ISBA 2022 World Meeting — International Society for Bayesian Analysis

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

Luca Presicce [1] Andrea Gilardi [1] Riccardo Borgoni [1] Jorge Mateu [2]

[1]Department of Economics, Management and Statistics, University of Milano - Bicocca

[2]Department of Mathematics, University Jaume I, Castellon, Spain

CONTACT ME

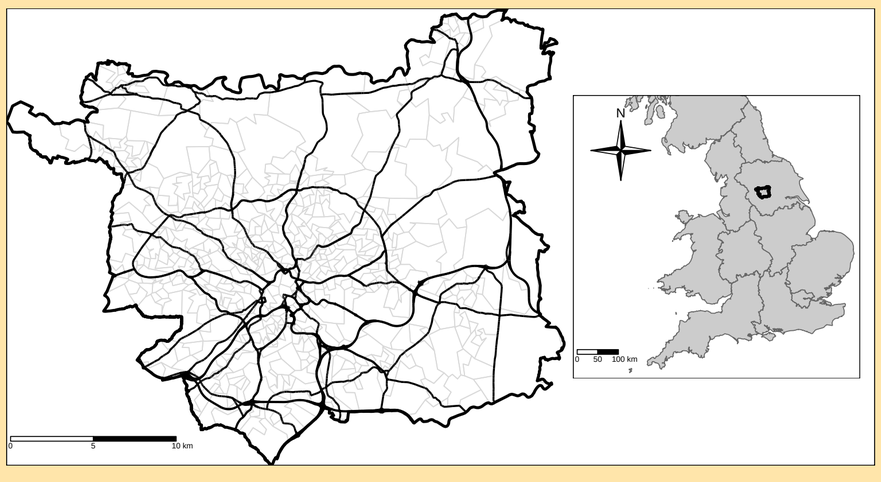
1. Introduction

We present a Bayesian hierarchical model to analyse car crashes occurrences at the network lattice level, represented by the road map of Leeds (UK), adjusting estimates for the presence of measurement error in the spatial covariates within the INLA framework.

- Traffic injuries have direct social costs and indirect economical consequences, especially in a city like Leeds that accounts for approximately **40%** of all car crashes in the West Yorkshire region.
- Spatial data are often prone to measurement error (ME) that can arise at different stages of the data collection process, such as unobservable effects that are only approximated by surrogate information or preferential sampling.
- We focus on a Bayesian hierarchical approach, where the estimation process is worked out using the Integrated Nested Laplace Approximation (INLA) and adjust the modelling structure using a reformulation of the hierarchy with augmented pseudo-observations.

2. Spatial domain

- The study area is the road network obtained from TomTom Move provider.
- This road network was treated and modelled as a spatial network structure.



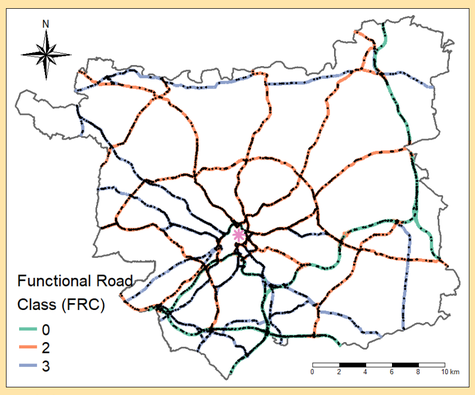
The black polygon denotes the geographical border of the City of Leeds (UK) with the street network adopted in this study, while the grey polygons denote the Lower Layer Super Output Areas (LSOA) in Leeds.

3. Data

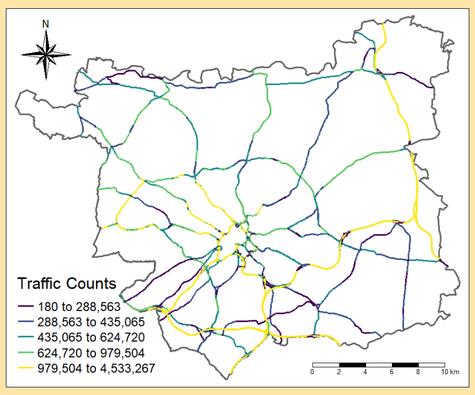
The data used in this research come from different sources.

- Context covariates (Z): socio-economic and demographic variables are from 2011 UK Census and recorded at the LSOA level. The data were downloaded from Nomis website : www.nomisweb.co.uk.
- Traffic volumes (X): the (unobservable) spatial covariate suffering from ME. It was approximated using traffic counts (W) shared by TomTom Move service and we assume this approximation suffers from unstructured and spatially structured random errors.
- Crash data (Y): in the road network of Leeds from 2011 to 2019 provided by the Department for Transport.

Since the datasets considered in this paper are provided by different agencies and institutions, some data preprocesing was necessary in order merge them into a unique usable file



(a) Black dots denote the locations of the car crashes.



(b) Traffic counts in Leeds during 2019

The FRC values describe the importance of each segment in a transportation system: 0 corresponds to motorways, 2 to major roads and 3 to secondary roads. On the right, segments are coloured according to the observed traffic volumes in 2019.

4a. Basic spatial model for crashes counts

The statistical model is a hierarchical regression model with latent gaussian components, having the structure summarised in following two section. The first stage is defined as

Outcome regression :

$$Y_i|\lambda_i \sim \text{Poisson}(e_i\lambda_i) \quad , \quad \ln\lambda = \ln e + \beta_0\mathbf{1} + \beta_x X + Z\beta_z + \nu$$

- Y number of crashes for road segment
- ν spatially structured $\mathcal{CAR}$ term such that:

$$\nu_i|\nu_{-i} \sim \mathcal{N}\left(\frac{1}{|\mathbf{N}_i|}\sum_{k\in\mathbf{N}_i}\nu_k \, , \, \frac{\tau_\nu^{-1}}{|\mathbf{N}_i|}\right)$$

- e length of road segments in meters
- X traffic volumes *(prone-error variable)*
- Z contextual covariates
- $|N_i|$ neighbourhood cardinality of segment i

4b. Further specifications and the Classical error model

The further stages are specified as follows

- Error model : $W = X + \xi + U \quad , \quad \xi \sim \mathcal{N}(0\,,\,\tau_w^{-1}) \quad , \quad u_i|u_{-i} \sim \mathcal{N}\left(\frac{1}{|\mathbf{N}_i|}\sum_{k\in\mathbf{N}_i}u_k\,,\,\frac{\tau_u^{-1}}{|\mathbf{N}_i|}\right)$
- Exposure model : $X = \alpha_0\mathbf{1} + Z\alpha_z + \varepsilon \quad , \quad \varepsilon \sim \mathcal{N}(0\,,\,\tau_x^{-1})$
- Latent random field : $\Omega = (\mathbf{x}^T, \beta_0, \beta_z^T, \alpha_0, \alpha_z^T)^T \quad , \quad \Omega \sim \mathcal{GMRF}$
- Hyperparameters : $\Theta = (\tau_w, \tau_x, \tau_\nu, \tau_u, \beta_x)^T$

Where the generic hyper parameter τ indicates the precision of a distribution, W denotes the road traffic covariate as reported by TomTom move provider.

The model has been estimated within the INLA paradigm. The results are reported in the three sections below

5. Measurement Error correction

The table below shows the posterior means and standard deviations (in brackets) associate to the traffic flow variable in three different specification of the model

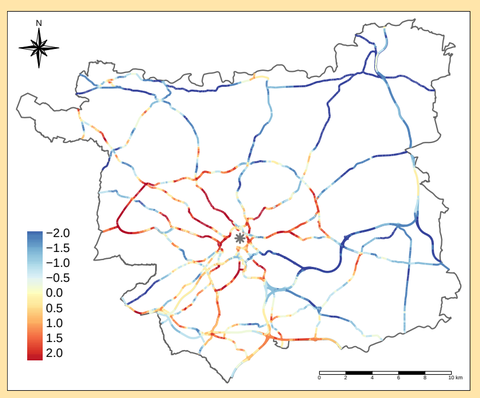
	Baseline	ME corrected	Spatial ME corrected
Traffic volumes (X)	0.319 (0.041)	3.990 (0.081)	7.956 (0.054)

Model specification (Computational time):

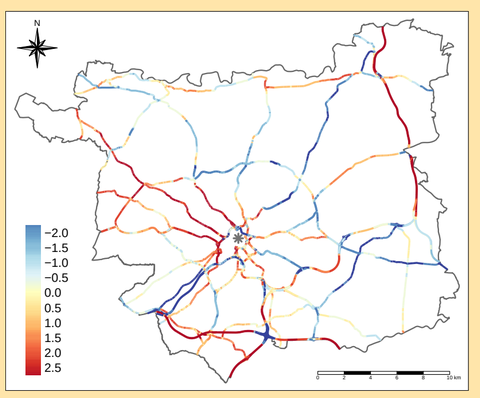
- Baseline : completely ignores the presence of ME (45 seconds).
- ME corrected : classical ME model (7 minutes).
- Spatial ME corrected : correction with a spatially structured random effect (10 minutes).

6. Random effects

The maps below reveals how the estimated posterior means of two spatial random effects ($\mathcal{CAR}$ components) in the model are distributed:



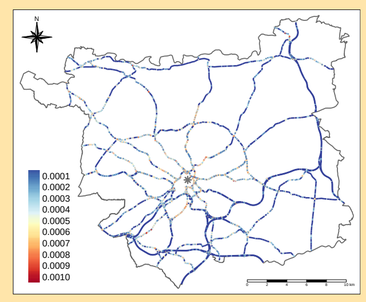
(c) Posterior means of spatial random effects in outcome model



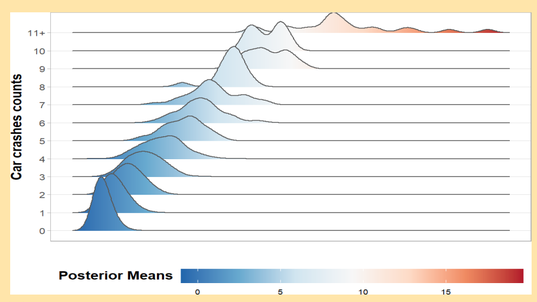
(d) Posterior means of spatial random effects in error model

7. Results and Validation

The following figures summarise the results obtained estimating the model and a validation procedure



(e) Estimated crash risk of each road segment ignoring the offset.



(f) Posterior distributions and posterior means of predicted crash counts.

8. Conclusion

- Our results highlight the importance of the ME terms as we can see in the section 5. In fact, we can notice the severity of the attenuation bias, which is typical in presence of measurement errors.
- We believe that these results are particularly important in real situations since they demonstrate that naive models may provide misleading guidance for policy evaluation if practitioners do not take into account all sources of errors properly.
- Finally, we tested the robustness of our modelling strategy exploring different hyperprior distributions for the key parameters of the ME model. We found that it is quite robust to variations in the prior assigned to τ_ξ and τ_ε and slightly more sensitive on assumptions regarding the strength of the relationship between road traffic and crashes counts β_x.

9. References

[1] A. Gilardi, J. Mateu, R. Borgoni, and R. Lovelace. Multivariate hierarchical analysis of car crashes data considering a spatial network lattice. *Journal of the Royal Statistical Society Series A (Statistics in Society) - 10.1111/rssa.12823*, 2020.

[2] S. Muff, A. Riebler, L. Held, H. Rue, and P. Saner. Bayesian analysis of measurement error models using integrated nested laplace approximations. *Journal of the Royal Statistical Society: Series C: Applied Statistics*, pages 231–252, 2015.

[3] H. Rue, S. Martino, and N. Chopin. Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):319–392, 2009.