Lattice models for spatial data on linear networks

Presenter: Andrea Gilardi
Contact: andrea.gilardi@unimib.it
University of Milano - Bicocca

About me

- I'm a research fellow at the University of Milano - Bicocca, Department of Economics, Management and Statistics.
- I'm interested in spatial and spatio-temporal statistic for events on linear networks.
- I'm a passionate R user focusing on georeferenced data. I also maintain two R  : *{osmextract}* and *{sfnetworks}*!

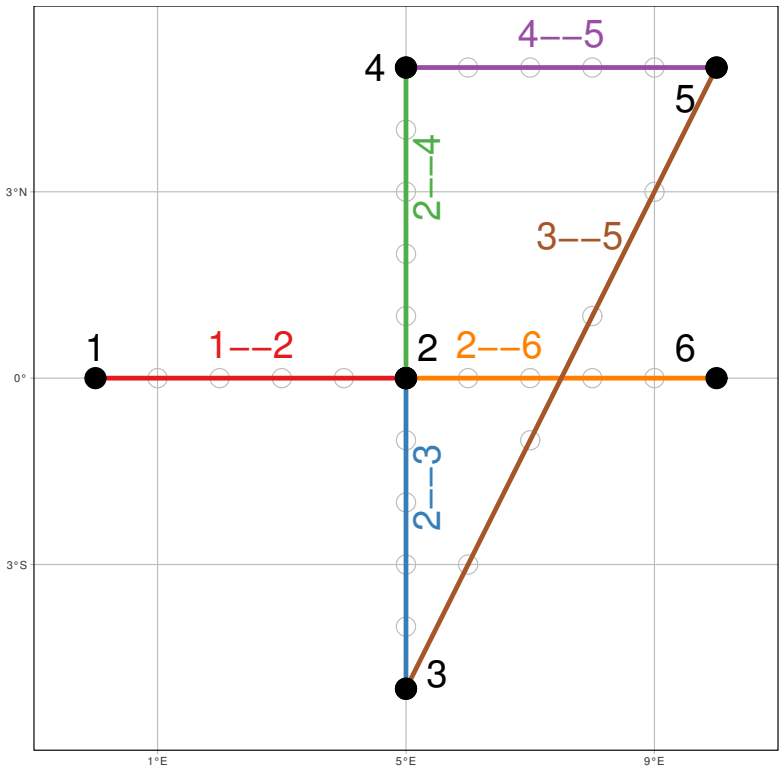
Outline of this presentation

- Introduction:
  - So what is a linear network? And a spatial network lattice?
  - The case study.
- Then I will present two projects named:
  - Multivariate hierarchical analysis of car crashes data considering a spatial network lattice;
  - Measurement error models for spatial network lattice data.
- Summary and conclusions

Introduction

- In the last years, we observed a surge of interest in the statistical analysis of spatial data lying on linear networks.
- Car crashes, vehicle thefts, and ambulance interventions are the most typical examples, whereas the edges of the network represent an abstraction of roads, rivers or railways.
- The most important characteristic of these events is that they are constrained to lie on a restricted spatial domain, which cannot be ignored for proper statistical modelling.

What is a linear network?



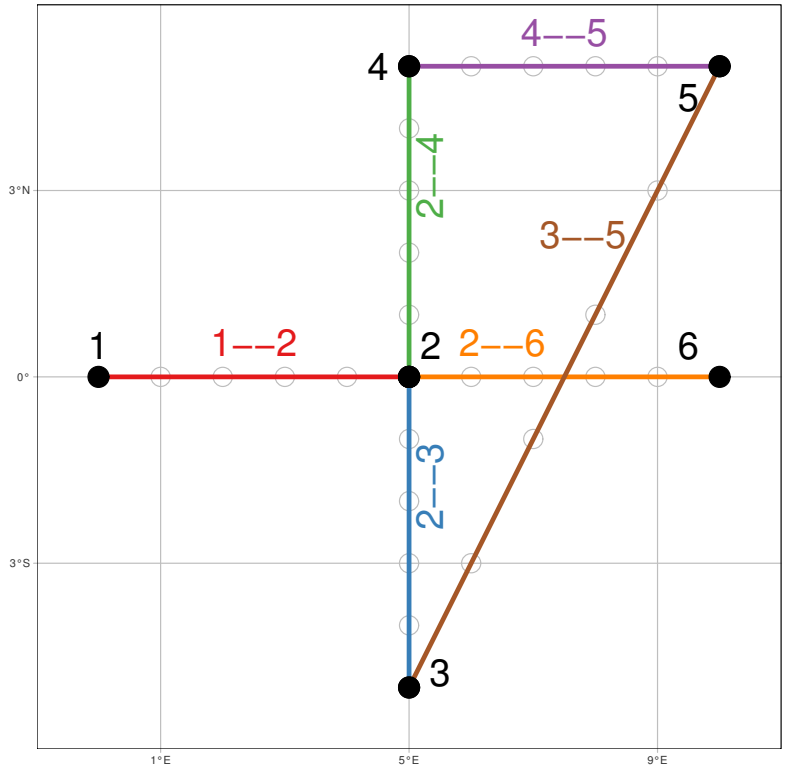
Spatial Component

```
Simple feature collection with 5 features
geometry type: LINESTRING
dimension: XY
1 LINESTRING (0 0, 1 0, ...)
2 LINESTRING (5 0, 5 -1, ...)
3 LINESTRING (5 0, 5 1, ...)
4 LINESTRING (5 5, 6 5, ...)
5 LINESTRING (5 0, 6 0, ...)
...
```

Graph Component

```
6 vertices
[1] 1 2 3 4 5 6
6 edges
[1] 1--2 2--3 2--4 4--5 2--6 3--5
```

And a spatial network lattice?



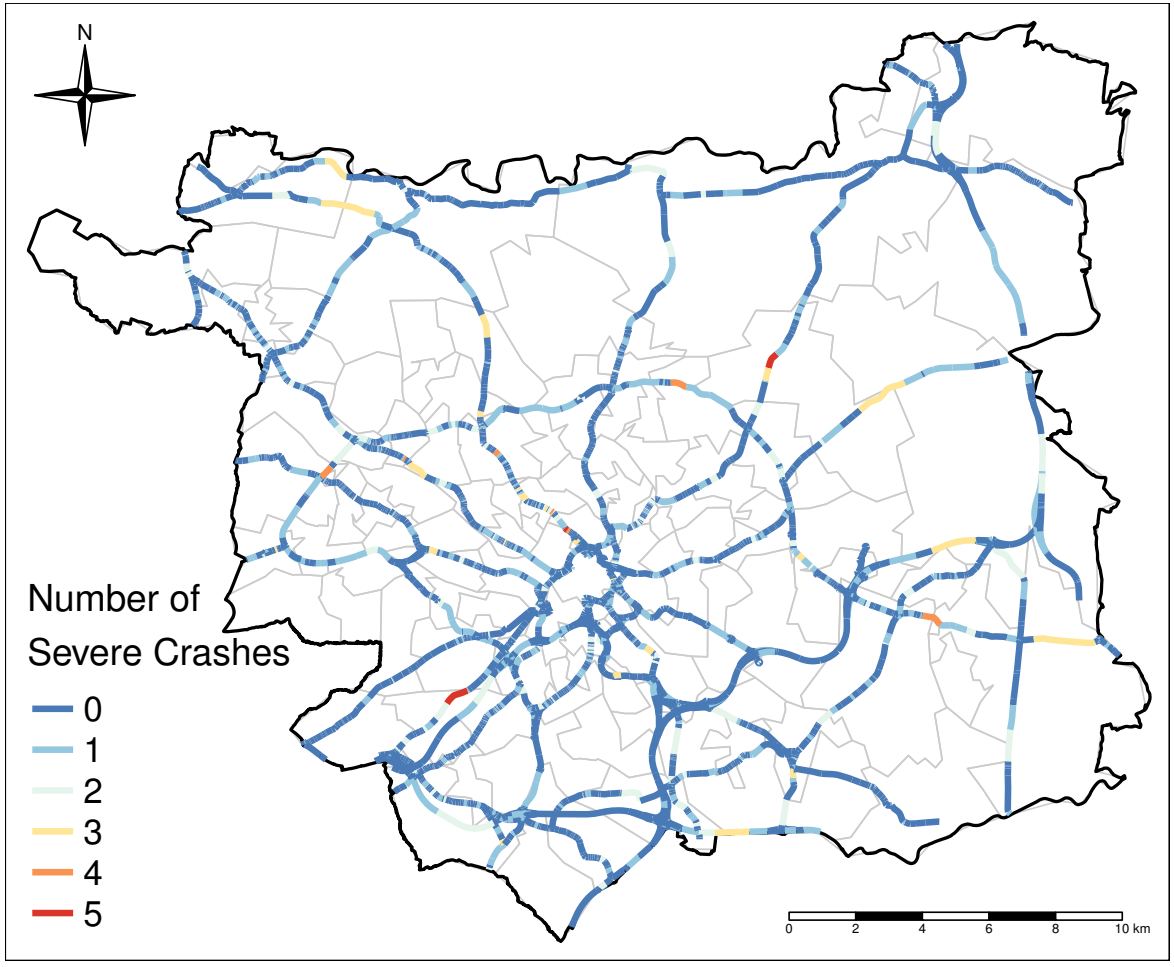
$$
\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{array}
\begin{array}{c}
\begin{array}{cccccc} 1 & 2 & 3 & 4 & 5 & 6 \end{array} \\
\begin{pmatrix}
\cdot & 1 & 1 & \cdot & 1 & \cdot \\
1 & \cdot & 1 & \cdot & 1 & 1 \\
1 & 1 & \cdot & 1 & 1 & \cdot \\
\cdot & \cdot & 1 & \cdot & \cdot & 1 \\
1 & 1 & 1 & \cdot & \cdot & \cdot \\
\cdot & 1 & \cdot & 1 & \cdot & \cdot
\end{pmatrix}
\end{array}
$$

Case study: car crashes in Leeds (UK)

- We analysed the car crashes that occurred in the road network of Leeds (UK) from 2011 to 2019. The sample included approximately 4000 events.
- The locations and the severity levels (coded as *slight* or *severe*) were obtained from an official database.
- The computational representation of the street network was built using data derived from Ordnance Survey (first project) and TomTom move (second project).
- They two networks are composed by approximately 4000 segments covering 800km.

Case study: car crashes in Leeds (UK)

Severe car crashes counts in the street segments of Leeds



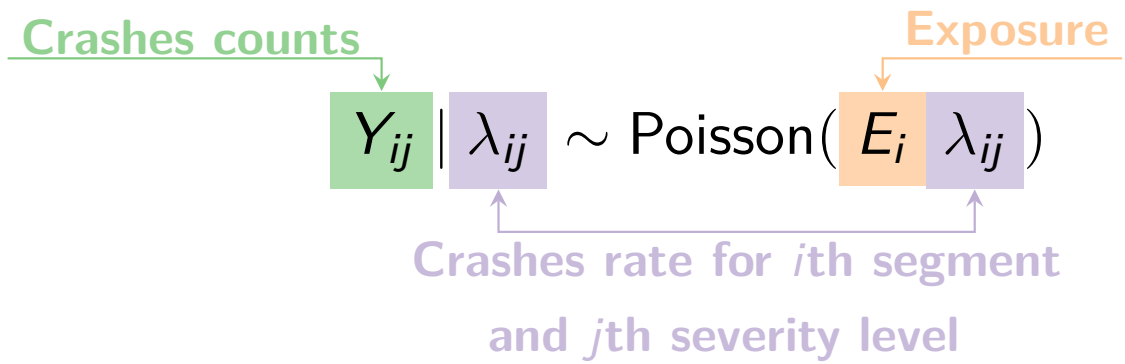
Multivariate hierarchical analysis of car crashes data considering a spatial network lattice

Joint work with Jorge Mateu, Riccardo Borgoni, and Robin Lovelace

The Bayesian Hierarchical model

- We modelled the data using a multivariate Bayesian hierarchical model with spatially structured and unstructured random effects.

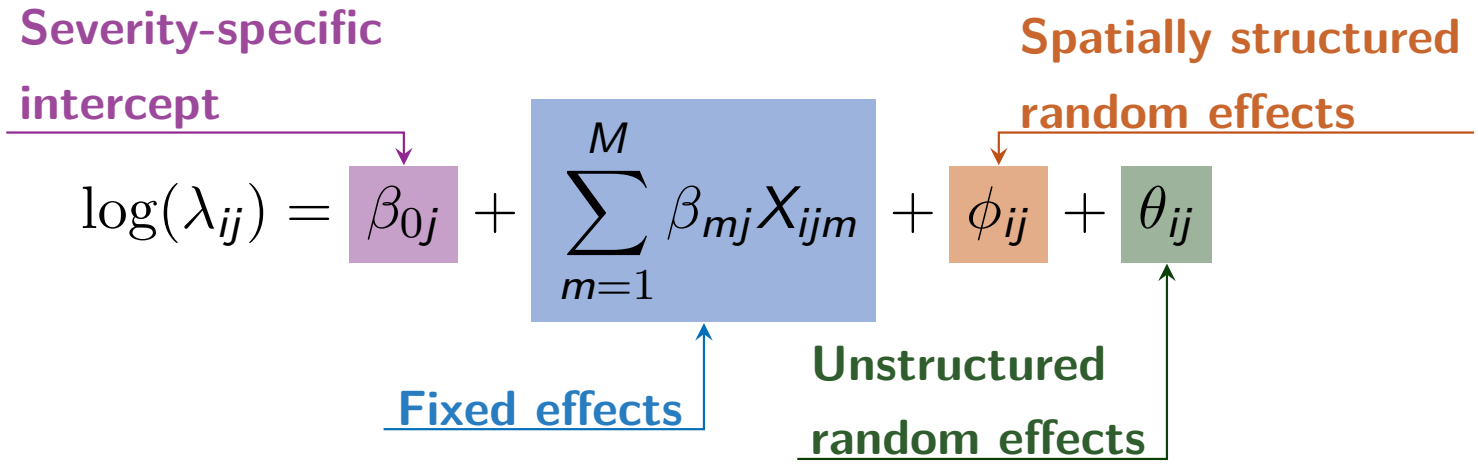
- In the first level of the hierarchy, we assumed that



- We have $i = 1, \ldots, 3661$ units and $j \in \{1, 2\}$ severity levels.

The Bayesian Hierarchical model (cont)

- At the second stage of the hierarchy, we assumed a log-linear structure on λ_{ij}:



- We tested six increasingly complex specifications for ϕ_{ij} and θ_{ij}, but first let me set up the terminology!

IMCAR prior

- Given a matrix $\mathbf{\Phi} = \{\phi_{ij}\}$, the Intrinsic Multivariate Conditional AutoRegressive (**IMCAR**) distribution is defind using a set of multivariate conditional distributions:

$$\mathbf{\Phi}_{i\cdot} \,|\mathrm{vec}(\, \mathbf{\Phi}_{-i\cdot}\,);\, \Omega \;\sim N_J\left(\frac{\sum_{i' \in \partial_i} \mathbf{\Phi}'_{i'\cdot}}{m_i};\, \frac{\Omega^{-1}}{m_i}\right).$$

$\mathbf{\Phi}_{i\cdot}$: The ith row of $\mathbf{\Phi}$

$\mathbf{\Phi}_{-i\cdot}$: The matrix $\mathbf{\Phi}$ without the ith row

Ω: $J \times J$ precision matrix

∂_i: Neighbours of unit i

m_i: N° of neighbours of unit i

- The vec operator row-binds the columns of a matrix.

IIMCAR prior

- The Independent Intrinsic Multivariate Conditional AutoRegressive (**IIMCAR**) prior represents a particular case of **IMCAR** obtained by setting

$$\Omega^{-1} = \begin{bmatrix} \sigma^2_{\phi_1} & 0 \\ 0 & \sigma^2_{\phi_2} \end{bmatrix}$$

- IIMCAR assumes independence between the J levels.

- Both distributions suffer from rank-deficiency problems that are solved by imposing sum-to-zero constraints.

PMCAR prior

- Given a matrix $\mathbf{\Phi} = \{\phi_{ij}\}$, the Proper Multivariate Conditional AutoRegressive (**PMCAR**) distribution can be characterised by:

$$\mathbf{\Phi}_{i\cdot}|\text{vec}(\mathbf{\Phi}_{-i\cdot}); \Omega \sim N_J\left(\frac{\boldsymbol{\rho}\sum_{i'\in\partial_i}\mathbf{\Phi}'_{i'\cdot}}{m_i}; \frac{\Omega^{-1}}{m_i}\right).$$

- The parameter $\boldsymbol{\rho}$ controls the strength of the spatial dependence, while all the other parameters have the same interpretation as before.

- The joint distribution is proper if $|\boldsymbol{\rho}| < 1$.

IPMCAR prior

- Analogously to the previous case, the Independent Proper Multivariate Conditional AutoRegressive (**IPMCAR**) prior is defined as a particular case of (**PMCAR**) obtained by setting

$$\Omega^{-1} = \begin{bmatrix} \sigma^2_{\phi_1} & 0 \\ 0 & \sigma^2_{\phi_2} \end{bmatrix}$$

- It also assumes independence between the two severity levels.

Summary

Now we can characterise the six different specifications we tested:

	Spatial effects	Unstructured effects
(A)	IIMCAR	Bivariate Gaussian with $\rho_\theta = 0$
(B)	IPMCAR	Bivariate Gaussian with $\rho_\theta = 0$
(C)	IMCAR	Bivariate Gaussian with $\rho_\theta = 0$
(D)	PMCAR	Bivariate Gaussian with $\rho_\theta = 0$
(E)	IMCAR	Bivariate Gaussian
(F)	PMCAR	Bivariate Gaussian

Hyperprior distributions

- In all cases, we adopted a Wishart hyperprior with parameters 2 and $\boldsymbol{I}_2$ (the identity matrix of order 2) for the variance covariance matrices of the two multivariate random effects.

- For the spatial autocorrelation parameter, we set $\rho \sim \text{Uniform}(0, 1)$ to avoid counter-intuitive spatial repulsive behaviours.

Results

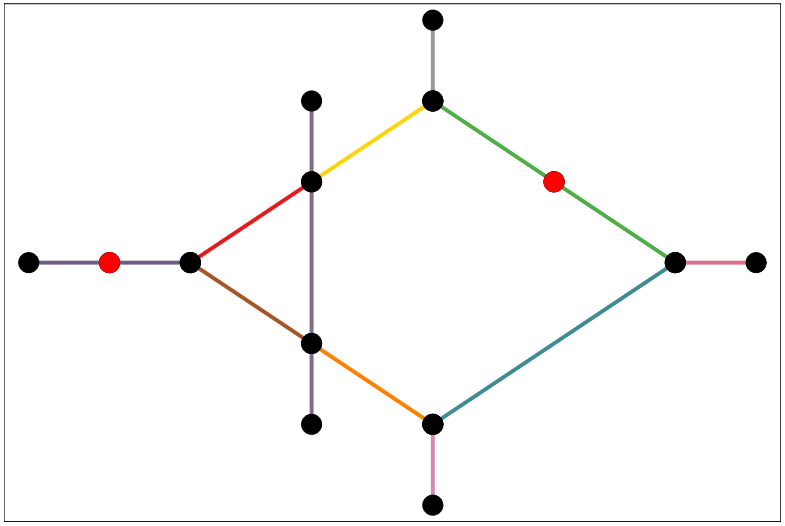
- Considering the scale of the problem, we adopted the Integrated Nested Laplace Approximation (INLA) approach instead of classical MCMC sampling.
- We included five covariates representing some social and physical characteristics of the road segments.
- We will now focus only on the MAUP analysis!

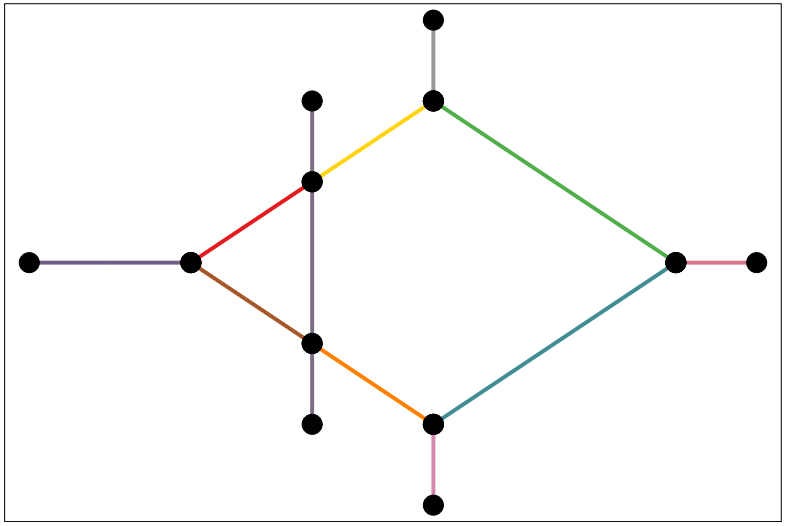
The Modifiable Areal Unit Problem

- Spatial aggregation leads to a well-known problem in geographical analysis, the Modifiable Areal Unit Problem (**MAUP**).
- MAUP implies that the size of the spatial units impacts on the statistical analysis.
- Hence, conclusions drawn at one scale of spatial aggregation might not necessarily hold at another scale.

MAUP on a linear network (cont)

A toy example is sketched in the following figures.





MAUP on a linear network (cont)

- The contracted network included approximately 2700 segments (instead of 3661).
- The estimates of fixed effects were not influenced by the new configuration and did not change in magnitude, sign, or significance.
- On the other hand, the new approach presents a slightly greater spatial uncertainty than model (F), in particular for $\sigma^2_{\phi_1}$ and $\sigma^2_{\phi_2}$, but similar posterior distributions for predicted values.

MAUP on a linear network: Conclusions

- A few authors explored the importance of MAUP for road safety modelling at the areal level.
- They conclude saying that MAUP affects the magnitude and significance of the estimates for both fixed and random effects.
- Our results tell a different story and suggest that a network lattice approach should not be ignored when analysing phenomena that naturally occur on a network.

Measurement error models for spatial network lattice data

Joint work with Riccardo Borgoni, Luca Presicce, and Jorge Mateu

Introduction

- There exists a vast literature that links traffic accidents to a variety of factors such as vehicle characteristics, social and environmental conditions, or traffic volumes.
- Nevertheless, it is quite difficult to obtain precise traffic figures at the road segment level.
- Therefore, several authors adopted alternative ways (e.g. OD data) to approximate the traffic flows.

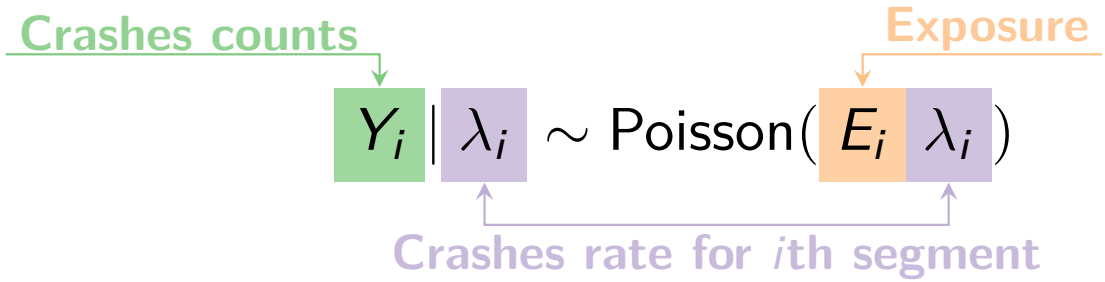
Traffic estimates from mobile data

For example:

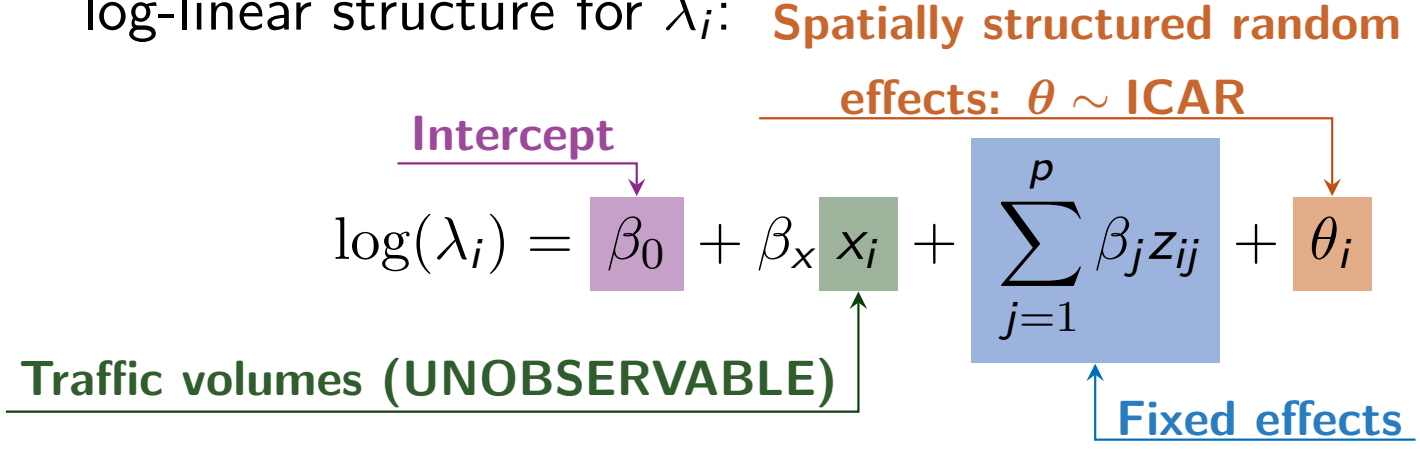
Source: https://www.tomtom.com/products/road-traffic-data-analytics/

The (ideal) modelling strategy

- In the first level of the hierarchy, we assume that



- Similarly to the previous example, the second level defines a log-linear structure for λ_i:



A first approximation

- Clearly, the model introduced in the previous frame cannot be estimated since x_i is not observable.

- As a first approximation, we might substitute the traffic measure derived from TomTom data directly into the model:

$$Y_i|\lambda_i \sim \text{Poisson}(e_i\lambda_i)$$

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

Traffic volumes obtained from TomTom data

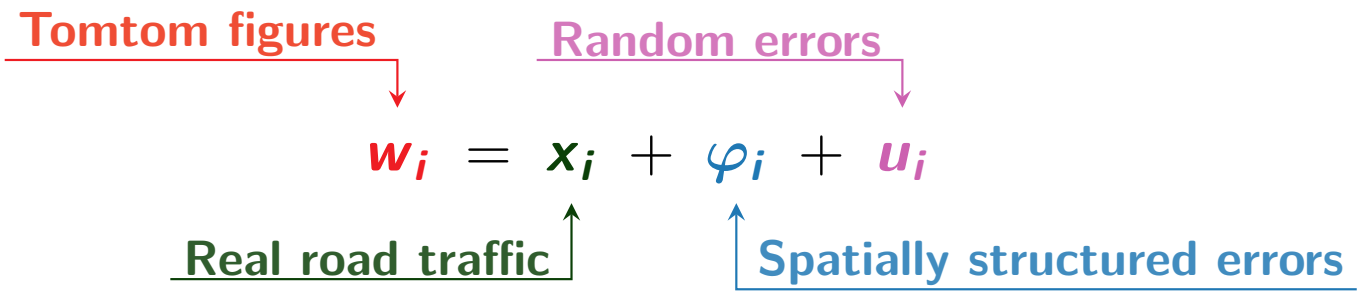
Is that a good idea?

Spatial Classical ME Model

- The previous model ignores that traffic figures derived from TomTom devices may suffer from (Spatial) Measurement Error (**ME**).

- It is well known that ignoring ME may lead to biased point estimates (i.e. attenuation bias) and problematic statistical inference.

- We propose to adopt a Spatial Classical ME correction:



Spatial Classical ME Model (cont)

- We adopt the following Bayesian hierarchical model.
- The first stage is specified as before

$$Y_i|\lambda_i \sim \text{Poisson}(e_i\lambda_i)$$

while in the second stage we assume:

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

Results

- We adopted the INLA via a re-parametrization with augmented 0 pseudo-observations.

- The posterior means/sd of β_x are as follows:

Baseline	Classical ME	Spatial ME
0.319	3.990	7.564
(0.041)	(0.081)	(0.054)

- Keeping all the other quantities as fixed, an increment of 100,000 annual traffic units correspond to an increase of car crashes rate equal to 1.046, 1.768 and 3.116 units.

Summary and conclusions

- We presented a multivariate Bayesian hierarchical model for the analysis of car crashes data on a spatial network lattice.
- We found that the data has complex spatial patterns that require the presence of interactions between the two severity levels. Moreover, the results are more robust than equivalent planar models.
- We also discussed a procedure to include spatial covariates recorded with measurement error and exemplified the problem with traffic flow data.

Next steps

- In the future we plan to enhance the first model considering also the temporal dimension of the data (although that induces a lot of sparsity).

- The second project can be enhanced in several ways:
  - Analyse the flows in the log scale?
  - Non-linear specification for the traffic covariate? U-shape?
  - Multivariate ME models?

The end! Grazie per l'attenzione :)

References I

Adrian Baddeley, Gopalan Nair, Suman Rakshit, Greg McSwiggan, and Tilman M Davies. "Analysing point patterns on networks—A review". In: *Spatial Statistics* 42 (2021), p. 100435.

Sudipto Banerjee, Bradley P Carlin, and Alan E Gelfand. *Hierarchical modeling and analysis for spatial data*. Chapman and Hall/CRC, 2003.

Marc Barthélemy. "Spatial networks". In: *Physics reports* 499.1-3 (2011), pp. 1–101.

L Bernadinelli, Cristian Pascutto, NG Best, and WR Gilks. "Disease mapping with errors in covariates". In: *Statistics in medicine* 16.7 (1997), pp. 741–752.

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.

Raymond J Carroll, David Ruppert, Leonard A Stefanski, and Ciprian M Crainiceanu. *Measurement error in nonlinear models: a modern perspective*. Chapman and Hall/CRC, 2006.

KV Mardia. "Multi-dimensional multivariate Gaussian Markov random fields with application to image processing". In: *Journal of Multivariate Analysis* 24.2 (1988), pp. 265–284.

References II

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

Stan Openshaw. "The modifiable areal unit problem". In: *Quantitative geography: A British view* (1981), pp. 60–69.

Edzer J Pebesma et al. "Simple features for R: standardized support for spatial vector data.". In: *R J.* 10.1 (2018), p. 439.

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.

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.