	10,030	3996	14, 026
6	Urban	6304	7246	13, 550
7	road_class_4	7208	3107	10, 315
8	Bldg_class_5	7480	2705	10, 185
9	road_class_2	7994	1559	9553
10	Priority Road	6302	2238	8540
11	Traffic lights	6381	1380	7761
12	pop_class_4	5847	1186	7033
13	Lane_cat_3	3959	2746	6705
14	pop_class_3	5961	626	6587
15	Signal Complexity	4848	1303	6151
16	pop_class_5	2138	3755	5893
17	Lane_Cat_2	5213	634	5847
18	Sharp_curve	4531	864	5395
19	spd_lim_cat(30,50]	3687	1462	5149
20	Bldg_class_4	3924	1143	5067
21	Bldg_class_2	4102	846	4948
22	Bldg_class_3	3985	687	4672
23	Pedestrian crossings	3403	994	4397
24	Roundabouts	3644	496	4140
25	pop_class_2	3423	678	4101
26	Stops	2360	800	3160
27	Shape Points	2137	710	2847
28	No overtaking	2353	238	2591
29	Crossings	1201	546	1747

The list is sorted according to the overall total number of occurrences

and local intricacies, underlines its importance in advancing the discussion on road safety analysis and policy formulation. However, although we acknowledge the significance of the dataset used in this paper, an extensive, open-source, geocoded repository of car crashes in Italy, it is crucial to recognise also its inherent limitations. The reliance on police-reported accidents implies a potential under-representation of the complete spectrum of accidents occurring in the region, as it excludes those unreported to authorities. This exclusion is believed to predominantly encompass accidents of lesser severity, not resulting in injuries or fatalities. Socially and statistically, these unreported incidents are expected to have a comparatively lower impact on individuals' lives and the healthcare system.

Indeed, it is reasonable to assume that events resulting in serious adverse outcomes, such as injuries or fatalities, are less likely to go unreported and therefore have a higher representation in the data set. Therefore, while recognising the inherent limitations, in particular the extent

of unreported events, the focus remains primarily on accidents with significant societal and individual consequences, with the aim of providing meaningful evidence to improve road safety policies.

Appendix A Official websites for car crashes location

We list some official websites with car crashes location

- **Austria** *Statistik Austria - Road Traffic Accidents:* https://www.statistik.at/en/statistics/tourism-and-transport/accidents/road-traffic-accidents
- **Belgium**
  - Statbel - Road Accidents 2021: https://statbel.fgov.be/en/open-data/road-accidents-2021
  - Statbel - Road Accidents 2022: https://statbel.fgov.be/en/open-data/road-accidents-2022
- **Denmark** StatBank Denmark: https://www.statbank.dk/20056
- **France** Data.gouv.fr - Annual Databases of Traffic Accidents:https://www.data.gouv.fr/en/datasets/bases-de-donnees-annuelles-des-accidents-corporels-de-la-circulation-routiere-annees-de-2005-a-2021/
- **Germany**
  - Destatis - Road Traffic Accidents: https://www-genesis.destatis.de/genesis/online?language=en&sequenz=statistikTabellen&selectionname=46241#abreadcrumb
  - Unfallatlas - Open Data 2022: https://unfallatlas.statistikportal.de/_opendata2022.html It is also available shape file with the correct location of all the claims occurred in Germany from 2016 to 2021.
- **Greece** Hellenic Statistical Authority: https://www.statistics.gr/en/statistics/-/publication/SDT04/-
- **Italy** Italian National Office of Statistics: https://www.istat.it/it/archivio/286933
- **Portugal** Statistics Portugal: https://www.ine.pt/xportal/xmain?xpid=INE&xpgid=ine_base_dados&bdpagenumber=1&bdnivelgeo=00&atributoordenar=null&atributoordem=null&contexto=bd&bdtemas=00&bdfreetext=accident&bdind_por_pagina=15
- **Spain** Ministry of Transport, Mobility and Urban Agenda: https://apps.fomento.gob.es/BDOTLE/inicioBD.aspx?s=4
- **Sweden** Swedish Transport Administration: https://www.trafa.se/en/road-traffic/road-traffic-injuries/
- **Switzerland** Swiss Federal Statistical Office: https://www.pxweb.bfs.admin.ch/pxweb/it/px-x-1106010100_101/px-x-1106010100_101/px-x-1106010100_101.px/
- **United Kingdom** Data.gov.uk - Road Safety Data (Updated as of 29 November 2023): https://www.data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-safety-data

Acknowledgements The authors thank the anonymous referees for their work and for the detailed comments and punctual advices that have allowed to improve the scientific quality of the paper. The authors acknowledge funding by the European Union - Next Generation EU. Project PRIN 2022 " *Building resilience to emerging risks in financial and insurance markets*". Project code: 2022FWZ2CR - CUP J53D23004560008. The views and opinions expressed are only those of the authors and do not necessarily reflect those of the European

Union or the European Commission. Neither the European Union nor the European Commission can be held responsible for them.

Author Contributions Authors contributed equally to this work

Funding Open access funding provided by Università Cattolica del Sacro Cuore within the CRUI-CARE Agreement.

Declarations

Conflict of interest The authors declare that they have no Conflict of interest.

Consent for publication All authors consent to the publication

References

Aguero-Valverde, J., & Jovanis, P. (2006). Spatial analysis of fatal injury crashes in Pennsylvania. *Accident Analysis and Prevention, 38*, 618–625. https://doi.org/10.1016/j.aap.2005.12.006

Ang, Q. W., Baddeley, A., & Nair, G. (2012). Geometrically corrected second order analysis of events on a linear network, with applications to ecology and criminology. *Scandinavian Journal of Statistics, 39*(4), 591–617.

Baddeley, A., & Nair, G. (2012). Approximating the moments of a spatial point process. *Stat, 1*(18), 30. https://doi.org/10.1002/sta4.5

Baddeley, A., Nair, G., & McSwiggan, G. (2020). Estimation of relative risk for events on a linear network. *Statistics and Computing, 30*, 469–484. https://doi.org/10.1007/s11222-019-09889-7

Baddeley, A., Nair, G., Rakshit, S., & McSwiggan, G. (2017). "Stationary" point processes are uncommon on linear networks: Point processes on linear networks. *Stat, 6*, 68–78. https://doi.org/10.1002/sta4.135

Baddeley, A., Nair, G., Rakshit, S., McSwiggan, G., & Davies, T. (2021). Analysing point patterns on networks—A review. *Spatial Statistics, 42*, 100435. https://doi.org/10.1016/j.spasta.2020.100435

Barrington-Leigh, C., & Millard-Ball, A. (2017). The world's user-generated road map is more than 80% complete. *PLoS ONE, 12*(8), 1–20. https://doi.org/10.1371/journal.pone.0180698

Barua, S., El-Basyouny, K., & Islam, M. T. (2014). A full bayesian multivariate count data model of collision severity with spatial correlation. *Analytic Methods in Accident Research, 3–4*, 28–43. https://doi.org/10.1016/j.amar.2014.09.001

Berman, M., & Turner, R. (1993). Approximating point process likelihoods with glim. *Insurance: Mathematics and Economics, 13*, 147. https://doi.org/10.1016/0167-6687(93)90845-G

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological), 36*(2), 192–236.

Bivand, R., Gómez-Rubio, V., & Rue, H. (2015). Spatial data analysis with R-INLA with some extensions. *Journal of Statistical Software, 63*(20), 1–31. https://doi.org/10.18637/jss.v063.i20

Borgoni, R., Gilardi, A., & Zappa, D. (2021). Assessing the risk of car crashes in road networks. Social Indicators Research **156**https://doi.org/10.1007/s11205-020-02295-x

Boulieri, A., Liverani, S., Hoogh, K. D., & Bliardo, M. (2016). A space-time multivariate Bayesian model to analyse road traffic accidents by severity. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 180*, 119–139. https://doi.org/10.1111/rssa.12178

Briz-Redon, A., Martinez Ruiz, F., & Montes, F. (2019). Investigation of the consequences of the modifiable areal unit problem in macroscopic traffic safety analysis: A case study accounting for scale and zoning. *Accident Analysis & Prevention, 132*, 105276. https://doi.org/10.1016/j.aap.2019.105276

Briz-Redón, A., Martínez-Ruiz, F., & Montes, F. (2019). Spatial analysis of traffic accidents near and between road intersections in a directed linear network. *Accident Analysis & Prevention, 132*, 105252. https://doi.org/10.1016/j.aap.2019.07.028

Brunsdon, C., Fotheringham, A. S., & Charlton, M. (1998). Spatial nonstationarity and autoregressive models. *Environment and Planning A, 30*(6), 957–973.

Cantaluppi, G., Giardino, G., & Zappa, D. (2023). A comparison of geospatial models for car crash risk. In: Bucci, A., et al. (eds.) *11th Scientific meeting of the SIS Group "statistics for the evaluation and quality in services"* Book of Short Papers (pp. 464–469). il Viandante, Chieti, Italia.

Choudhari, T., & Maji, A. (2019). Socio-demographic and experience factors affecting drivers' runoff risk along horizontal curves of two-lane rural highway. *Journal of Safety Research, 71*, 1–11. https://doi.org/10.1016/j.jsr.2019.09.013

Christensen, J., Wilson, A., Bastien, C., & Kayvantash, K. (2022). Efficient crash structure design for road traffic accidents of tomorrow. *International Journal of Crashworthiness, 28*, 1–20. https://doi.org/10.1080/13588265.2022.2114577

Davies, W. W. (1944). Road accidents and road structure. *Nature, 153*(3881), 330–333. https://doi.org/10.1038/153330a0

Freedman, D. A. (1999). *Ecological inference and the ecological fallacy*. https://api.semanticscholar.org/CorpusID:2810476

Gilardi, A., Borgoni, R., Presicce, L., & Mateu, J. (2023). Measurement error models for spatial network lattice data: Analysis of car crashes in Leeds. *Journal of the Royal Statistical Society Series A: Statistics in Society, 186*(3), 313–334. https://doi.org/10.1093/jrsssa/qnad057

Gilardi, A., Mateu, J., Borgoni, R., & Lovelace, R. (2022). Multivariate Hierarchical Analysis of car crashes data considering a spatial network lattice. *Journal of the Royal Statistical Society Series A: Statistics in Society, 185*(3), 1150–1177. https://doi.org/10.1111/rssa.12823

Glaser, S. (2017). *A review of spatial econometric models for count data*. https://api.semanticscholar.org/CorpusID:158877844

Golgher, A. B., & Voss, P. R. (2016). How to interpret the coefficients of spatial models: Spillovers, direct and indirect effects. *Spatial Demography, 4*, 175–205.

Gomes, M. J. T. L., Cunto, F., & da Silva, A. R. (2017). Geographically weighted negative binomial regression applied to zonal level safety performance models. *Accident Analysis & Prevention, 106*, 254–261. https://doi.org/10.1016/j.aap.2017.06.011

Gómez-Rubio, V., Bivand, R. S., & Rue, H. (2020). Bayesian model averaging with the integrated nested laplace approximation. *Econometrics, 8*(2), 23. https://doi.org/10.3390/econometrics8020023

Gómez-Rubio, V., Bivand, R. S., & Rue, H. (2021). Estimating spatial econometrics models with integrated nested laplace approximation. *Mathematics, 9*(17), 2044.

Greibe, P. (2003). Accident prediction models for urban roads. *Accident Analysis & Prevention, 35*(2), 273–285. https://doi.org/10.1016/S0001-4575(02)00005-2

Gschloßl, S., & Czado, C. (2007). Spatial modelling of claim frequency and claim size in non-life insurance. *Scandinavian Actuarial Journal*. https://doi.org/10.1080/03461230701414764

Hodges, J., Carlin, B., & Fan, Q. (2003). On the precision of the conditionally autoregressive prior in spatial models. *Biometrics, 59*, 317–22. https://doi.org/10.1111/1541-0420.00038

ISTAT - Italian National Institute of Statistics: Road accidents data. (2021). Data on road accidents for the year 2021. https://www.istat.it/it/files//2022/07/REPORT_INCIDENTI_STRADALI_2021_EN.pdf

Jensen, J. L., & Møller, J. (1991). Pseudolikelihood for exponential family models of spatial point processes. *The Annals of Applied Probability, 1*(3), 445–461.

Kılıç, B., Hacar, M., & Gülgen, F. (2023). Effects of reverse geocoding on openstreetmap tag quality assessment. *Transactions in GIS*. https://doi.org/10.1111/tgis.13089

Korn, U. (2021). A simple method for modeling changes over time. *Variance, 14*(1), 1–13.

Lindgren, F., & Rue, H. (2015). Bayesian spatial modelling with R-INLA. *Journal of Statistical Software*, *63*(19), 1–25. https://doi.org/10.18637/jss.v063.i19

Lord, D., & Mannering, F. (2010). The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives. *Transportation Research Part A: Policy and Practice, 44*(5), 291–305. https://doi.org/10.1016/j.tra.2010.02.001

Mackay, M. (1994). Engineering in accidents: Vehicle design and injuries. *Injury, 25*, 615–21. https://doi.org/10.1016/0020-1383(94)90037-X

McSwiggan, G. (2019). *Spatial point process methods for linear networks with applications to road accident analysis*. Technical report.

Miaou, S.-P., Song, J. J., & Mallick, B. (2003). Roadway traffic crash mapping: A space-time modeling approach. *Journal of Transportation Statistics, 6*, 33–57.

Mooney, P., & Minghini, M. (2017). *A review of OpenStreetMap data* (pp. 37–59). https://doi.org/10.5334/bbf.c

Murakami, D., Tsutsumida, N., Yoshida, T., Nakaya, T., Lu, B., & Harris, P. (2023). A linearization for stable and fast geographically weighted Poisson regression. *International Journal of Geographical Information Science, 37*, 1–22. https://doi.org/10.1080/13658816.2023.2209811

Nakaya, T., Fotheringham, A. S., Brunsdon, C., & Charlton, M. (2005). Geographically weighted Poisson regression for disease association mapping. *Statistics in Medicine, 24*(17), 2695–2717.

OpenStreetMap contributors. (2017). Planet dump. Retrieved from https://planet.osm.org. https://www.openstreetmap.org

Padgham, M., & Rudis, B. (2017). Osmdata: Import 'OpenStreetMap' data as simple features or spatial objects. R package version 0.1.3. https://CRAN.R-project.org/package=osmdata

Pirdavani, A., Bellemans, T., Brijs, T., & Wets, G. (2014). Application of geographically weighted regression technique in spatial analysis of fatal and injury crashes. *Journal of Transportation Engineering, 140*(8), 04014032.

R Core Team. (2022). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. R Foundation for Statistical Computing. https://www.R-project.org/

Rakshit, S., McSwiggan, G., Nair, G., & Baddeley, A. (2021). Variable selection using penalised likelihoods for point patterns on a linear network. *Australian and New Zealand Journal of Statistics, 63*(3), 417–454. https://doi.org/10.1111/anzs.12341

Savolainen, P. T., Mannering, F. L., Lord, D., & Quddus, M. A. (2011). The statistical analysis of highway crash-injury severities: A review and assessment of methodological alternatives. *Accident; Analysis and Prevention, 43*(5), 1666–1676. https://doi.org/10.1016/j.aap.2011.03.025

Smeed, R. J. (1949). Some statistical aspects of road safety research. *Journal of the Royal Statistical Society. Series A (General), 112*(1), 1–34. https://doi.org/10.2307/2984177

Tang, X., Bi, R., & Wang, Z. (2023). Spatial analysis of moving-vehicle crashes and fixed-object crashes based on multi-scale geographically weighted regression. *Accident Analysis & Prevention, 189*, 107123. https://doi.org/10.1016/j.aap.2023.107123

Tufvesson, O., Lindström, J., & Lindström, E. (2019). Spatial statistical modelling of insurance risk: a spatial epidemiological approach to car insurance. *Scandinavian Actuarial Journal, 2019*, 1–15. https://doi.org/10.1080/03461238.2019.1576146

Unterfinger, M., & Possenriede, D. (2023). hereR: 'sf'-Based Interface to the 'HERE' REST APIs. R package version 1.0.0. https://munterfi.github.io/hereR/

Vega, S. H., & Elhorst, J. (2015). The SLX model. *Journal of Regional Science, 55*, 339. https://doi.org/10.1111/jors.12188

Xu, P., & Huang, H. (2015). Modeling crash spatial heterogeneity: Random parameter versus geographically weighting. *Accident Analysis & Prevention, 75*, 16–25. https://doi.org/10.1016/j.aap.2014.10.020

Ziakopoulos, A., & Yannis, G. (2020). A review of spatial approaches in road safety. *Accident, Analysis and Prevention, 135*, 105323. https://doi.org/10.1016/j.aap.2019.105323

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B (Statistical Methodology), 67*(2), 301–320.