# 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

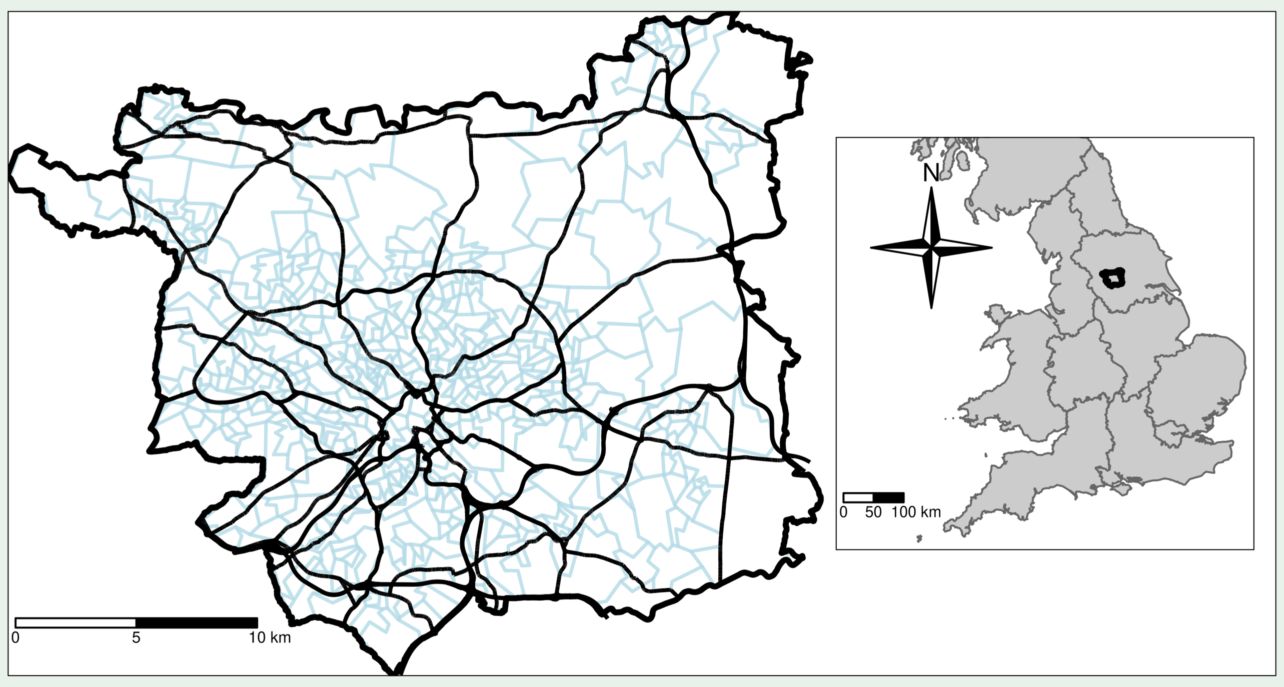
## 1. Introduction

We present a model to analyze car crash occurrences at the network lattice level, represented by the road map of Leeds (UK), adjusting the estimates for the presence of measurement error (ME) in the spatial covariates.

- Traffic injuries have direct social costs and indirect economical consequences, especially for Leeds which accounts for ≈ **40%** of all car crashes of West Yorkshire.
- ME can arise at different stages of the data collection, particularly with spatial data.
- We focus on a Bayesian hierarchical framework estimated by the Integrated Nested Laplace Approximation (INLA).

## 2. Spatial domain

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

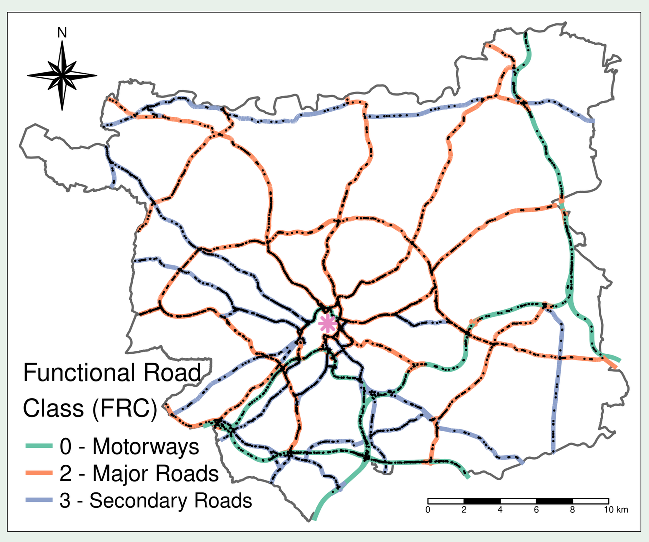



The black polygon denotes the border of Leeds (UK) with the street network adopted in this study, while the light-blue polygons denote the Lower Layer Super Output Areas (LSOA) in Leeds.

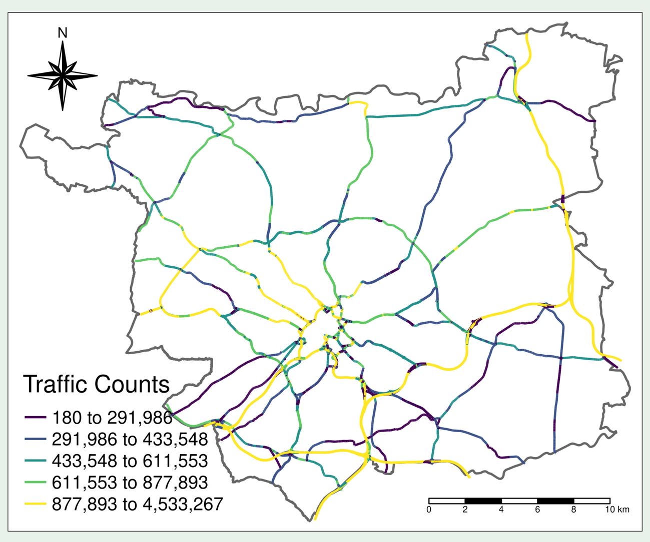
## 3. Data

The data used in this research come from different sources.

- Context covariates $(Z)$: socio-economic and demographic variables from 2011 UK Census recorded at LSOA levels. (downloaded from the Nomis website: www.nomisweb.co.uk.)
- Traffic volumes $(X)$: the (unobservable) spatial covariate suffering from ME and approximated using traffic counts $(W)$ shared by the TomTom Move service.
- Crash data $(Y)$: in the road network from 2011 to 2019 provided by the Department for Transport.



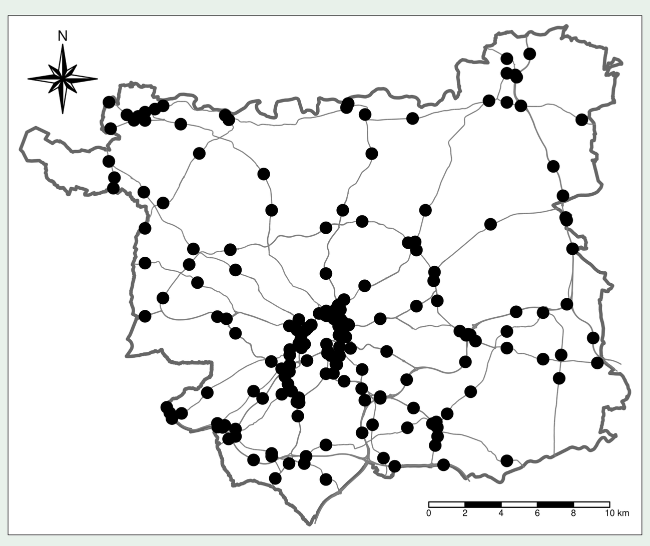

(a) Black dots denote the locations of the car crashes.



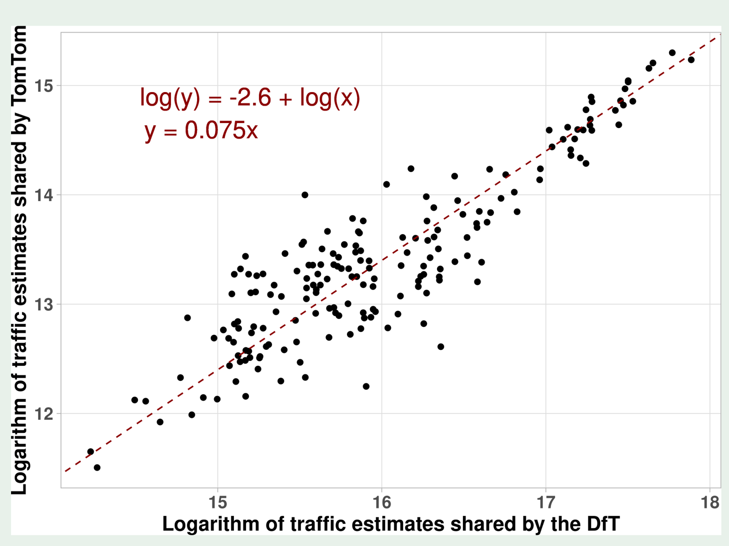

(b) Traffic counts in Leeds during 2019

## 4. Complementary traffic estimates

- The maps below display the descriptive analysis of the complementary traffic data from road control units.
- This dataset was considered for the hyperpriors specification, through reasoning on interest phenomena and its link with the TomTom data.



(a) Locations of the *count points* in Leeds.




(b) Scatterplot displaying the two sets of traffic estimates.

## 5. Statistical methods

The statistical model is a hierarchical regression with latent Gaussian components, having the structure summarised in the following three stages.

- Outcome regression :

$$y_i|\lambda_i \sim \text{Poisson}(e_i\lambda_i), \qquad i = 1, \dots, n$$

$$\ln(\lambda_i) = \beta_0 + x_i^\intercal \beta_x + z_i^\intercal \beta_z + \theta_i, \qquad i = 1, \dots, n$$

$$\theta_i | \{\theta_{i'}, i' \in \partial_i\}; \tau_\theta \sim \mathcal{N}\left(\frac{1}{|\partial_i|}\sum_{i' \in \partial_i} \theta_{i'} \ , \ \frac{\tau_\theta^{-1}}{|\partial_i|}\right), \qquad i = 1, \dots, n$$

$$\beta_0, \beta_x \sim \mathcal{N}(0, 50), \qquad \beta_z = (\beta_1, \dots, \beta_p) \sim \mathcal{N}_p(0_p, 50\mathbb{I}_p)$$

- Error model and Exposure model :

$$\ln(w_i) = \tilde{\rho}_0 + \ln(x_i) + \frac{1}{\tau_u}\left[\sqrt{1-\phi}\,\tilde{u}_i + \sqrt{\phi}\,\tilde{\nu}_i\right], \qquad i = 1, \dots, n$$

$$\ln(x_i) = \alpha_0 + \tilde{z}_i^\intercal \alpha_z + \varepsilon_i, \qquad i = 1, \dots, n$$

$$\tilde{u}_i \sim \mathcal{N}(0, 1), \quad \tilde{\nu}_i \sim \text{ICAR}(1), \qquad i = 1, \dots, n$$

$$\alpha_0 \sim \mathcal{N}(0, 50), \qquad \alpha_z = (\alpha_1, \dots, \alpha_q) \sim \mathcal{N}_q(0_q, 50\mathbb{I}_q)$$

$$\varepsilon_i \sim \mathcal{N}(0\ ,\ \tau_x^{-1}), \qquad i = 1, \dots, n$$

- Hyperprior distributions :

$$\tau_\theta \sim \text{Gamma}\left(1, 5e^{-5}\right)$$

$$-\tilde{\rho}_0 \sim \text{log-}\mathcal{N}(2.6, 0.067), \quad \tau_u \sim \text{Gamma}(96.12, 12.27), \quad \phi \sim UC(0, 1)$$

$$\tau_x \sim \text{Gamma}(96.2, 58.64)$$

## 6. Measurement Error correction

The table below shows the posterior means (and standard deviations) associated with the ME prone variable in three specifications of the model

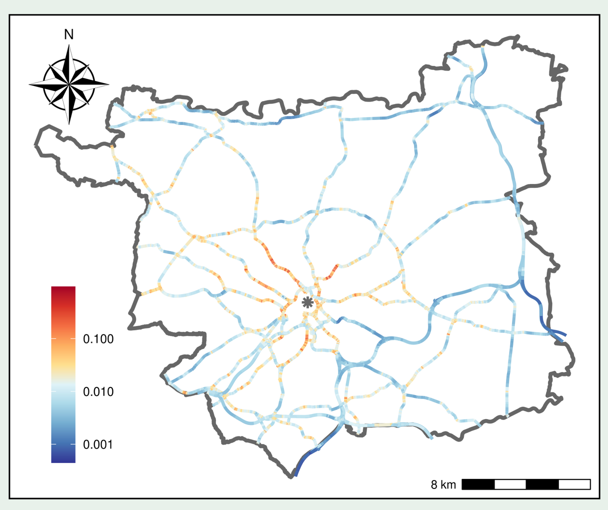
| | Baseline | ME corrected | Spatial ME corrected |
|---|---|---|---|
| Traffic volumes (X) | 0.216 (0.040) | 0.402 (0.070) | 0.506 (0.058) |

Model specification (Computational time):

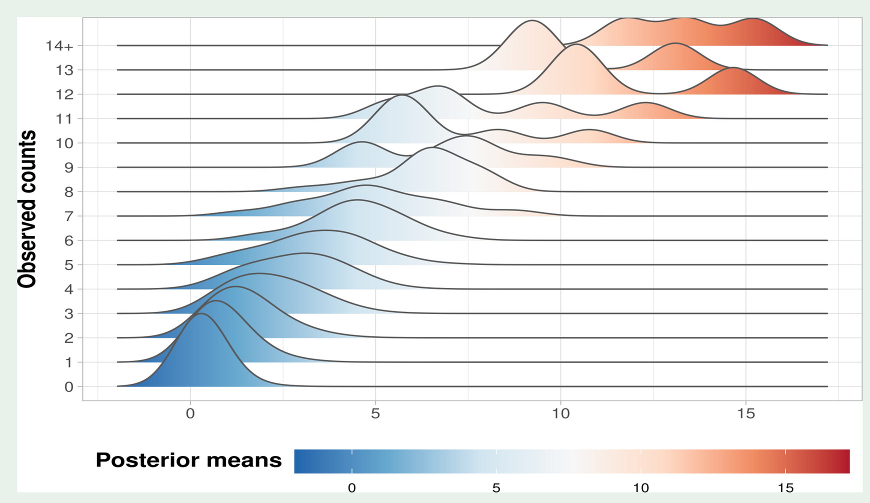
- Baseline : completely ignores the presence of ME (45 sec).
- ME corrected : classical ME model (3 min).
- Spatial ME corrected : ME correction with a spatially structured random effect (6 min).

## 7. Results and Validation

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



(c) Posterior means of $\lambda_i$ for all street segments in the city network.




(d) Density curves comparing posterior means of predicted car crash counts and observed counts.

## 8. Conclusion

- A multiplicative structure for the ME model allows the introduction of the parameter $\tilde{\rho}_0$ which estimates the proportion of observed traffic volume (in **log** scale).
- We believe these results are particularly important in real situations since they demonstrate naive models may provide misleading guidance for policy evaluation.
- We tested the temporal stability of our results under three scenarios, noticing that, in general, the posteriors are stable.

## 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.