Measurement Error Models for Spatial Network Lattice Data: Analysis of Car Crashes in Leeds

Andrea Gilardi, Riccardo Borgoni, Luca Presicce and Jorge Mateu

Abstract

Road casualties represent an alarming concern for modern societies, especially in poor and developing countries. In the last years, several authors developed sophisticated statistical approaches to help local authorities implement new policies and mitigate the problem. These models are typically developed taking into account a set of socio-economic or demographic variables, such as population density and traffic volumes. However, they usually ignore that the external factors may be suffering from measurement errors, which can severely bias the statistical inference. This paper presents a Bayesian hierarchical model to analyse car crashes occurrences at the network lattice level taking into account measurement error in the spatial covariates. The suggested methodology is exemplified considering all road collisions in the road network of Leeds (UK) from 2011 to 2019. Traffic volumes are approximated at the street segment level using an extensive set of road counts obtained from mobile devices, and the estimates are corrected using a measurement error model. Our results show that omitting measurement error considerably worsens the model's fit and attenuates the effects of imprecise covariates.

Keywords: Bayesian Hierarchical Models, Car Crashes, GPS Traffic Devices, Network Lattice, Measurement Error, Spatial Networks

1 Introduction

According to the World Health Organisation [57, 56], the global number of road casualties is steadily increasing since 2016 and, in the last few years, reached unacceptably high level. Car crashes are the leading cause of death for children and young people aged 5-29 years and the eight cause of death in all age groups. Traffic injuries have direct social costs and indirect economical consequences (such as an adverse impact on the burden of hospitalisation or an increased health expenditure) that represent, on average, 3% of the annual GDP [58].

These problems are slightly less severe in more developed countries (mainly in Europe or North America), but the situation is still alarming considering that people from lower socio-economic backgrounds are always more likely to be involved in traffic casualties. Moreover, the burden of road accidents is disproportionately borne by the vulnerable road users (such as pedestrians, cyclists and motorcyclists), and, according to a mental health study, 39.2 percent of car crashes survivors develop post-traumatic stress disorder and require psychological assistance [12]. Hence, local and global authorities define road safety plans as an *"unfinished agenda"*, demanding for innovative approaches and evidence-based interventions [58].

The first papers were developed using the areal approach [38, 1, 7, 14], whereas the network lattice strategy gained popularity during the last years thanks to increasing computing capabilities and the rapid development of open-source spatial databases (such as Open Street Map) that provided the starting point for creating street networks at a wide range of spatial scales [6, 5, 36, 13, 15, 26]. Both frameworks employ spatial smoothing techniques to simplify the estimation process, borrowing strength from neighbouring sites. However, in this paper we adopt the network lattice approach because it provides spatially disaggregated results at the street segment level that can be more informative from a social and policy perspective. Moreover, the road infrastructure and the traffic volumes, which are key ingredients for road safety models, can be included more naturally in statistical models developed over a network lattice. We refer to Lord and Mannering [35], Savolainen et al. [46], and Ziakopoulos and Yannis [60] for more historical details, alternative modelling strategies and additional considerations.

Following the development of statistical methodologies over the years, the road safety analysts focused on the study of car crashes determinants. In fact, there is a vast literature that links traffic accidents to a variety of factors, such as vehicles characteristics [33], environmental conditions [47, 2], drivers behaviour [28], and, as already mentioned, the road design [39]. Amongst the various potential causes, road traffic is certainly of particular interest since a robust understanding of the relationship between casualties and traffic conditions is necessary to improve traffic management and reduce the crashes frequencies. A number of papers addressed this issue, typically reporting that higher traffic flows are associated with higher number of collisions [18, 30]. We refer to [53] for a recent discussion.

Road traffic is a dynamic phenomenon, evolving both in space and time [19]. It is typically difficult to obtain precise measurements of traffic volumes and, for this reason, several authors adopted alternative proxies derived from travel surveys or origin-destination tables. Nevertheless, there is an increasing evidence suggesting that traffic conditions can be accurately estimated using vehicular GPS data, i.e. moving devices acting as sensors [22, 55]. These new technologies represent a cheap and easy-to-use alternative to expensive and

time-consuming questionnaires. Moreover, they eased the collection process even at a very granular spatial domain, creating new opportunities to investigate in depth the relationship between traffic flow and road casualties. For example, Stipancic, Miranda-Moreno, and Saunier [50] and Petraki, Ziakopoulos, and Yannis [42] used GPS data to correlate collision severity and frequencies with quantitative measures of congestion derived from smartphone-collected GPS data. Following this latter approach, in this paper we approximate traffic volumes at the street segment level using estimates derived from TomTom Move service [51]. The main advantage of our workflow over the previous proposals is that TomTom collects data with global coverage using billion of GPS devices. Therefore, it provides better approximations of local traffic flows than any ad-hoc application.

Spatial data are often prone to measurement error (ME) and GPS devices represent no exception. ME can arise at different stages of the data collection process and is typically linked to various sources such as: a) instrumental imprecision in the measurement of physical attributes; b) alignment and harmonisation of characteristics recorded at different spatial scales or domains; c) unobservable effects that are only approximated by surrogate information; d) preferential sampling or incomplete observations. Considering the context analysed in this paper, the traffic volumes computed from GPS devices may suffer from ME since, usually, only a small fraction of the vehicles circulating on a road network is equipped with a GPS receiver. Therefore, the estimates could suffer from underreporting, which may not be homogeneous in all parts of the network (i.e. it might have a spatial structure).

It is well known that ignoring ME leads to severe distortions in the statistical inference. The estimate of the regression coefficient associated to the imprecise covariate can be biased downwards (attenuation) or upwards (reverse attenuation), and even the effects of error-free regressors can be distorted, where the direction of the bias depend on the correlations between the variables. Furthermore, ME may cause a loss of power for detecting signals, whereas relevant features in the data can also be masked. [17] call these effects the *"Triple Whammy of Measurement Error"*.

Nevertheless, there are only a few studies that addressed this problem, adjusting for measurement error in spatial modelling. For example, the seminal papers by [9] and [59] introduced Bayesian spatial and spatio-temporal models for disease mapping with errors in the covariates. [32] developed spatial linear mixed models for a covariate observed with error, whereas [29] proposed a semiparametric regression model to correct ME in the spatial explanatory variables. To the best of our knowledge, there are no papers in the literature that consider the problem of adjusting measurement errors in the covariates of a spatial regression model for lattice data. Given the severe implication of ME, we believe this is a relevant gap in the literature that is worth filling. Hence, the objective of this paper is to

define a statistical model to estimate the car crashes rates at a very detailed spatial resolution (i.e. the network lattice) using extensive information on traffic volumes that were derived from GPS data adjusting the statistical model to account for a spatial ME component. The suggested methodology is exemplified analysing the car accidents that occurred in the road network of Leeds (UK).

We focused on a Bayesian hierarchical approach where prior knowledge can be easily incorporated in the model. The estimation process is worked out using the Integrated Nested Laplace Approximation (INLA) [45, 34], which is an alternative to MCMC inference for the class of latent Gaussian models. Avoiding sampling, the INLA methodology can be used for efficient model fitting even in presence of large datasets and it makes prior sensitivity analysis and model comparisons more feasible. Similarly to Muff et al. [40], we adjust our modelling structure using a reformulation of the hierarchy with augmented pseudo-observations.

The rest of the paper is structured as follows. In Section 2, the data sources are presented, as well as the methods adopted to integrate them into a unique dataset. In Section 3, we introduce the modelling strategy and the classical framework for ME. Section 4 shows the main results of our analysis, whereas discussion and conclusions in Section 5 end the paper.

2 Data

The studies developed in this paper required the union of several datasets obtained from different sources. As already mentioned, we analysed car crashes that occurred in the road network of Leeds from 2011 to 2019. The city of Leeds was selected because it is the most important city in the West Yorkshire region and it accounts for approximately 40% of all car crashes in that area. The road network and the traffic volumes, which represent the spatial domain and the covariate suffering from measurement error, were obtained from TomTom Move service[1], a *"self-service product that gives direct access to the industry's largest historical traffic database"* [51]. The study area and the road network are depicted in Figure 1, where the inset map is used to geolocate the city with respect to England. Finally, we considered a set of socio-economic and demographic variables from 2011 UK Census that were recorded at the LSOA level (see below).

The datasets considered in this paper are provided by a range of different agencies and institutions. This required us some data preprocessing in order to harmonise and convert them into a usable format. Hence, in the remaining part of this section, we briefly introduce the data providers and describe the preprocessing steps that were necessary to combine these

[1]URL: https://move.tomtom.com/. Data downloaded in July 2021.

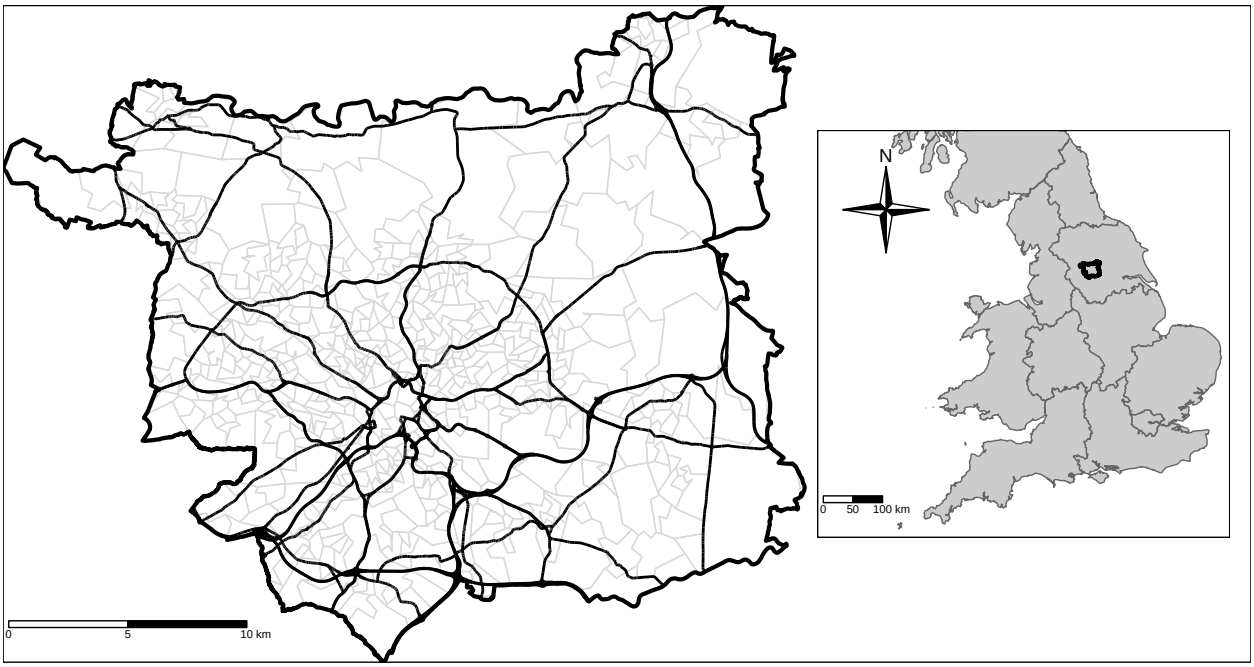


Figure 1: The black polygon denotes the geographical border of the City of Leeds (UK), while the grey polygons denote the Lower Layer Super Output Areas (LSOA) in Leeds. The black segments represent the street network adopted in this study, whereas the inset map is used to locate the study-area with respect to England.

datasets into a unique structure suitable for the statistical analysis detailed in Section 3.

2.1 Road network and traffic volumes

The road network was built using data downloaded from TomTom Move provider. Starting from 2008, TomTom collects anonymized GPS location data and gives its users access to *"the largest car-centric traffic database of more than 14 trillion anonymously collected real trip data points"*[2]. The TomTom Traffic Stats service (which is part of TomTom Move) can be used to download traffic data, either via a web portal or an API, using customised queries selecting particular geographical areas (e.g. Leeds) or time periods. It offers three types of analysis named *Route Analysis* (which returns average speed and traffic counts for a given route), *Area Analysis* (which returns average speeds, travel times, and traffic counts for all street segments in a given area), and *Traffic Density* (which is similar to *Area Analysis* but it focuses only on traffic counts). In all cases, the results are supported by a geographical database that describes the spatial dimension of the query, typically as a collection of geolocated segments.

[2]Sources: https://support.move.tomtom.com/ts-introduction/ and https://support.move.tomtom.com/products/traffic-stats/.

Hence, using the *Area Analysis* service, we downloaded the street segments that compose the most important roads in Leeds and the corresponding traffic volumes. More precisely, TomTom developed a set of rules to rank all road segments according to their importance in a transportation system using a value (named *Functional Road Class*, FRC) going from 0 (Motorways) to 8 (Least important roads). We focused our analysis on a subset of segments, selecting only those that are internally classified by TomTom as *Motorways*, *Major Roads* or *Secondary Roads*. The chosen segments represent only a subset of the complete road network but, according to our exploratory analysis, approximately 50% of all car crashes registered in Leeds from 2011 to 2019 occurred in their proximity. Additional details on the internal road classes defined by TomTom are reported at the following link: https://developer.tomtom.com/traffic-stats/support/faq/what-are-functional-road-classes-frc.

The road network downloaded from TomTom Traffic Stats was composed by 8814 segments of different lengths, ranging from 3m to 2753m with an average value of 67.5m (sd = 108m). According to TomTom documentation, the road segments have a random length and they are created internally at every location where a road attribute (e.g. road class or speed limit) changes. The road network covers approximately 595km. After downloading the raw geographic database, we removed all sets of isolated segments to simplify the estimation of the statistical model detailed in Section 3 [27, 24]. More precisely, the raw network was composed by a big group of connected segments and two isolated smaller clusters with 13 and 3 segments, respectively. After removing these two groups, we computed a (sparse) binary adjacency matrix among the segments since they represent the elementary units for the statistical model introduced below. This matrix describes the lattice connectivity and characterises the spatial dimension of the data. It is one of the key ingredients to estimate the spatial random effects characterised in the next section. The processed road network and the FRC values are depicted in Figure 2a. We notice that the spatial network spreads uniformly over the entire municipality territory. The shape of the motorway and the arterial thoroughfares reaching the city centre can also be clearly distinguished. We refer to Gilardi et al. [26] for more details on the pre-processing steps. Finally, it should be noted that, to avoid links among overlapping segments lying at different heights (e.g. bridges and underpasses), all operations detailed before were run assuming that two segments were connected if and only if they share a point in the union of their boundaries.

The most important benefit of road networks downloaded from TomTom Traffic Stats is that the provider links the geographic data to traffic counts that, according to the official documentation, are estimated at the segment level using *"anonymous signals from connected car in-dash navigation systems, portable navigation devices and anonymous GPS-equipped mobile phones"*. They are collected from *"600 million devices in use globally, generating over*

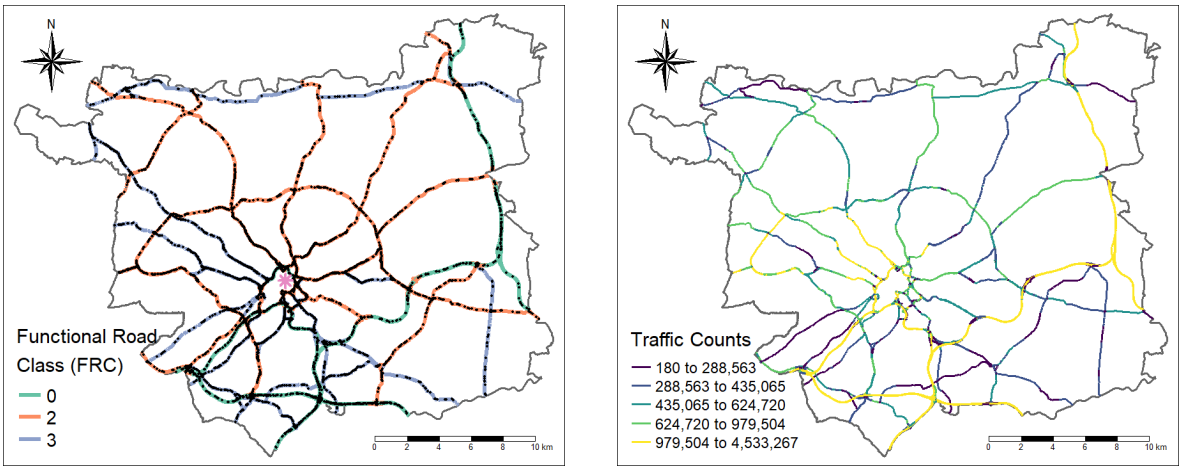
3.5 billion kilometres of GPS measurements every single day in over 80 countries" [3]. Clearly, the estimates obtained from TomTom Move suffer some underreporting being a sample of the true traffic volumes and they might be affected by spatially structured and unstructured random errors. Hence, this variable will be included in the road safety models using a spatial measurement error correction, as explained in Section 3. Figure 2b displays the estimated traffic volumes that occurred in the road network of Leeds from 1st of January 2019 to 31st of December 2019. First, we notice that the traffic counts are provided at an extremely detailed level of spatial resolution that allows us to define a measurement error model at the network lattice level. Then, we can see that the values have a large variability, ranging from 180 to, approximately, 4.5M units. Unsurprisingly, the highest values are recorded for motorways and major roads, indicating a high correlation level between the two thematic values (i.e. FRC classes and TomTom counts). Finally, due to traffic data availability, we developed our analysis considering only one year of historical data. However, it should be considered that annual traffic flows are reasonably stable, hence we do not expect any major difference if average values calculated over a larger period, as the one considered for accident data, were included in the statistical model [14].

2.2 Car crashes data

In this paper, we focused on the car crashes that occurred between the 1st of January 2011 and the 31st of December of 2019 in the road network of Leeds. More precisely, starting from a database[4] shared by the Department for Transport (DfT) that contains all geo-located traffic collisions in England during the last years, we filtered only the events that occurred inside the polygon of Leeds. Furthermore, given that the TomTom network represents only a subset of the complete city network and that, according to some working papers shared by the DfT [20, 21], the car crashes locations are typically provided at 10m or less resolution, all road collisions that occurred farther than 10m from the closest street segments were excluded from the analysis since we assumed they occurred in other segments not included in the network. The final sample is composed of 7234 events, and they are reported as black dots in Figure 2a. Then, we projected all crash locations to the nearest point of the road network, counting the occurrences on all street segments. These values represent the response variable for the statistical model defined below. We end this section pointing out that the DfT database contains only road collisions that involved at least one personal injury and became known to the Police forces within thirty days of the occurrence. Moreover, the

[3]Source: https://support.move.tomtom.com/faq-data-source-quality/

[4]Data available at the following url: https://data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-safety-data. Downloaded in November 2021.



(a) Street segments and car crashes locations. The pink star denotes the city centre.

(b) Traffic counts measured by TomTom in Leeds during 2019

Figure 2: The road network downloaded from TomTom portal. (a) The FRC values describe the importance of each segment in a transportation system. The value 0 corresponds to motorways, 2 is for major roads and 3 for secondary roads. The black dots denote the location of the car crashes that occurred in the network from 2011 to 2019. (b) Segments are coloured according to the estimates of traffic volumes in 2019. We can notice there is a strong relationship between the two thematic values (i.e. FRC and traffic counts).

English Government does not enforce people to report all personal injury accidents to the police. Hence, we acknowledge that the crashes counts may suffer, to some extent, from under-reporting and we refer to Savolainen et al. [46] for an extensive discussion on this problem.

2.3 Socio-economic variables

The statistical models described in Sections 3 and 4 take into account a set of socio-economic covariates obtained from 2011 UK census. The data were downloaded from the Nomis website[5], a webpage maintained by the University of Durham which provides all the official statistics related to census and labour market. In particular, following the literature on car crashes and road safety analysis, we decided to include three variables, namely the *population density* (given by the ratio of the number of inhabitants in a given region and its area in squared metres), the *proportion of young people* (given by the ratio of the people aged 18 to 29 and the total population), and the *proportion of home workers* (given by the ratio of home workers population and total workers population). These variables measure socio-demographic factors that cannot be fully explained using road-specific covariates like the

[5]URL: www.nomisweb.co.uk. Data downloaded in July 2021.

road type.

However, the census variables obtained from Nomis webpage were recorded at the Lower Layer Super Output Areas (LSOA) level, i.e. geographical areas designed by the Office for National Statistics to improve the reporting of small area statistics and census data estimates [41]. Hence, the socio-economic covariates and the road network were spatially misaligned and they were merged using an overlay operation: we assigned to each street segment the census values of the LSOA that intersects the largest fraction of the segments. The road network and the LSOA in Leeds are depicted in Figure 1.

3 Statistical methods

This section introduces the statistical models that were developed to analyse the road casualty occurrences. We started from a three-level Bayesian hierarchical model which is presented in Section 3.1. This model does not include any measurement error correction and represents the baseline model in our analysis. Thereafter, Sections 3.2 and 3.3 define two types of extensions. The first one enhances the baseline model including a (classical) measurement error term, whereas the second one further improves the correction allowing for potential spatial dependence in the ME by including a spatially structured random effect. Finally, we briefly discuss a few techniques for comparing the three approaches and some computational details behind the estimation of ME models.

3.1 Baseline model

Let $y_i,\ i = 1, \ldots, n$ denote the number of car crashes that occurred in the ith road segment. Following a classical hypothesis in the road safety literature (see, for example, Ziakopoulos and Yannis [60, Section 3]), in the first stage of the hierarchy we assume that

$$y_i|\lambda_i \sim \text{Poisson}(e_i\lambda_i), \tag{1}$$

where λ_i represents the car crashes rate and e_i is an offset parameter that we set equal to the geographical length of each segment. As we mentioned in the previous section, the road segments have different lengths; hence, the offset values account for the fact that longer street segments are expected to have a higher collision risk than shorter ones, guaranteeing comparable rates.

The second level of the hierarchy defines a log-linear structure for λ_i. More precisely, we

assume that

$$\log(\lambda_i) = \beta_0 + \beta_x x_i + \sum_{j=1}^{p} \beta_j z_{ij} + \theta_i, \tag{2}$$

where β_0 denotes the intercept, x_i represents the unobserved true traffic volumes with parameter β_x, $\{z_{i1}, \ldots, z_{ip}\}$ is a set of error-free covariates and $\{\beta_1, \ldots, \beta_p\}$ are the corresponding coefficients. Clearly, model (2) cannot be estimated since x_i is not observed. In this paper we assume that the traffic estimates obtained from TomTom Move provider represent a proxy, say w_i, for the real traffic volumes. Hence, after substituting x_i with w_i, the log-linear structure for λ_i writes as

$$\log(\lambda_i) = \beta_0 + \beta_w w_i + \sum_{j=1}^{p} \beta_j z_{ij} + \theta_i, \tag{3}$$

where β_w denotes the coefficient associated to the proxy measure. Finally, θ_i denotes a spatially structured random effect that is modelled using an Intrinsic Conditional Auto-Regressive (ICAR) distribution [10, 11].

The ICAR distribution is a common tool to induce spatial dependence in statistical models developed for lattice data. It is is typically defined through a set of conditional distributions

$$\theta_i | \{\theta_{i'}, i' \in \partial_i\}; \tau_\theta \sim N\left(\frac{1}{m_i}\sum_{i' \in \partial_i} \theta_{i'}; \frac{1}{m_i \tau_\theta}\right), \tag{4}$$

where m_i and ∂_i represent the cardinality and the indices of the set of neighbours for unit i, respectively. As we briefly mentioned in Section 2.1, these quantities are derived from a (sparse) binary adjacency matrix, denoted by $\boldsymbol{W}$, that summarises the neighbouring structures among the street segments. Besag [10] showed that the conditional distributions in (4) yield a joint multivariate distribution that can be expressed as

$$\boldsymbol{\theta} | \tau_\theta \sim N\left(\mathbf{0}; [\tau_\theta(\boldsymbol{D} - \boldsymbol{W})]^{-1}\right), \tag{5}$$

where $\boldsymbol{D} = \text{diag}(m_1, \ldots, m_n)$. It is possible to prove that the variance-covariance matrix in Equation (5) is not positive definite, a problem that is usually solved imposing a sum-to-zero constraint on each component (i.e. each cluster of connected road segments) of the vector $\{\theta_1, \ldots, \theta_n\}$ [27]. We refer to Banerjee, Carlin, and Gelfand [4], Martínez-Beneito and Botella-Rocamora [37], and references therein for more details on the ICAR prior.

The third level completes the definition of the baseline hierarchical model specifying the prior distributions for each parameter in (2). Considering that, before estimating the model, all fixed effects were scaled to zero mean and unit variance, we assigned a $N(0, 50)$ prior to

β_0, β_w, and $\beta_j,\ j = 1, \ldots, p$ and a logGamma$(1, 5e-05)$ prior to $\log(\tau_\theta)$.

3.2 Classical measurement error model

As mentioned at the beginning of this section, the first extension improves over the baseline correcting the traffic volume estimates with a *classical measurement error model*. In general terms, the classical ME model assumes that a covariate, say x_i, can only be observed via a proxy, say w_i, such that

$$w_i = x_i + u_i, \quad i = 1, \ldots, n. \tag{6}$$

The terms $\{u_1, \ldots, u_n\}$ represent the random errors and, as in the previous case, the index n denotes the number of statistical units. The random errors are assumed to be independent and normally distributed with zero mean and precision τ_u, i.e. $u_i \overset{\text{i.i.d}}{\sim} N(0, 1/\tau_u)$. Moreover, the classical ME model also assumes that the error terms are independent of the true covariate, here denoted by x_i, and, when working in a regression setting, it is also assumed they are independent of the response variable and any other covariate included in the regression model. We refer to [17] and [16] for a thorough introduction to ME models.

In the case study presented in this paper, we assume that the traffic volumes detected using mobile devices can approximate the unobservable measurements. We also expect that the errors obtained in the estimation process are independent of the actual traffic flows, the number of casualties in each road segment (i.e. the response variable of the Bayesian hierarchical model), as well as any other road-related or demographic covariate. Hence, the classical ME model specified above provides an adequate framework for the case at hand. Moreover, following the approach in Muff et al. [40, Section 4.1], we assume that the unobserved covariate has a Gaussian distribution with precision τ_x and mean μ_i that depends on a set of predictors, say $\tilde{z}_{ij}$.

We can now present the extension of the baseline model. To account for potential ME, the baseline model in Equation (1) is enriched by two ME terms and writes as

$$\begin{cases} y_i|\lambda_i \sim \text{Poisson}(e_i\lambda_i) \\ x_i|\,\mu_i \sim N(\mu_i, 1/\tau_x) \\ w_i|x_i \sim N(0, 1/\tau_w) \end{cases}, \tag{7}$$

where the index i ranges from 1 to n. The first expression, which is usually named as *regression* or *outcome model* in the ME literature, can be interpreted as before, whereas the second and third expressions define the ME structure. They are typically referred to as *exposure model* and *error model*, respectively. In this paper, the variable x_i denotes the

(unobserved) real traffic volumes, while w_i are the volume's estimates derived from TomTom data. The parameters τ_x and τ_w represent the precisions of the two terms.

In the second level of the hierarchy, previously described by Equation (3), we assume that

$$\begin{cases} \log(\lambda_i) = \beta_0 + \beta_x x_i + \sum_{j=1}^{p} \beta_j z_{ij} + \theta_i \\ x_i = \mu_i + \varepsilon_i \text{ and } \mu_i = \alpha_0 + \sum_{j=1}^{q} \alpha_j \tilde{z}_{ij} \\ w_i = x_i + u_i \end{cases}, \tag{8}$$

where the coefficients $\{\beta_x, \beta_0, \beta_1, \ldots, \beta_p\}$ and the variables $\{x_i, z_{ij}, \theta_i\}$ are introduced above. The parameter α_0 denotes the intercept in the exposure model, $\{\alpha_1, \ldots, \alpha_q\}$ is another set of coefficients and $\{\tilde{z}_{i1}, \ldots, \tilde{z}_{iq}\}$ are the corresponding error-free variables. The terms u_i and ε_i, which are assumed to be independent of each other, denote zero-mean normally-distributed unstructured error components having precisions τ_u and τ_ε, respectively.

We can notice that the main difference between the baseline model and the first extension is that the former simply substitutes the proxy variable to the unobserved measurements, while the latter defines a precise ME model to include this approximation.

The third level of the hierarchy specifies the prior distributions for all parameters included in the regression model. In particular, following the same reasoning as before, we assigned a $N(0, 50)$ prior to β_0, β_x, $\{\beta_1, \ldots, \beta_p\}$, α_0, $\{\alpha_1, \ldots, \alpha_q\}$, and a logGamma$(1, 5e-05)$ prior to the logarithm of τ_θ. The parameters τ_u and τ_ε, which represent the uncertainty in the exposure and error components, were modelled using the so-called *Penalised Complexity* (PC) priors. PC priors were defined in Simpson et al. [48] starting from a list of desiderata (e.g. robustness or invariance regarding reparametrisations) and, as the name suggests, they penalise departures from a base model (see Definition 1 in the original paper). We adopted these priors since they provide an attractive alternative to classical weakly-informative priors and can be constructed using probability statements on the parameters space. These are extremely appealing features considering that we are working in a ME context where prior specification plays a crucial role.

In the univariate case, the PC priors can be defined by two parameters, say σ_0 and α, that control the mass in the tail of the distribution, giving an upper bound for the range of sensible values. For example, considering the precision parameter for unstructured Gaussian random effects, such as ε_i and u_i, the PC prior, denoted in the rest of the paper as $PC_\tau(\sigma_0, \alpha)$, can be induced via the following probability statement

$$\mathbb{P}\left(\frac{1}{\sqrt{\tau}} > \sigma_0\right) = \alpha \iff \mathbb{P}(\sigma > \sigma_0) = \alpha.$$

We refer to Simpson et al. [48] for more details. In this paper, considering the high correlation between traffic values and road types, displayed in Figure 2, and that all variables are scaled to zero mean and unit variance, we decided to adopt a $PC_\tau(1, 0.1)$ prior for τ_ε and $PC_\tau(2, 0.1)$ for τ_u. In these cases, we are assuming that, in the standardised scale, the values of $\sigma_\varepsilon = \frac{1}{\sqrt{\tau_\varepsilon}}$ and $\sigma_u = \frac{1}{\sqrt{\tau_u}}$ are smaller than 1 and 2, respectively.

3.3 Spatial classical measurement error model

The second extension assumes that the measurement error can also include a spatially structured component which encompasses spatial regularities that are not appropriately accounted by the measurable covariates. We will refer to this type of structure using the term *spatial classical measurement error model*. In general terms, it writes as

$$w_i = x_i + u_i + \varphi_i; \quad i = 1, \ldots, n. \tag{9}$$

Equation (9) presents an additional term with respect to Equation (6) named φ_i. It accounts for this extra source of spatial variability, adding a spatial smoothing effect to the unobserved covariate. Considering the spatial nature of the data considered in this paper, we believe that the spatial ME model provides an ideal framework for appropriately approximating the (unobserved) road traffic volumes at the segment level using data derived from mobile devices.

Under these assumptions, the model previously described by Equation (7), can be extended as follows

$$\begin{cases} \log(\lambda_i) = \beta_0 + \beta_x x_i + \sum_{j=1}^{p} \beta_j z_{ij} + \theta_i \\ x_i = \mu_i + \varepsilon_i \text{ and } \mu_i = \alpha_0 + \sum_{j=1}^{q} \alpha_j \tilde{z}_{ij} \\ w_i = x_i + u_i + \varphi_i \end{cases} . \tag{10}$$

The term φ_i denotes a spatially structured random effect that is modelled using an ICAR prior, whereas all the other parameters are interpreted as before. We adopted the same priors as in the first extensions, and we assigned a logGamma$(1, 5e-05)$ prior to $\log(\tau_\varphi)$, i.e. the logarithm of the precision of the new spatial random effect.

3.4 Bayesian estimation of measurement error models

The three models described in the previous sections were estimated using the Integrated Nested Laplace Approximation (INLA), a popular alternative to classical MCMC sampling for a particular class of models, named Latent Gaussian Models (LGM) [45]. INLA provides

an attractive framework to model spatial dependence in lattice data via random effects that can be conveniently expressed as multivariate Gaussian distributions, typically with a sparse precision matrix [44, 45, 3]. The ICAR prior represents a classical example. Moreover, [40] recently showed that the INLA approach can be adjusted to accommodate ME terms through a *"reformulation with augmented pseudo-observations and a suitable extension of the latent field"*. Hence, following their recommendations, we modified the last two models presented above slightly readjusting the second level of the hierarchical structure.

In particular, Equation (8) now writes as

$$\begin{cases} \log(\lambda_i) = \beta_0 + \beta_x x_i + \sum_{j=1}^{p} \beta_j z_{ij} + \theta_i \\ 0 = -x_i + \mu_i + \varepsilon_i \text{ and } \mu_i = \alpha_0 + \sum_{j=1}^{q} \alpha_j \tilde{z}_{ij} \\ w_i = x_i + u_i \end{cases},$$

where we introduced a set of zero pseudo-observations on the left side of the second equation. A similar approach was adopted to adjust Equation (10). Using this formulation, the response variables can be specified via a response matrix that has one column for each separate equation and $3n$ rows. We refer to [40], where more details on the computing aspects are also discussed.

Finally, the models introduced in Sections 3.1, 3.2, and 3.3 were compared using Deviance Information Criterion (DIC) [49] and Watanabe–Akaike Information Criterion (WAIC) [54, 25]. These criteria represent a generalisation of classical Akaike information criterion (AIC) to Bayesian hierarchical models. They measure the adequacy of a model penalised by the number of effective parameters. In both cases, lower values of the index suggest a better fit of the model.

4 Results

In this section, we summarise the results obtained estimating the three models detailed before. As already mentioned, given the spatial nature of the problem and the network lattice constraints we decided to adopt the computationally efficient INLA framework via the homonymous R package [34, 43] following Muff et al. [40]. Using a laptop with an AMD Ryzen 5 3500U processor with Radeon Vega Mobile Gfx 2.10 GHz, four cores and 8GB of RAM, estimating the baseline model required approximately 45 seconds, whereas estimating the first and the second extension requires approximately 7 and 10 minutes, respectively.

Table 1: Posterior means and standard deviations (in brackets) of all covariates included in the log-linear model for λ_i considering the three regression models detailed in Section 3.

	Baseline	First extension	Second extension
Intercept	−4.980	−12.070	−19.871
	(0.128)	(0.165)	(0.239)
2nd Road Class	0.437	8.097	15.614
	(0.142)	(0.186)	(0.268)
3rd Road Class	0.347	9.064	20.084
	(0.164)	(0.205)	(0.312)
Speed Limit	−0.479	−0.911	−1.490
	(0.041)	(0.048)	(0.087)
Population Density	−0.017	0.029	0.025
	(0.024)	(0.023)	(0.024)
Young Pop. Ratio	0.003	0.045	0.036
	(0.051)	(0.043)	(0.044)
Smart-working Ratio	0.096	0.109	0.115
	(0.041)	(0.034)	(0.036)
Road traffic	0.319	3.990	7.956
	(0.041)	(0.081)	(0.054)

4.1 Fixed effects

The baseline model and the two extensions were specified including a set of common covariates in the log-linear structure for λ_i. In particular, besides the road traffic measurements, we considered two structural variables obtained from TomTom network data, namely the *road class* (i.e. the FRC values described in Section 2.1) and the *speed limit* (recorded as a numeric variables with values ranging from 18 to 116 km/h), as well as three socio-economic or demographic variables, namely the *population density*, the *ratio of young residents*, and the *percentage of people working mainly at or from home*. As explained in section 2.3, the last three variables were obtained from UK 2011 Census data. Traffic counts have been found strongly related to the road class and speed limits; hence, these two latter road-specific covariates were also included as fixed effects in the exposure model (i.e. the second expression in Equation (10)). Finally, since the considered covariates were collected at very different scales, all numeric explanatory variables were standardised to zero mean and unit variances before estimating the three models.

The posterior means and standard deviations for all covariates included in the regression and exposure models are summarised in Tables 1 and 2, respectively. The coefficient associated to the road traffic variable in the regression model is reported in the last two rows of the first table. Estimates strongly indicate the importance of the ME corrections and the strength

Table 2: Posterior means and standard deviations (in brackets) of all fixed effects included in the exposure considering the two extensions.

	First extension	Second extensions
Intercept	1.873	1.934
	(0.023)	(0.023)
2nd Road Class	−2.030	−1.973
	(0.025)	(0.024)
3rd Road Class	−2.298	−2.560
	(0.028)	(0.031)
Speed Limit	0.109	0.130
	(0.008)	(0.010)

Table 3: Posterior means and standard deviations (in brackets) computed using a model that has the same structure of the baseline model excluding the covariate related to traffic volumes.

Intercept	2n Road Class	3rd Road Class	Speed limit	Pop. Dens.	Young Ratio	Smart. Ratio
−4.334	−0.238	−0.484	−0.428	−0.018	−0.001	0.096
(0.094)	(0.111)	(0.122)	(0.040)	(0.024)	(0.051)	(0.042)

of the attenuation bias, especially for the baseline model. To interpret these figures we must refer to the standard deviation of the original variable that, in this case, was found approximately equal to 700,000. For example, keeping all the other quantities of the baseline model fixed, an increment of 100,000 annual traffic units in the original scale leads to an increase of the car crashes rate equal to $exp(0.319 * 100,000/700,000) \simeq 1.046$. Similar calculations provide an increase of the estimated car crash rate equal to $exp(3.990 * 100,000/700,000) \simeq 1.768$ and to $exp(7.956 * 100,000/700,000) \simeq 3.116$, for the two extended models.

Looking at the last four rows in Table 1, we can notice that the *population density* and *proportion of young residents* covariates were not found significant in any of the three models, while *smart-working Ratio* is found only slightly positively correlated with the car crashes rates. These results are consistent among the three different specifications. On the other hand, road-specific covariates were found extremely important both in the regression and exposure models. Looking at the estimates in Table 2, we can conclude that, unsurprisingly, the motorways are the most congested roads in Leeds. In fact, the coefficients associated to the classes named as *"2"* and *"3"* (which correspond to the major and secondary roads, respectively) were found negative, highlighting that, on average, the corresponding traffic volumes are lower than the reference class (i.e. the motorways). A similar conclusion is also suggested by the *speed limit* coefficient.

Table 4: Posterior means and standard deviations (in brackets) of the hyperparameters included in the three regression models detailed in Section 3.

	Baseline	First extension	Second extension
τ_θ	0.100	0.690	0.458
	(0.006)	(0.065)	(0.024)
τ_ε		22.830	108.843
		(1.372)	(6.015)
τ_u		2.910	16.317
		(0.0489)	(0.611)
τ_φ			0.767
			(0.037)

From Table 1, we can notice that the motorways are regarded as the safest road type among the three classes. This is a quite common finding in road safety analysis [23, 31, 14, 52, 26], and it can be explained by ad-hoc prevention measures often implemented by the road planners (e.g. presence of physical dividers between opposite directions and absence of road-side obstacles). Estimates across the three specifications show some instability since, as already mentioned, the FRC variable is strongly correlated with the traffic volumes. However, the two predictors always suggest the same potential impact i.e. the higher is the road traffic coefficient, the lower is the intercept of the model since the motorways represent the reference class and are typically associated to larger flows. Also the *speed limit* covariate supports a similar conclusion: keeping all the other variables fixed, we expect less road casualties in roads with higher speed limits. Similar findings are also reported by [30]. However, we found that fitting the baseline model ignoring the traffic volumes leads to an opposite ranking regarding the safeness of the highway classes. Estimates are detailed in Table 3. These estimates point out that excluding the traffic covariate can have a confounding effect on the coefficient of the FRC variable. These results show that the inclusion of traffic volumes and the definition of a proper ME structure are crucial steps for developing an accurate road safety model.

4.2 Random effects

Table 4 summarises the posterior means and standard deviations of all hyperparameters included in the three models. The first row contains the estimates of τ_θ, i.e. the precision of the spatially structured random effect included in the regression model. This is the only hyperparameter shared by all three specifications. Comparing the values of τ_u across different models, we observe a lower spatial uncertainty in the last two cases, which is probably linked to the inclusion of the ME structure that helps explaining part of the random variation. The

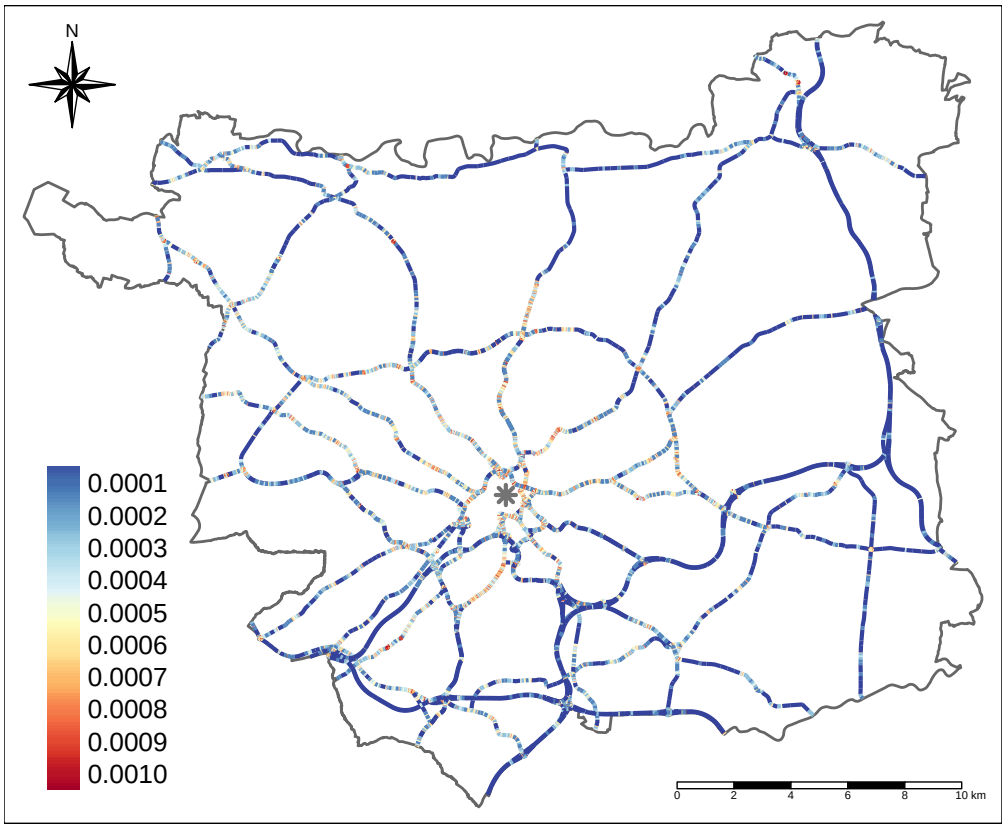


Figure 3: Choropleth map displaying the posterior means of λ_i for all street segments in the city network. The colours range from blue (lower risk) to red (higher risk). The grey star denotes the city centre.

second row reports the estimates of τ_ε, i.e. the precision of the unstructured random effect in the exposure model. Both precisions are found quite high, which is not unreasonable since the covariates were specified in the standardised scale and, as already mentioned, the traffic estimates are strongly correlated with the road types and the speed limits. In fact, these values are consistent with the considerations reported in Section 2.1 and the relationships displayed in Figure 2. Finally, the last two parameters, namely τ_u and τ_φ, denote the precisions of the unstructured and spatially structured random effects included in the error model, respectively. If we compare the posterior estimates across different extensions, we can observe the same behaviour discussed above. In fact, the third model presents a higher value of τ_u than the second one, which is probably due to the inclusion of spatially structured terms that help explaining part of the random variation.

4.3 Car crashes rates

The results reported in the previous subsections suggest that the two extensions overperform the baseline model. Hence, hereinafter we compare the first and second extensions using the DIC and WAIC criteria. We found that the inclusion of a spatial random effect in the error model greatly improves the fit (DIC = 36272.34 vs 12517.35 and WAIC = 35618.01 vs 11054.73). Therefore, we will discuss below the second extension.

We report in Figure 3 a choropleth map displaying the posterior means of λ_i, i.e. the road causality rates, for all street segments in the road network. The map was created using a palette of ten colours going from blue (lower risk) to red (higher risk). The lowest value corresponds to, approximately, one car crash every 10 kilometers while the higher value correspond to, approximately, 1 car crash every kilometer. Looking at the map, we can clearly recognise the shape of the motorways going through the city (see, e.g., Figure 2b), since the corresponding street segments are coloured as dark blue. A similar behaviour can be observed in the northern and north-eastern parts of the municipality. On the other hand, we can notice that a few segments close to the city centre (identified by a black grey star) and several arterial thoroughfares (e.g. *Scott Hall Road*, *Woodhouse Lane*, *Kristall Road*) represent the streets more prone to accidents in the municipality.

4.4 Model validation and sensitivity analysis

We validated the predictive capabilities of the selected model computing the posterior means of the car crashes rates, i.e. λ_i, and comparing these values (multiplied by the corresponding offsets) with the observed counts via a series of density curves. The results are reported in Figure 4, which clearly displays a good agreement between the two quantities, being lower car crashes frequencies associated with lower posterior means.

Finally, considering the importance of priors elicitation in a Bayesian framework, we performed a sensitivity analysis to evaluate the robustness of our results fitting the second extension model under different prior distributions for β_x, τ_ε and τ_u. We decided to focus on these parameters since they represent the key components of the ME correction. We varied one prior at a time, considering more and less diffuse specifications than the ones considered above. More precisely, we adopted a $N(0, 25)$ and $N(0, 100)$ prior for β_x, a $PC_\tau(0.5, 0.1)$ and $PC_\tau(2, 0.1)$ prior for τ_ϵ, and a $PC_\tau(1, 0.1)$ and $PC_\tau(3, 0.1)$ prior for τ_u. The results are reported in Tables 5 (fixed effects) and 6 (random effects). In both cases, the first column, denoted by index `(0)`, summarises the estimates obtained under default priors (i.e. the priors detailed in Section 3), while the other columns contain the results with the alternative specifications. Each column can be linked to the corresponding alternative prior using the

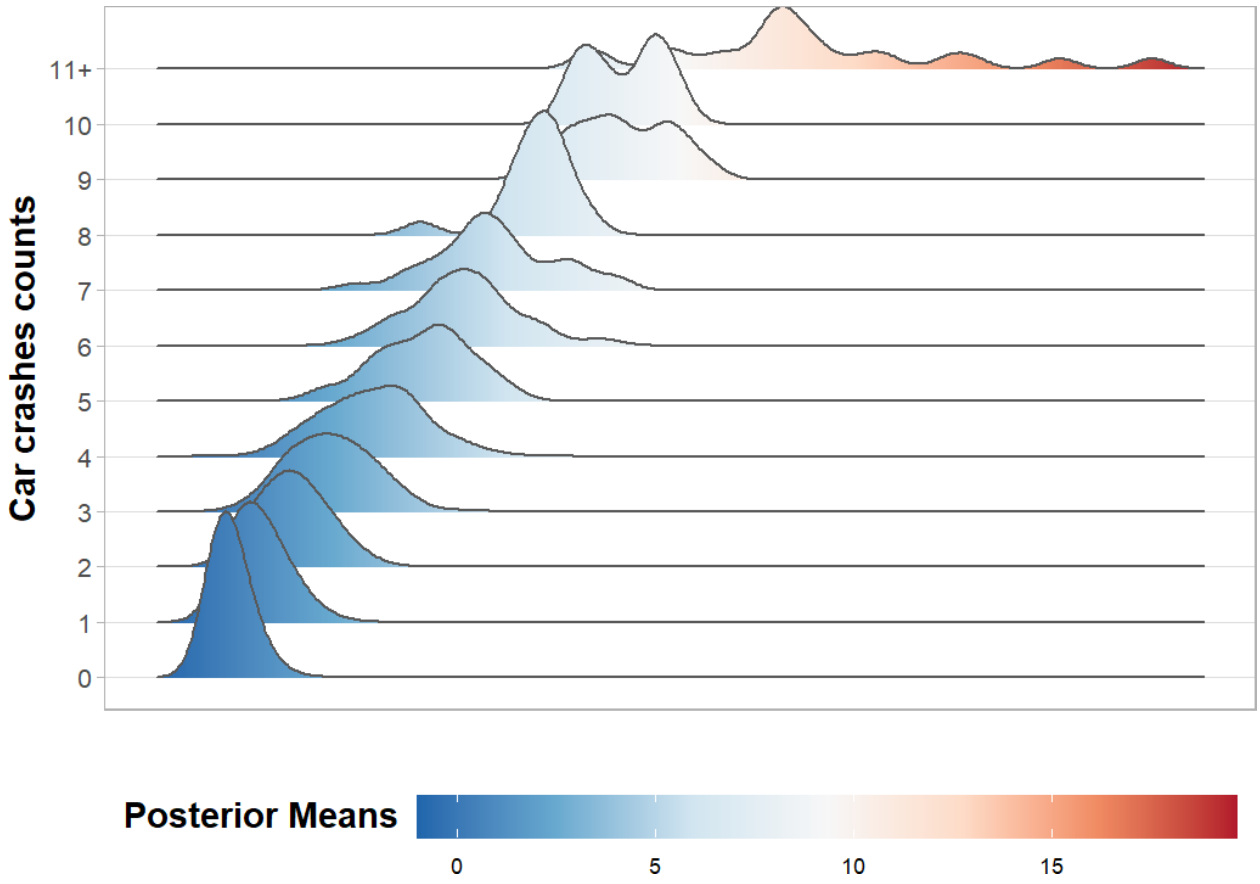


Figure 4: Density curves comparing posterior means of predicted car crashes counts and observed counts. The class `11+` summarises all street segments that registered eleven or more car crashes during the years 2011-2019. We decided to lump the last levels since they are extremely sparse in the data at end.

IDs detailed in the captions.

Columns from (3) to (6) of Table 5 show that the posterior distributions of the fixed effects are stable under different priors for τ_ε and τ_u, i.e. the precisions of the unstructured terms in the exposure and error model. Columns (1) and (2) show the results obtained by modifying the prior assigned to β_x and exhibit a slightly more erratic behaviour, underlying that the model is moderately more influenced by assumptions regarding the strength of the relationship between traffic volumes and crashes counts. This is not surprising considering the potential interactions amongst the variables associated to the traffic flows, the functional road class, and the speed limits. In fact, the same situation was also discussed when comparing the first and second extensions. Nevertheless, it should be noticed that in all cases we have got results very close to the ones obtained using the default priors.

Table 6 highlights that the last four specifications are virtually equivalent to the model with default priors also regarding the posterior distributions of the hyperparameters. More

Table 5: Posterior means and standard deviations (in brackets) of all fixed effects included in the regression model for the second extensions considering different priors than the default ones. Column `(0)` shows the results obtained using the priors detailed in Section 3, while the other columns report the results obtained considering the following priors: (1): $\beta_x \sim N(0, 10)$; (2): $\beta_x \sim N(0, 100)$; (3): $\tau_\epsilon \sim PC_\tau(2, 0.1)$; (4): $\tau_\epsilon \sim PC_\tau(0.5, 0.1)$; (5): $\tau_u \sim PC_\tau(3, 0.1)$; (6): $\tau_u \sim PC_\tau(1, 0.1)$.

	(0)	(1)	(2)	(3)	(4)	(5)	(6)
Intercept	−19.871	−18.272	−22.561	−19.754	−19.133	−20.236	−19.578
	(0.239)	(0.208)	(0.577)	(0.331)	(0.209)	(0.228)	(0.267)
2nd Road Class	15.614	13.988	19.387	15.499	14.857	15.990	15.321
	(0.268)	(0.235)	(0.605)	(0.356)	(0.239)	(0.260)	(0.293)
3rd Road Class	20.084	17.974	24.293	19.951	19.109	20.587	19.715
	(0.312)	(0.272)	(0.763)	(0.441)	(0.275)	(0.301)	(0.353)
Speed Limit	−1.490	−1.380	−1.614	−1.478	−1.438	−1.512	−1.466
	(0.087)	(0.079)	(0.116)	(0.087)	(0.084)	(0.088)	(0.085)
Population Density	0.025	0.023	0.030	0.027	0.021	0.027	0.028
	(0.024)	(0.024)	(0.024)	(0.023)	(0.024)	(0.023)	(0.023)
Young Pop. Ratio	0.036	0.035	0.040	0.040	0.032	0.040	0.040
	(0.044)	(0.044)	(0.043)	(0.043)	(0.045)	(0.043)	(0.043)
Smart-working Ratio	0.115	0.115	0.116	0.115	0.115	0.115	0.115
	(0.036)	(0.036)	(0.035)	(0.035)	(0.036)	(0.035)	(0.035)
Road traffic	7.956	6.903	9.031	7.748	7.494	7.840	7.657
	(0.054)	(0.113)	(0.363)	(0.151)	(0.079)	(0.144)	(0.118)

specifically, columns (1) and (2) point out that the estimates of the spatially structured terms, namely τ_θ and τ_φ, as well as the unstructured component of the error model, i.e. τ_u are quite stable, while there is a somewhat stronger variation for τ_ε.

5 Discussion and Conclusions

In this paper we considered the problem of adjusting measurement errors in the covariates of a spatial regression estimated at the network lattice level. We tackled this problem in a Bayesian framework comparing three increasingly complex hierarchical models. The first one completely ignores the problem of ME, the second one defines a classical ME model, whereas the last one enhances the correction with a spatially structured random effect. We exemplified the suggested methodologies analysing the distribution of all car crashes that occurred in the streets of Leeds from 2011 to 2019 using a network lattice approach. The traffic volumes, which represent a crucial component for a road safety model, were approximated using GPS data and corrected to adjust for underreporting. We found that,

Table 6: Posterior means and standard deviations (in brackets) of all hyperparameters included in the error model for the second extensions considering different priors than the default ones. The first column, denoted by index `(0)`, contains the results obtained using the priors detailed in Section 3, while the other columns report the results obtained varying the following priors: (1): $\beta_x \sim N(0, 10)$; (2): $\beta_x \sim N(0, 100)$; (3): $\tau_\epsilon \sim PC_\tau(2, 0.1)$; (4): $\tau_\epsilon \sim PC_\tau(0.5, 0.1)$; (5): $\tau_u \sim PC_\tau(3, 0.1)$; (6): $\tau_u \sim PC_\tau(0.75, 0.1)$.

	(0)	(1)	(2)	(3)	(4)	(5)	(6)
τ_θ	0.458	0.502	0.560	0.548	0.348	0.506	0.571
	(0.024)	(0.089)	(0.066)	(0.060)	(0.027)	(0.062)	(0.034)
τ_ε	108.843	83.840	143.558	100.985	109.719	102.070	95.120
	(6.015)	(4.235)	(11.006)	(6.860)	(6.754)	(5.374)	(4.002)
τ_u	16.317	16.839	15.807	17.147	16.238	16.750	17.332
	(0.611)	0.499)	(0.471)	(0.487)	(0.664)	(0.513)	(0.480)
τ_φ	0.767	0.790	0.771	0.773	0.774	0.759	0.743
	(0.037)	(0.028)	(0.027)	(0.026)	(0.033)	(0.024)	(0.024)

according to DIC and WAIC criteria, the spatial ME greatly improves over the classical framework.

Our results, which are summarised in Section 4, highlight the importance of the ME terms. In fact, Table 1 shows the posterior means and standard deviations of all fixed effects included in the regression model. Looking at the last row of the table, we can notice the severity of the attenuation bias, which is typical in presence of measurement errors. Road specific covariates, namely the road class and the speed limits, were found extremely important. From Table 2 we can deduce that, unsurprisingly, motorways register higher traffic volumes than other road types. Table 1 also highlights two important findings: a) there is a positive correlation between crashes counts and traffic volumes; b) the motorways are the safest road category. These are typical results in the road safety literature. However, we found that these results can be seriously attenuated (or even completely distorted) if one estimates the spatial regression model at the segment level ignoring a proper correction for road traffic volumes. We believe that these results are particularly important from a social perspective since they demonstrate that naive models may provide misleading guidance for policy evaluation if practitioners do not properly take into account all sources of errors.

Table 1 also suggests that the socio-demographic variables are not related with the car crashes occurrences. This is an unexpected finding, which may due to the overlay operations used to merge the census data with the street segments. In fact, the overlay creates a crude step-wise constant approximation of the socio-demographic data at the network lattice level (i.e. the same value is assigned to all segments in a given LSOA), implying that there might

be abrupt changes among neighbouring segments. There exists a vast literature on the spatial misalignment problem for polygonal areas (see, e.g., Banerjee, Carlin, and Gelfand [4, Chapter 7]), but, to the best of our knowledge, no-one extended these methods to the network framework. Model based approaches may provide a smoother approximation of the census covariates and a better understanding of their relationship with crashes counts. This aspect, despite deserving further investigations, is beyond the scope of this paper since we focused on the estimation of the relationship between car crashes and traffic counts.

Finally, we tested the predictive accuracy and the robustness of our modelling strategy exploring different hyperprior distributions for the key parameters of the ME model. We found that our results are quite robust to variations in the prior assigned to τ_u and τ_ε and slightly more sensitive on our assumptions regarding the strength of the relationship between traffic volumes and crashes counts.

The approaches presented in this manuscript can be extended in several directions. First, the temporal evolution of car crashes and traffic volumes could be considered, defining a spatio-temporal variation of the suggested models [14, 36]. Moreover, we could also examine the multivariate nature of road casualties, computing crashes counts after dividing the events according to one or more characteristics (e.g. the severity level or the number of cars involved) [6, 26]. We do not consider both extensions in the current paper since they introduce a lot of sparsity in the response variable, making the statistical inference even more challenging. These difficulties are even more pronounced when the ME model is coupled with spatial regression since the corrections introduce new parameters, adding more complexities to the estimation process. Finally, we note that there are situations where the classical ME model is not completely appropriate and other frameworks should be used instead, such as the so-called *"Berkson measurement error model"* [8]. The Berkson model could be also considered in the paradigm proposed in this paper, although the practical implementation for lattice data defined on a network requires further investigation. All these aspects are, however, beyond the scope of the present paper and material for future research.

Acknowledgements

We greatly acknowledge the DEMS Data Science Lab for supporting this work by providing computational resources.

References

[1] Jonathan Aguero-Valverde and Paul P Jovanis. "Spatial analysis of fatal and injury crashes in Pennsylvania". In: *Accident Analysis & Prevention* 38.3 (2006), pp. 618–625.

[2] Mircea-Paul Andreescu and David B Frost. "Weather and traffic accidents in Montreal, Canada". In: *Climate research* 9.3 (1998), pp. 225–230.

[3] Haakon Bakka et al. "Spatial modeling with R-INLA: A review". In: *Wiley Interdisciplinary Reviews: Computational Statistics* 10.6 (2018), e1443.

[4] Sudipto Banerjee, Bradley P Carlin, and Alan E Gelfand. *Hierarchical modeling and analysis for spatial data*. Chapman and Hall/CRC, 2015.

[5] Christopher Barrington-Leigh and Adam Millard-Ball. "The world's user-generated road map is more than 80% complete". In: *PloS one* 12.8 (2017), e0180698.

[6] Sudip Barua, Karim El-Basyouny, and Md Tazul Islam. "A full Bayesian multivariate count data model of collision severity with spatial correlation". In: *Analytic Methods in Accident Research* 3 (2014), pp. 28–43.

[7] Karim El-Basyouny and Tarek Sayed. "Urban arterial accident prediction models with spatial effects". In: *Transportation research record* 2102.1 (2009), pp. 27–33.

[8] Joseph Berkson. "Are there two regressions?" In: *Journal of the american statistical association* 45.250 (1950), pp. 164–180.

[9] L Bernadinelli et al. "Disease mapping with errors in covariates". In: *Statistics in medicine* 16.7 (1997), pp. 741–752.

[10] Julian Besag. "Spatial interaction and the statistical analysis of lattice systems". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 36.2 (1974), pp. 192–225.

[11] Julian Besag and Charles Kooperberg. "On conditional and intrinsic autoregressions". In: *Biometrika* 82.4 (1995), pp. 733–746.

[12] Edward B Blanchard et al. "Psychiatric morbidity associated with motor vehicle accidents." In: *Journal of nervous and mental disease* (1995).

[13] Riccardo Borgoni, Andrea Gilardi, and Diego Zappa. "Assessing the risk of car crashes in road networks". In: *Social Indicators Research* 156.2 (2021), pp. 429–447.

[14] Areti Boulieri et al. "A space–time multivariate Bayesian model to analyse road traffic accidents by severity". In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 180.1 (2017), pp. 119–139.

[15] Álvaro Briz-Redón, Jorge Mateu, and Francisco Montes. "Modeling accident risk at the road level through zero-inflated negative binomial models: A case study of multiple road networks". In: *Spatial Statistics* 43 (2021), p. 100503.

[16] John P Buonaccorsi. *Measurement error: models, methods, and applications*. Chapman and Hall/CRC, 2010.

[17] Raymond J Carroll et al. *Measurement error in nonlinear models: a modern perspective*. Chapman and Hall/CRC, 2006.

[18] Li-Yen Chang. "Analysis of freeway accident frequencies: negative binomial regression versus artificial neural network". In: *Safety science* 43.8 (2005), pp. 541–557.

[19] Glen M D'Este, Rocco Zito, and Michael AP Taylor. "Using GPS to measure traffic system performance". In: *Computer-Aided Civil and Infrastructure Engineering* 14.4 (1999), pp. 255–265.

[20] Department for Transport. *Instructions for the Completion of Road Accident Reports from non-CRASH Sources*. 2011. URL: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/230596/stats20-2011.pdf.

[21] Department for Transport. *STATS19 review: Final recommendations*. 2018. URL: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1001195/stats-19-review-final-report.pdf.

[22] Nour-Eddin El Faouzi, Henry Leung, and Ajeesh Kurian. "Data fusion in intelligent transportation systems: Progress and challenges–A survey". In: *Information Fusion* 12.1 (2011), pp. 4–10.

[23] Benoìt Flahaut. "Impact of infrastructure and local environment on road unsafety: Logistic modeling with spatial autocorrelation". In: *Accident Analysis & Prevention* 36.6 (2004), pp. 1055–1066.

[24] Anna Freni-Sterrantino, Massimo Ventrucci, and Håvard Rue. "A note on intrinsic conditional autoregressive models for disconnected graphs". In: *Spatial and spatio-temporal epidemiology* 26 (2018), pp. 25–34.

[25] Andrew Gelman, Jessica Hwang, and Aki Vehtari. "Understanding predictive information criteria for Bayesian models". In: *Statistics and computing* 24.6 (2014), pp. 997–1016.

[26] Andrea Gilardi et al. "Multivariate hierarchical analysis of car crashes data considering a spatial network lattice". In: *arXiv preprint arXiv:2011.12595v2* (2021).

[27] James S Hodges, Bradley P Carlin, and Qiao Fan. "On the precision of the conditionally autoregressive prior in spatial models". In: *Biometrics* 59.2 (2003), pp. 317–322.

[28] L.John Horwood and David M Fergusson. "Drink driving and traffic accidents in young people". In: *Accident Analysis & Prevention* 32.6 (2000), pp. 805–814. ISSN: 0001-4575. DOI: https://doi.org/10.1016/S0001-4575(00)00005-1. URL: https://www.sciencedirect.com/science/article/pii/S0001457500000051.

[29] Md Hamidul Huque et al. "Spatial regression with covariate measurement error: A semiparametric approach". In: *Biometrics* 72.3 (2016), pp. 678–686.

[30] Maria-Ioanna M Imprialou et al. "Re-visiting crash–speed relationships: A new perspective in crash modelling". In: *Accident Analysis & Prevention* 86 (2016), pp. 173–185.

[31] Ming–Der Li et al. "Differences in urban and rural accident characteristics and medical service utilization for traffic fatalities in less-motorized societies". In: *Journal of safety research* 39.6 (2008), pp. 623–630.

[32] Yi Li, Haicheng Tang, and Xihong Lin. "Spatial Linear Mixed Models with Covariate Measurement Errors". In: *Statistica Sinica* 19.3 (2009), pp. 1077–1093.

[33] Anders Lie et al. "The Effectiveness of Electronic Stability Control (ESC) in Reducing Real Life Crashes and Injuries". In: *Traffic Injury Prevention* 7.1 (2006), pp. 38–43.

[34] Finn Lindgren and Håvard Rue. "Bayesian Spatial Modelling with R-INLA". In: *Journal of Statistical Software* 63.19 (2015), pp. 1–25. URL: http://www.jstatsoft.org/v63/i19/.

[35] Dominique Lord and Fred Mannering. "The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives". In: *Transportation research part A: policy and practice* 44.5 (2010), pp. 291–305.

[36] Xiaoxiang Ma, Suren Chen, and Feng Chen. "Multivariate space-time modeling of crash frequencies by injury severity levels". In: *Analytic Methods in Accident Research* 15 (2017), pp. 29–40.

[37] Miguel A Martínez-Beneito and Paloma Botella-Rocamora. *Disease mapping: from foundations to multidimensional modeling*. CRC Press, 2019.

[38] Shaw-Pin Miaou, Joon Jin Song, and Bani K Mallick. "Roadway traffic crash mapping: a space-time modeling approach". In: *Journal of transportation and Statistics* 6 (2003), pp. 33–58.

[39] John Milton and Fred Mannering. "The relationship among highway geometrics, traffic-related elements and motor-vehicle accident frequencies". In: *Transportation* 25.4 (1998), pp. 395–413.

[40] Stefanie Muff et al. "Bayesian analysis of measurement error models using integrated nested Laplace approximations". In: *Journal of the Royal Statistical Society: Series C: Applied Statistics* (2015), pp. 231–252.

[41] Office for National Statistics. *A Beginner Guide to UK Geography (2021) v1.0*. 2021. URL: https://geoportal.statistics.gov.uk/documents/ons::a-beginners-guide-to-uk-geography-2021-v1-0/about.

[42] Virginia Petraki, Apostolos Ziakopoulos, and George Yannis. "Combined impact of road and traffic characteristic on driver behavior using smartphone sensor data". In: *Accident Analysis & Prevention* 144 (2020), p. 105657.

[43] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2021. URL: https://www.R-project.org/.

[44] Havard Rue and Leonhard Held. *Gaussian Markov random fields: theory and applications*. CRC press, 2005.

[45] Håvard Rue, Sara Martino, and Nicolas Chopin. "Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 71.2 (2009), pp. 319–392.

[46] Peter T Savolainen et al. "The statistical analysis of highway crash-injury severities: a review and assessment of methodological alternatives". In: *Accident Analysis & Prevention* 43.5 (2011), pp. 1666–1676.

[47] Daniel Shefer and Piet Rietveld. "Congestion and safety on highways: towards an analytical model". In: *Urban Studies* 34.4 (1997), pp. 679–692.

[48] Daniel Simpson et al. "Penalising model component complexity: A principled, practical approach to constructing priors". In: *Statistical science* (2017), pp. 1–28.

[49] David J Spiegelhalter et al. "Bayesian measures of model complexity and fit". In: *Journal of the royal statistical society: Series b (statistical methodology)* 64.4 (2002), pp. 583–639.

[50] Joshua Stipancic, Luis Miranda-Moreno, and Nicolas Saunier. "Impact of Congestion and Traffic Flow on Crash Frequency and Severity: Application of Smartphone-Collected GPS Travel Data". In: *Transportation Research Record* 2659.1 (2017), pp. 43–54. DOI: 10.3141/2659-05.

[51] TomTom. *Product: TomTom Move*. Data downloaded on the 28th of July 2021. 2021. URL: https://move.tomtom.com/.

[52] U.S. Department of Transportation. *Rural/Urban Comparison of Traffic Fatalities*. 2017. URL: https://crashstats.nhtsa.dot.gov/Api/Public/ViewPublication/812741.

[53] Peter Wagner, Ragna Hoffmann, and Andreas Leich. "Observations on the Relationship between Crash Frequency and Traffic Flow". In: *Safety* 7.1 (2021).

[54] Sumio Watanabe and Manfred Opper. "Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory." In: *Journal of machine learning research* 11.12 (2010).

[55] Dawn Woodard et al. "Predicting travel time reliability using mobile phone GPS data". In: *Transportation Research Part C: Emerging Technologies* 75 (2017), pp. 30–44.

[56] World Health Organization. *European regional status report on road safety 2019*. Licence: CC BY-NC-SA 3.0 IGO. 2020. URL: https://www.euro.who.int/en/publications/abstracts/european-regional-status-report-on-road-safety-2019.

[57] World Health Organization. *Global status report on road safety 2018*. 2018. URL: https://www.who.int/violence_injury_prevention/road_safety_status/2018/en/.

[58] World Health Organization. *Road traffic injuries*. 2021. URL: https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries.

[59] Hong Xia and Bradley P Carlin. "Spatio-temporal models with errors in covariates: mapping Ohio lung cancer mortality". In: *Statistics in medicine* 17.18 (1998), pp. 2025–2043.

[60] Apostolos Ziakopoulos and George Yannis. "A review of spatial approaches in road safety". In: *Accident Analysis & Prevention* 135 (2020), p. 105323.