Hierarchical spatial network models for road accident risk assessment

Diamante, 12 June 2025

Gian Paolo Clemente

Professore Associato e Attuario

Università Cattolica del Sacro Cuore, Milano

gianpaolo.clemente@unicatt.it

UNIVERSITÀ CATTOLICA del Sacro Cuore

Joint work con Francesco Della Corte e Diego Zappa (Università Cattolica del Sacro Cuore, Milano)

Aim of the work

- Our **aim** is to exploit roads characteristics, traffic, socio-demographic local data and the location of past accidents to estimate the risk of getting car crashes for any edge of a (local or even nationwide) road network.
- **Possible benefits**
  - For policymakers: more efficient use of public resources to reduce the risk of accidents (i.e. where is it necessary to invest?)
  - For civil engineers: evidence of what are the main factors that may impact the risk of an accident (i.e. are roundabouts riskier than traffic lights?)
  - For everyday use: which roads are safer?
  - For insurance companies: how to link the risk of drivers' trajectories to expected frequency (blackboxes recordings are necessary)

Contribute to the literature and methods

- **In particular we focus on "where policyholders drive"**
  - We **do not** consider here (research is in progress) other features that can be detected by telematic data and that can affect the risk as:
    - Driving behaviour (see, e.g., Ayuso et al. (2016), Gao and Wuthrich (2018), Arumugam, Bhargavi (2019), Gao and Wuthrich (2019), Gao et al. (2019), Huang, Meng (2019), Narvaez et al. (2019), Wuthrich and Buser (2019), Gao, Meng, Wuthrich (2022), Ziakopoulos et al. (2024), …)
    - Driving habits as KM, daytime, weather conditions, etc. (see, e.g., Ayuso et al. (2018), Verbelen et al. (2018), Perez-Marin and Guillen (2019), Guillen et al. (2021))

- **We follow a combined approach:**
  - **A modification of the conditional autoregressive modelling** (see Boulieri et al. (2016), Gilardi et al. (2022)) **incorporating spatial lagged effects,** will be applied in order to assess the risk on the basis of a set of features related to the characteristics of the streets.
  - **From the spatial object we build a weighted network**, where vertices and arcs correspond to geographical elements as junctions and roads and where the assessed risk of each segment is used as a weight.
  - **A two-stage mixed geographically weighted Poisson regression** (see Murakami et al. (2023), Briz-Redón et al. (2019) Gomes et al. (2017)) to unveil local heterogeneity.

Which "ingredients" do we need ?

Road details (e.g. Open Street Map)

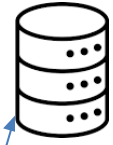
Traffic source (e.g. Google, other providers)

Demographic database (population density, building density, commuting people)

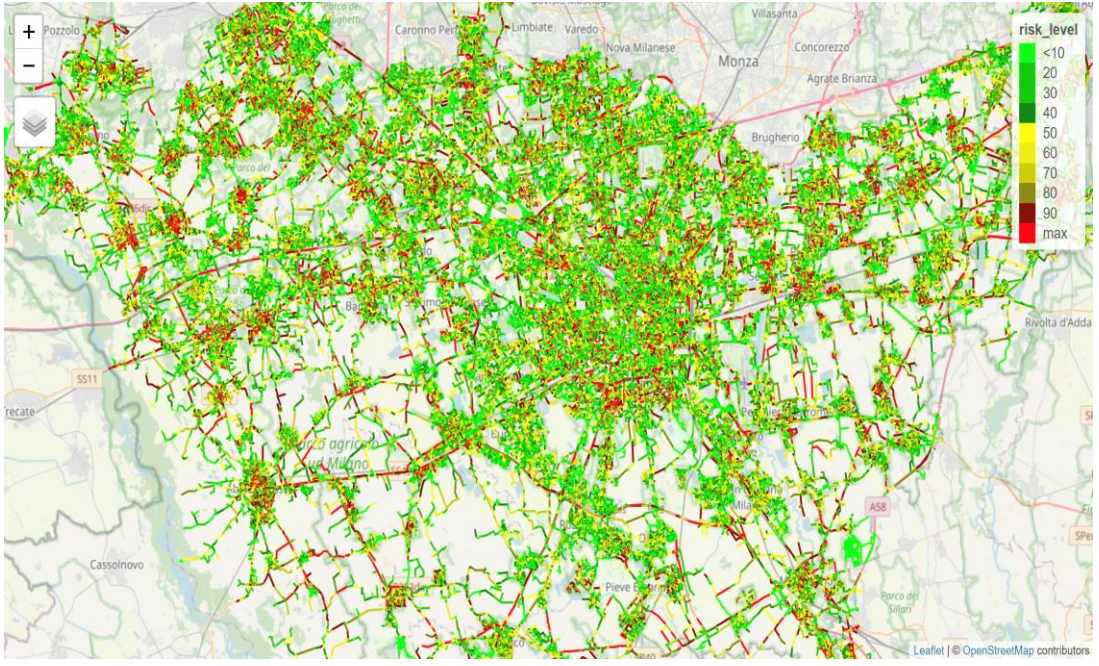
Region/municipality/ZIP boundaries (from ISTAT, other private sources)

Location of accidents (e.g. from company, open data)

Weather conditions



$$E(\#Accidents) = g^{-1}(features, \theta)$$



Main issues related to data

***Road* details** (e.g. Open Street Map)

- links details are often unbalanced because of missing information
- some very relevant details (i.e. number of crossings) are not available and must be ad hoc estimated

Traffic source (e.g. Google, other providers)

- high quality open access data are barely available
- the size of datasets are in many terabytes even for short time periods

Demographic databases (population, building density, commuting people)

- they are not available at the link level but mainly at a small area level

Region/municipality/ZIP boundaries (from ISTAT, other sources)

- what is the optimal subregion to fit data?

Location of accidents (e.g. from company, open data)

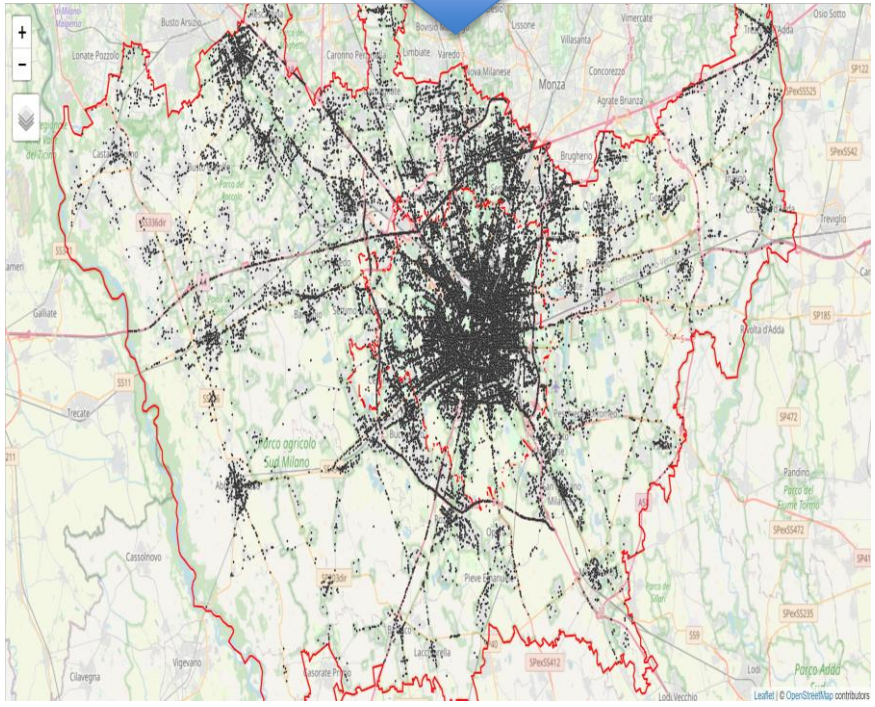
- In general, dataset contains location of accidents.
- Reverse geocoding (i.e. lat/long coordinates) algorithms are in some cases necessary but often with limited precisions

The number of road crossings is not directly available in the OSM database.
For each road, we computed it as the number of segments that have in common one coordinate with that road.
This method represents an approximation of the true crossings (for instance, two roads at different level one above the other through a bridge) but it returns in general an estimate quite close to reality.

Coordinates of accidents are not always strictly in line with a segment. Approximations are due to proxies implicit into the reverse geocoding algorithm or to errors in the registration of accident locations. We project (orthogonally) that coordinates onto the closest segment

The data
City of Milan and province

Data regard accidents that resulted in fatalities or injuries of at least one person. We display below the location for Milan and province



id_link	highway (type)	highway (length)	URBAN	# junctions	# pedestrian crossings	# traffic lights	# car crashes
1	Tertiary	119	Y	5	0	0	**0**
2	Secondary	309	Y	5	2	0	**0**
3	Primary	11.3	N	4	0	3	**0**
4	Primary	11.3	Y	5	1	0	**2**
5	Primary	150	Y	7	2	1	**0**
6	Secondary	35.4	N	6	0	0	**1**
7	Secondary	67.9	Y	6	0	3	**0**
8	Tertiary	97.7	Y	6	1	0	**3**
9	Motorway	157	N	4	0	0	**0**
10	Other	150	N	6	0	1	**1**

For each OSM segment save/compute

- Type of road (highway)
- Features (if available) e.g. surface, maxspeed, lit…
- Number of junctions (computed exogenously: proxy very close to reality)
- Number of traffic lights
- Number of pedestrian crossings

The model currently applied

We split the domain into subregions. We report the results obtained considering sub-areas based on cities and also ZIP codes for Milan

We propose a method that combines *Conditionally autoregressive models* (CAR) with two types of *Spatial Lag Models* (SLM and SLX)

- Let $\boldsymbol{Y} = (Y_1, Y_2, \dots, Y_n)^T$ be the vector of random variables (claim counts) observed in n different regions (in our case, links).

- In general the average count of a spatially dependent variable can be modeled as:

$$Y_i | \{y_j, j \in N(i)\} \sim Poisson(E_i \mu_i)$$

where $N(i)$ is the neighbourhood of node i, E_i is the exposure.

The model currently applied

Considering n records (i.e. links) and k covariates,we set:

$$\log(\boldsymbol{\mu}) = \left(\rho \boldsymbol{W}\boldsymbol{y} + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{W}\boldsymbol{X}_{(-1)}\boldsymbol{\eta}\right)$$

With:

- $\boldsymbol{W}$ proximity matrix (spatial weight matrix)
- $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$ vectors of coefficients (fixed effect and spatial random effects)
- $\boldsymbol{X}_{(-1)}$ indicates the $\mathbf{X}$ matrix without the column associated to the intercept.
- ρ spatial autocorrelation parameter

The model currently applied

The expected count or intensity λ is then obtained as:

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

$\boldsymbol{W} = [w_{i,j}]$ spatial weight matrix (n x n)

Elements $w_{i,j}$ are based on a *bi-square kernel function* (h=1km) (see Bidanset, Lombard (2014)):

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

Weights are then normalized to assure that the sum for row is equal to 1.

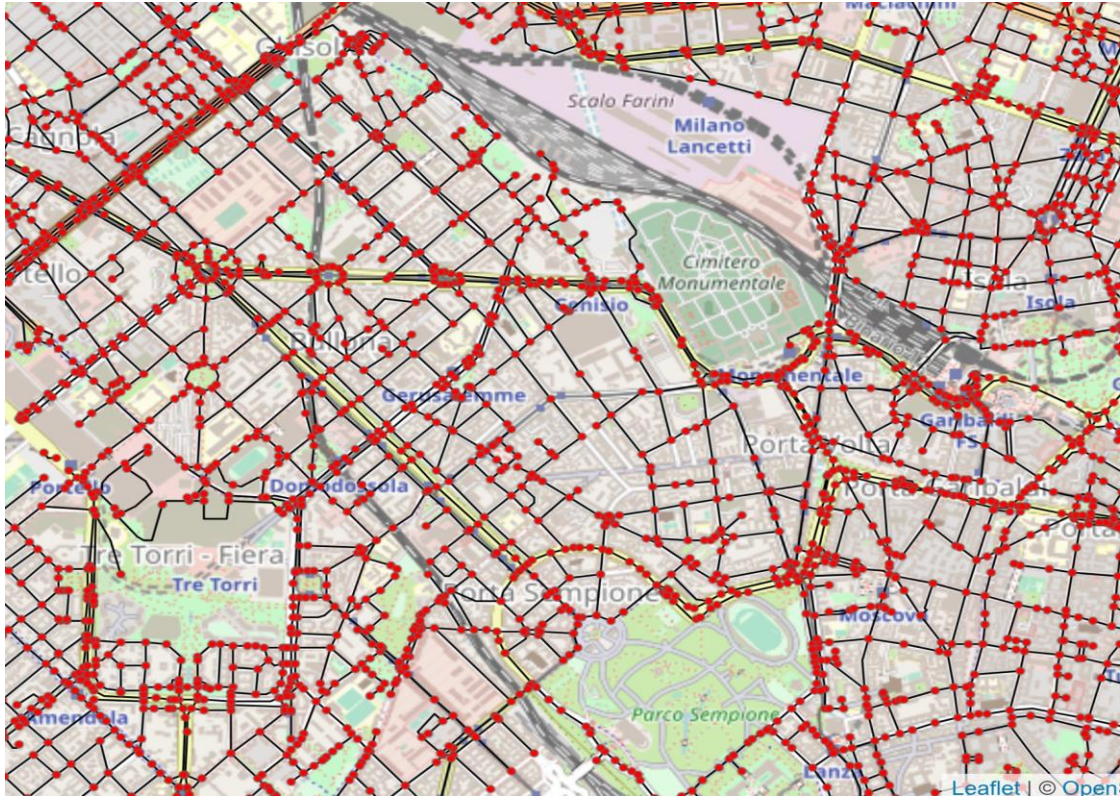
Offset: VehicleMilesTravelled (VMT) = #vehicles * length (total km travelled for each segment) or *length,* if traffic is not available

Distances

- To compute distances, we convert the street network in a *graph* **focusing on a "junction graph"** (see, e.g., Marshall et al., 2018), where each segment is an arc and nodes are given by junctions (or by termination of closed streets).

- Formally, given the street network, we build a graph $G = (V; E)$ where V and E are respectively the set of n vertices and m arcs. Two nodes are adjacent if there is an arc $(i, j)\ \epsilon\ E$ (i.e. a road segment) connecting them

- In particular, we consider at moment ***a directed and weighted network*** G_w equal to G, where each arc *is weighted with the length of the segment*.



- Distances between two roads have been computed by adding centroid to each segment and by computing the directed weighted shortest path between two centroids.

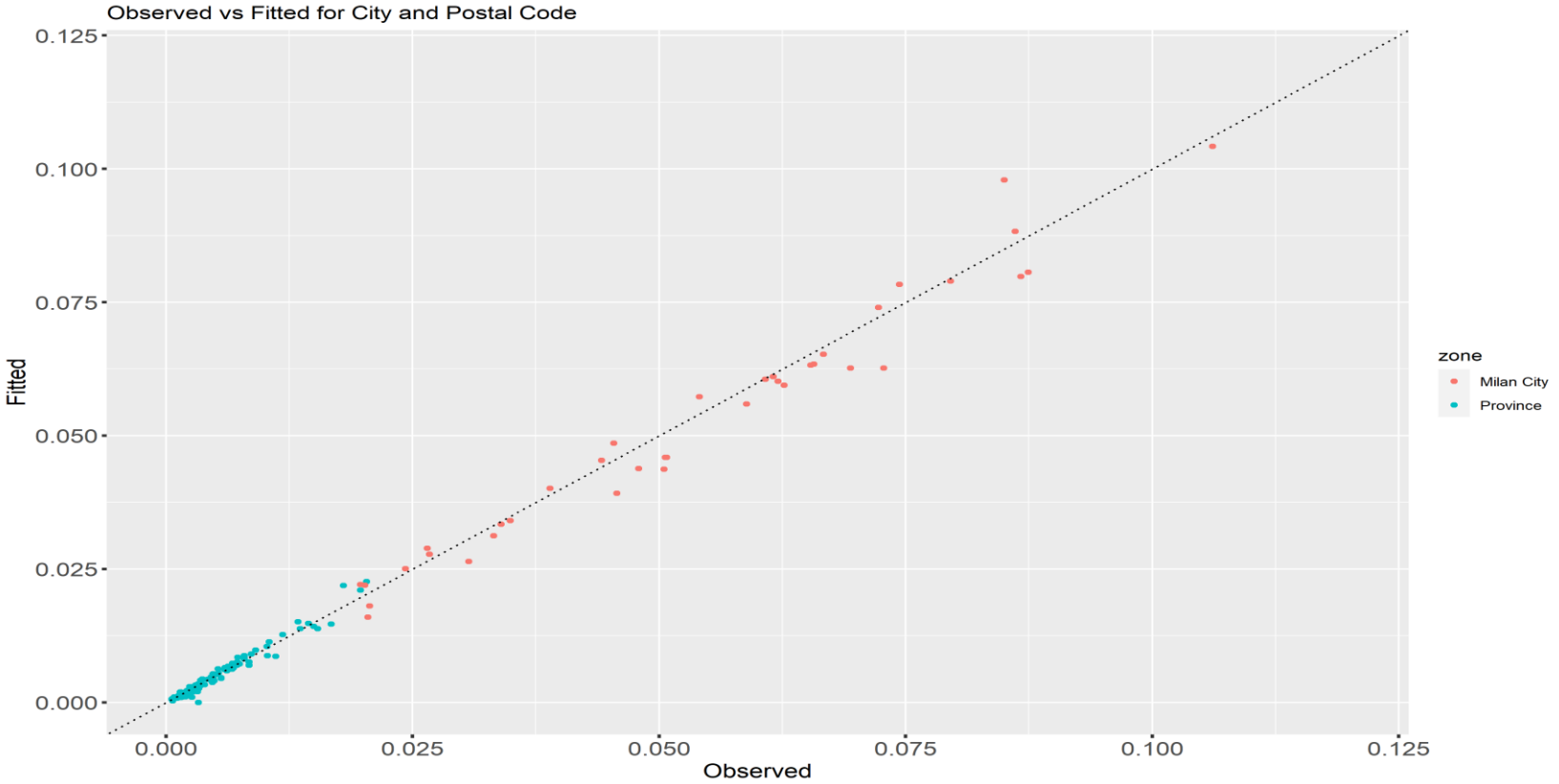
- The **shortest path problem** is the problem of finding a path between two nodes in a graph such that the sum of the weights of its constituent links is minimized:

$$d_{ij} = min\left\{ \sum_{(u,v \in P)} l(u,v) | P \in P(i,j) \right\}$$

Where:

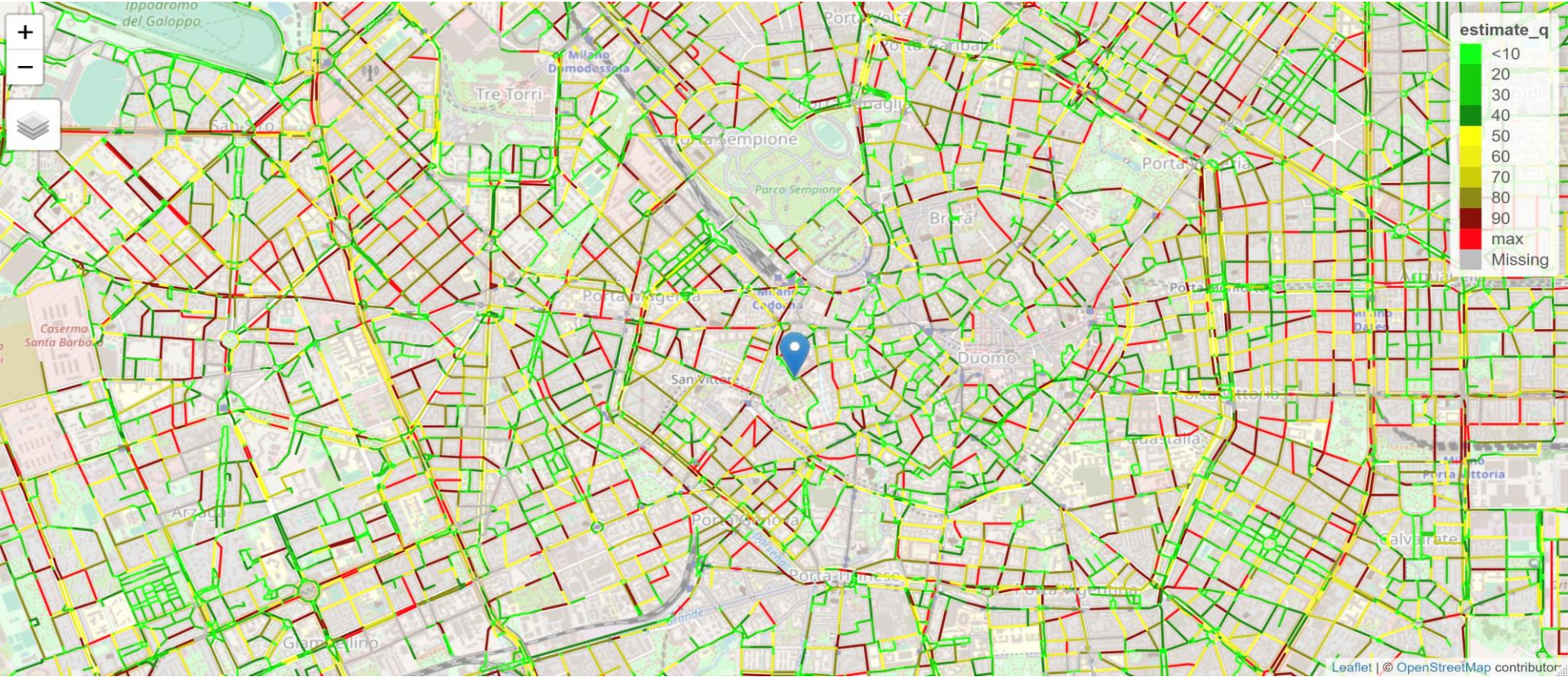
- $l(u,v)$ is the length of the link connecting two neighbours u, v
- P is a directed path from node i to node j, i.e. a sequence of directed links that allow to go from i to j
- $P(i,j)$ is the set including all directed paths from i to j

Main results:
Model behaviour



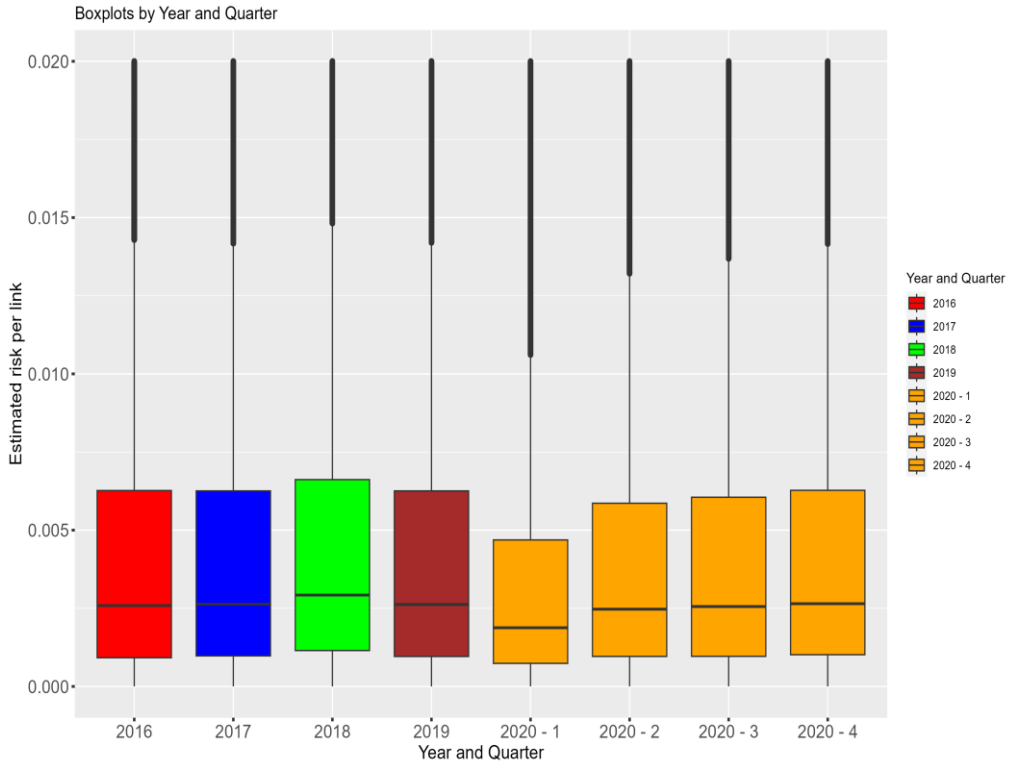
First results

Example of map of the risk Center of Milan

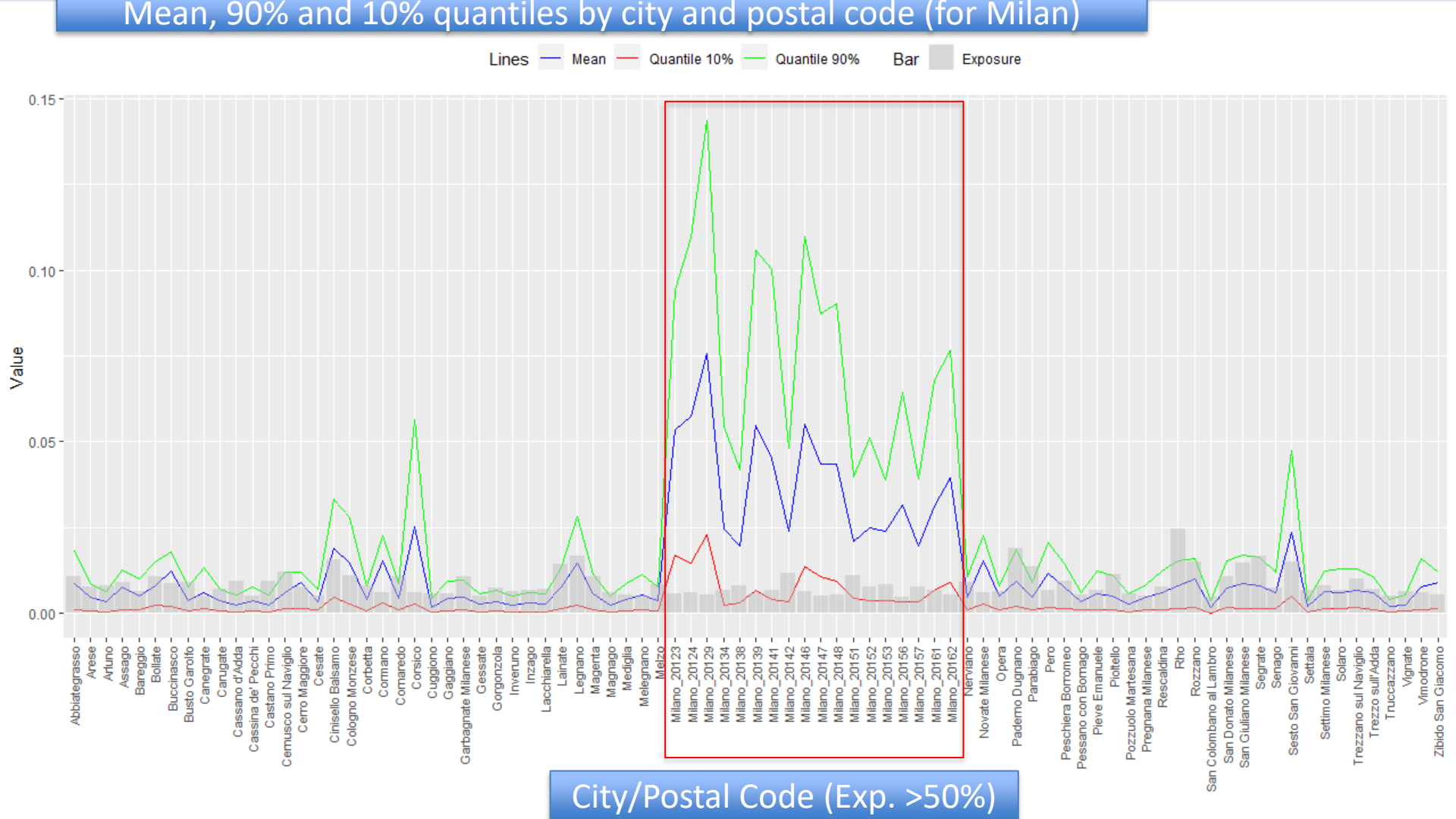


Risk based on time

Risk wrt to time

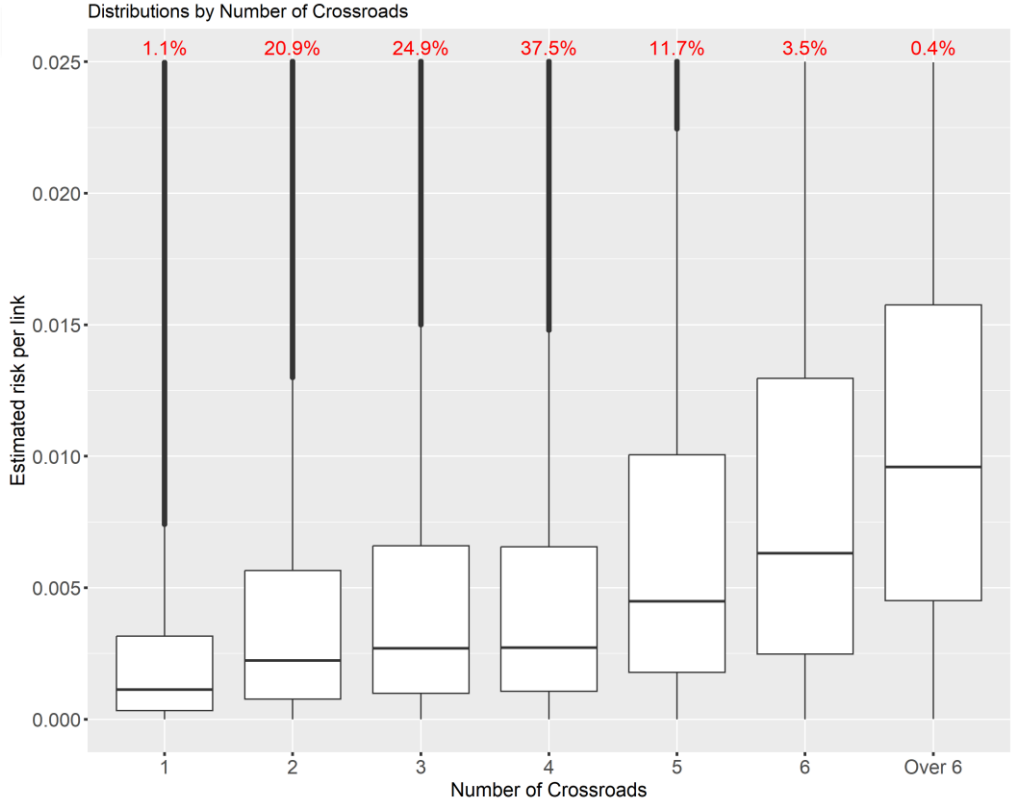


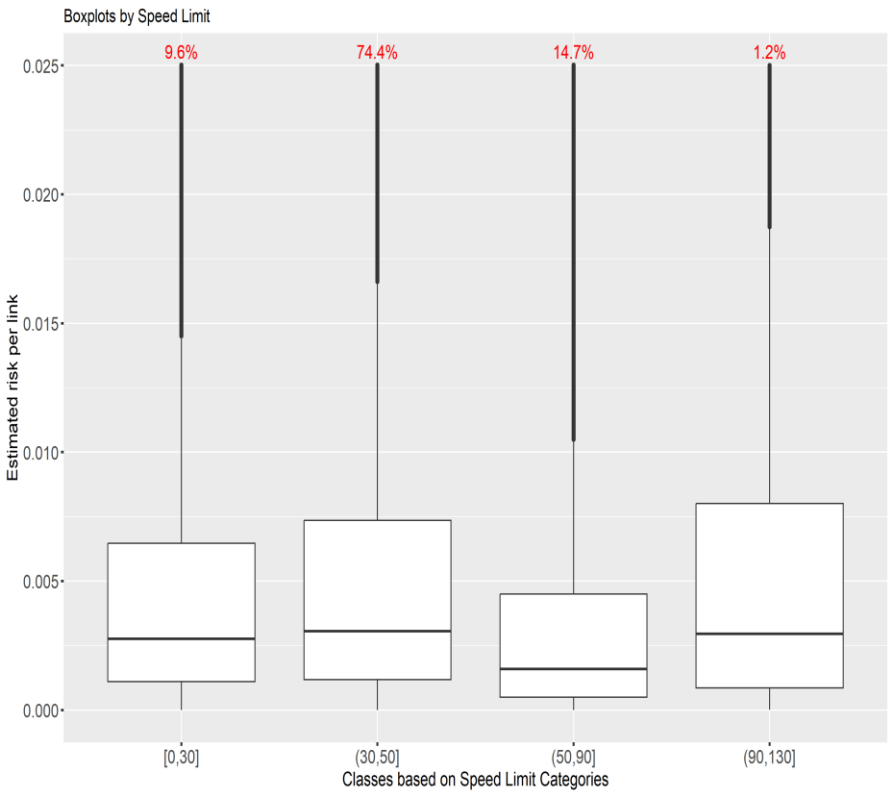
Mean, 90% and 10% quantiles by city and postal code (for Milan)



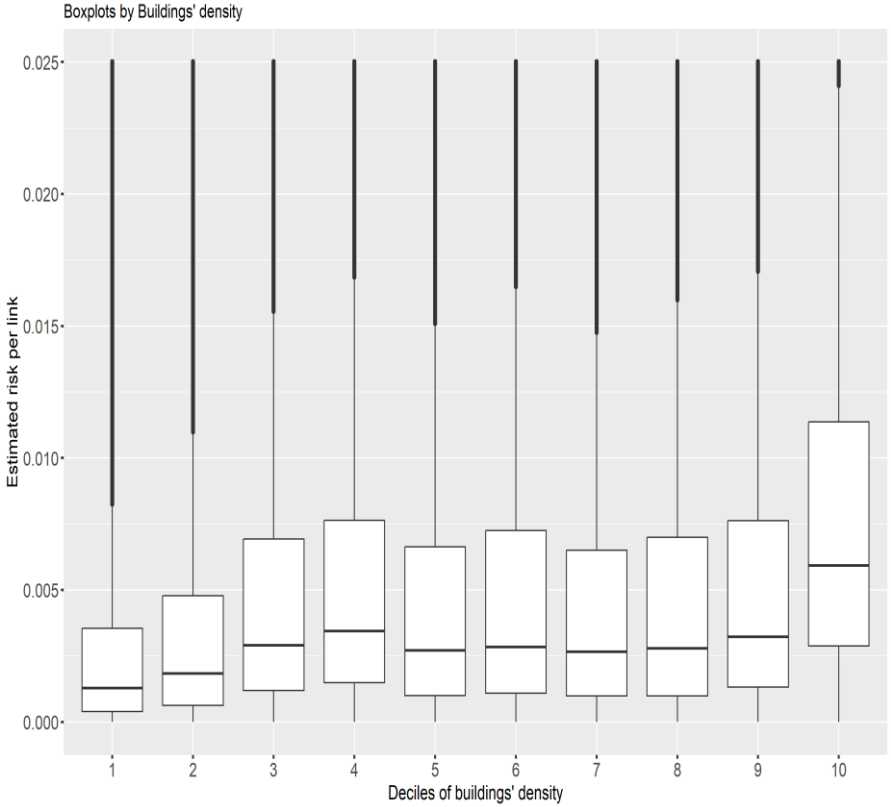
City/Postal Code (Exp. >50%)

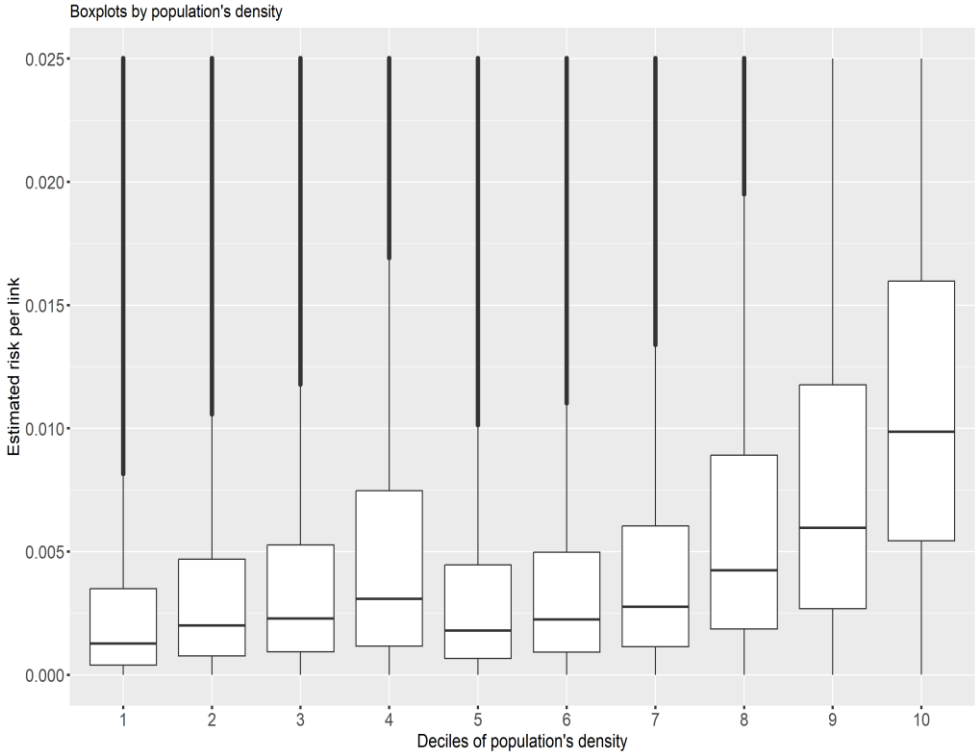
Complexity of the street



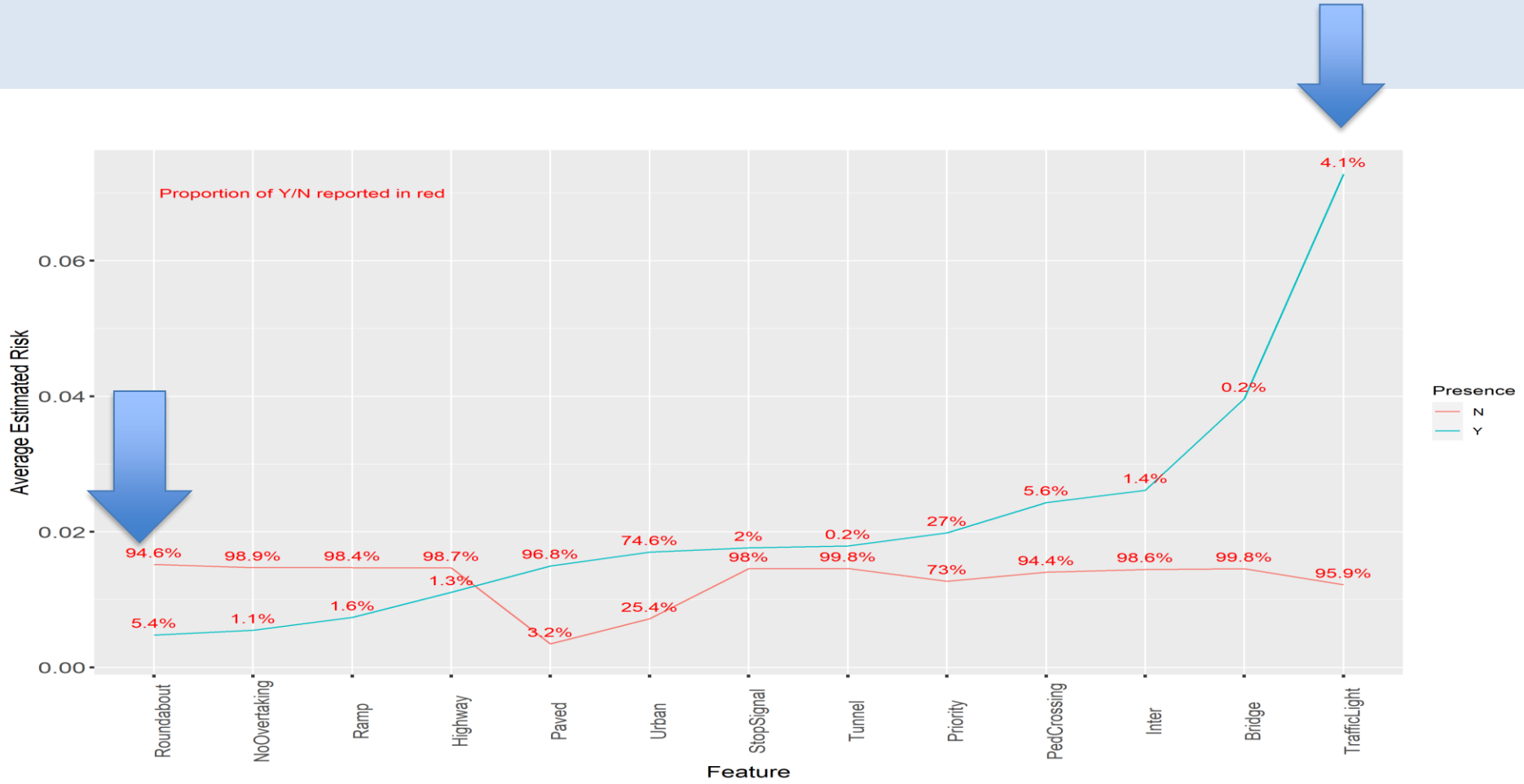


Density
Population and Buildings





Other characteristics



GWPR

We consider the Geographically Weighted Poisson regression (GWPR):

$$Y_j \sim Poisson\left[E_j \exp\left(\sum_k \beta_k(u_i, v_i)\, x_{ik}\right)\right]$$

GWPR estimates local coefficients by maximizing a locally weighted Poisson likelihood, where the weights come from the bi-square function with bandwith $h = h_j$

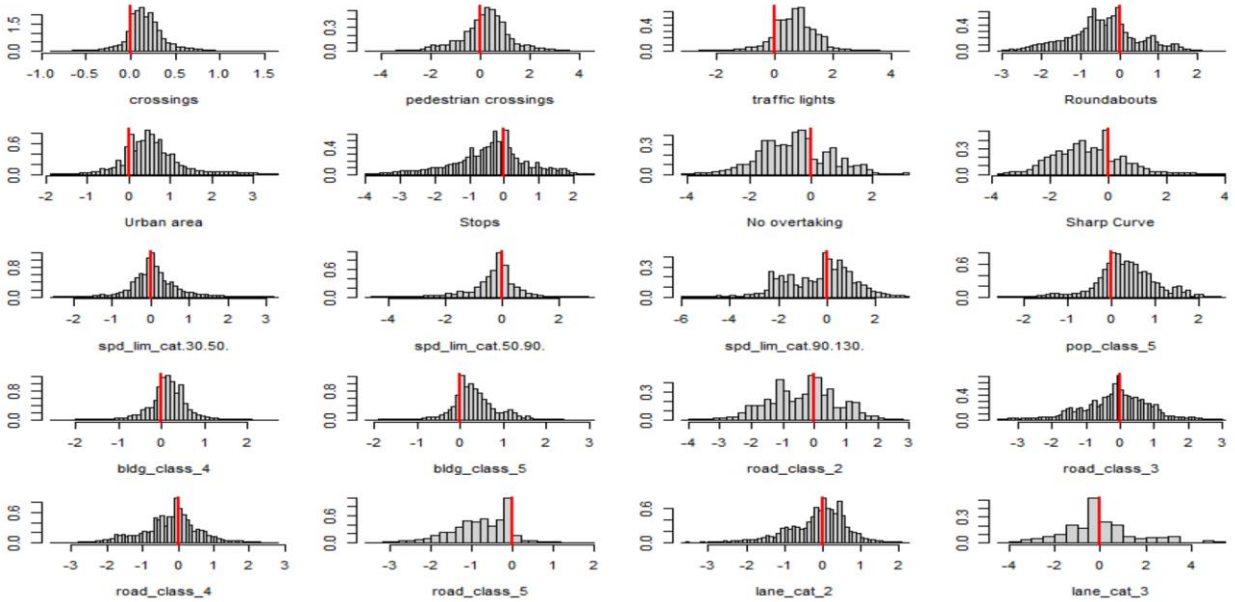
We apply a two stage GWP regression. A sketch is:

1. For each position lat-long, say $\boldsymbol{u}$, select h_j and fit a penalized GWPR model.
2. Divide the explanatory variables into $\boldsymbol{X}_{(-z)}$ and $\boldsymbol{Z}$, where $\boldsymbol{Z}$ contains the not significant variables (*irrelevant to the problem or local*).
3. Apply a penalized (not geographical) elastic net to $\boldsymbol{Z}$. Let $\boldsymbol{Q}$ be the variables selected.
4. In case $\boldsymbol{Q}$ is not empty, fit a penalized mixed-GWPR with

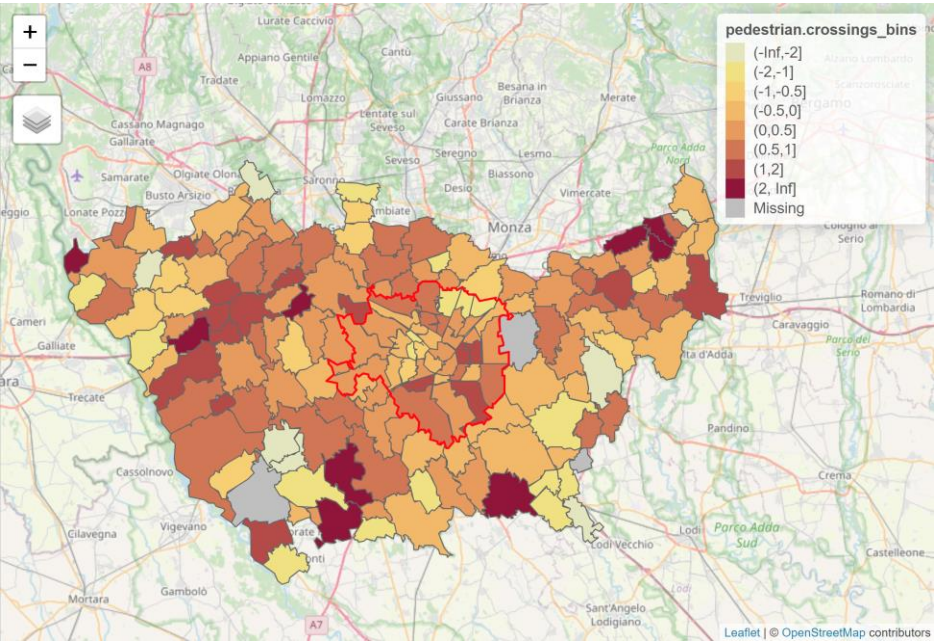
$$\log\left(\boldsymbol{\mu}(u)\right) = \boldsymbol{X}_{(-z)}\boldsymbol{\beta}_{(-z)}(u) + \boldsymbol{Q}\boldsymbol{\delta}$$

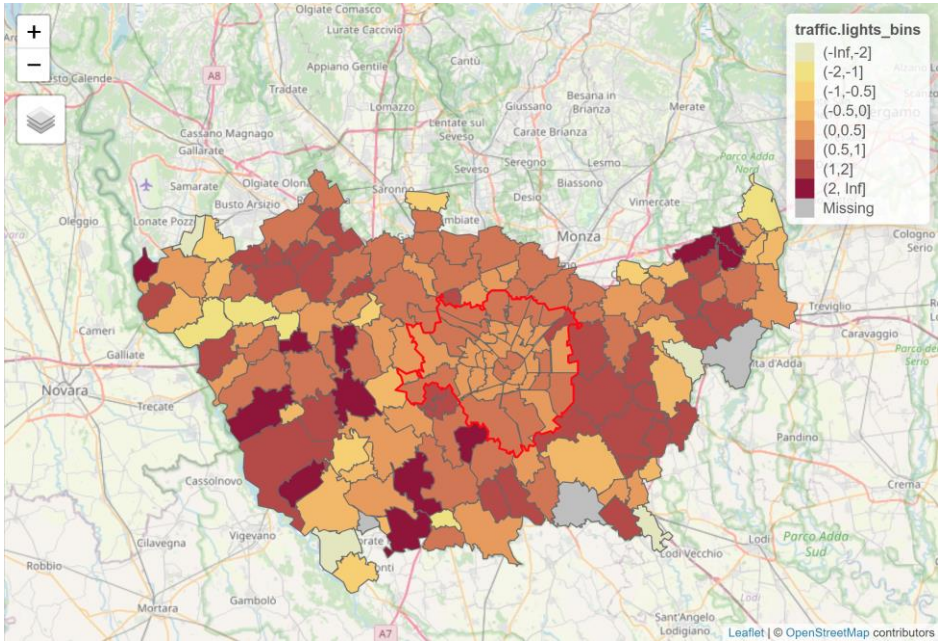
with geographical weights applied only to $\boldsymbol{X}_{(-z)}$

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



Relevance of features at zone level





Choropleth maps depicting the varying contributions of two distinct features to the risk assessment across different areas.

Conclusions

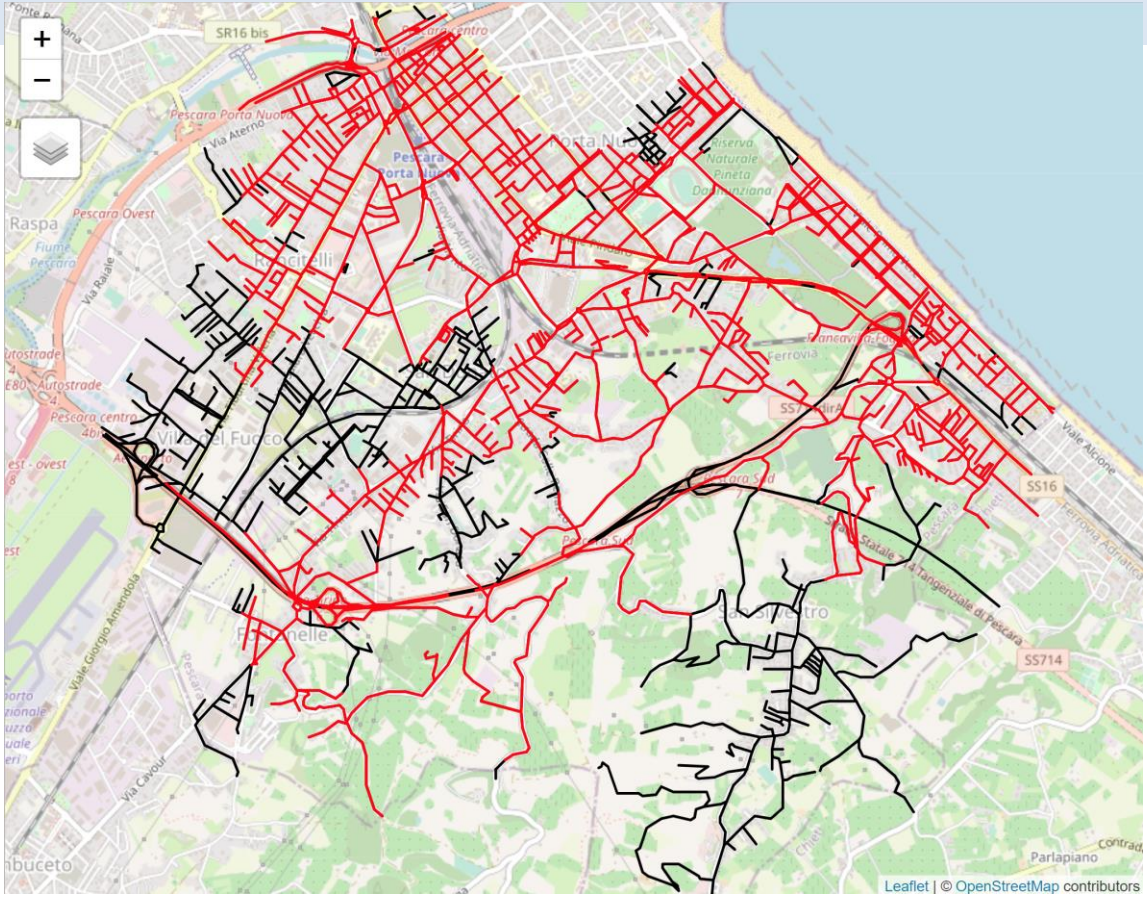
- The proposed approach exploits the use of open-source data to estimate the risk related to where the policyholder drives.

- It is a work in progress and several points are under investigation. At moment, we are evaluating the possibility of:
  - Validating the model using training and testing
  - Consider time dependence (scarcity of data per time unit might be present)
  - Testing ***Graph Neural networks*** including spatial dependence
  - Evaluating which improvements these results can offer for insurance pricing.
  - Improving results using other data (e.g. average speed per link)
  - Testing the model using data of other countries

Main References

- Assunção R., Azevedo Costa M., Oliveira Prates M., and Silva e Silva L.G.(2014) Spatial Analysis, in A. Charpentier (2014), Computational Actuarial Science with R, Chapman & Hall/CRC press,
- Barua, S., El-Basyouny, K., Islam, M.T., (2014). A full bayesian multivariate count data model of collision severity with spatial correlation. Analytic Methods in Accident Research 3, 28-43
- Blier-Wong, C., Cossette, H., Lamontagne, L., Marceau, E. (2022), Geographic ratemaking with spatial embeddings, Astin Bullettin
- Borgoni, R., Gilardi, A., Zappa, D. (2020), Assessing the Risk of Car Crashes in Road Networks, Social Indicators Research
- Boskov M. and Verrall R. J. (1994) Premium Rating at Geographic Area Using Spatial Models. ASTIN Bulletin, 24, pp 131-143
- Clemente G.P., Della Corte, F., Zappa, D. (2024), Hierarchical Spatial Network Models for Road Accident Risk Assessment, Annals of Operation Research
- Gilardi,A., Mateu, J., Borgoni, R., Lovelace, R. (2022), Multivariate hierarchical analysis of car crashes data considering a spatial network lattice, Journal of the Royal Statistical Society Series A: Statistics in Society
- Marshall et al. (2018), Street Network Studies: from Networks to Models and their Representations, Network and Spatial Economics
- Rashmi, R. et al. (2019), Analysis of Road Networks Using the Louvain Community Detection Algorithm, Soft Computing for Problem Solving.
- Tufvesson, O. et al. (2019) Spatial statistical modelling of insurance risk: a spatial epidemiological approach to car insurance, Scandinavian Actuarial Journal, 2019:6, 508-522
- Yao J. (2016) Clustering in General Insurance Pricing. In E. Frees, G. Meyers, & R. Derrig (Eds.), Predictive Modeling Applications in Actuarial Science (International Series on Actuarial Science, pp. 159-179). Cambridge: Cambridge University Press.
- Wuthrich, M. V., and C. Buser (2019), Data analytics for non-life insurance pricing, g. Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2870308

Validation procedure: train/testing



Main issues

- i.i.d. Train/testing selection cannot be applied
- Subgraphs must be connected and edges direction must return a feasible network
- Exact (or almost exact) desired train set dimension is in some cases not possible

=> The two subnetworks are correlated

Additional aspects based on network theory

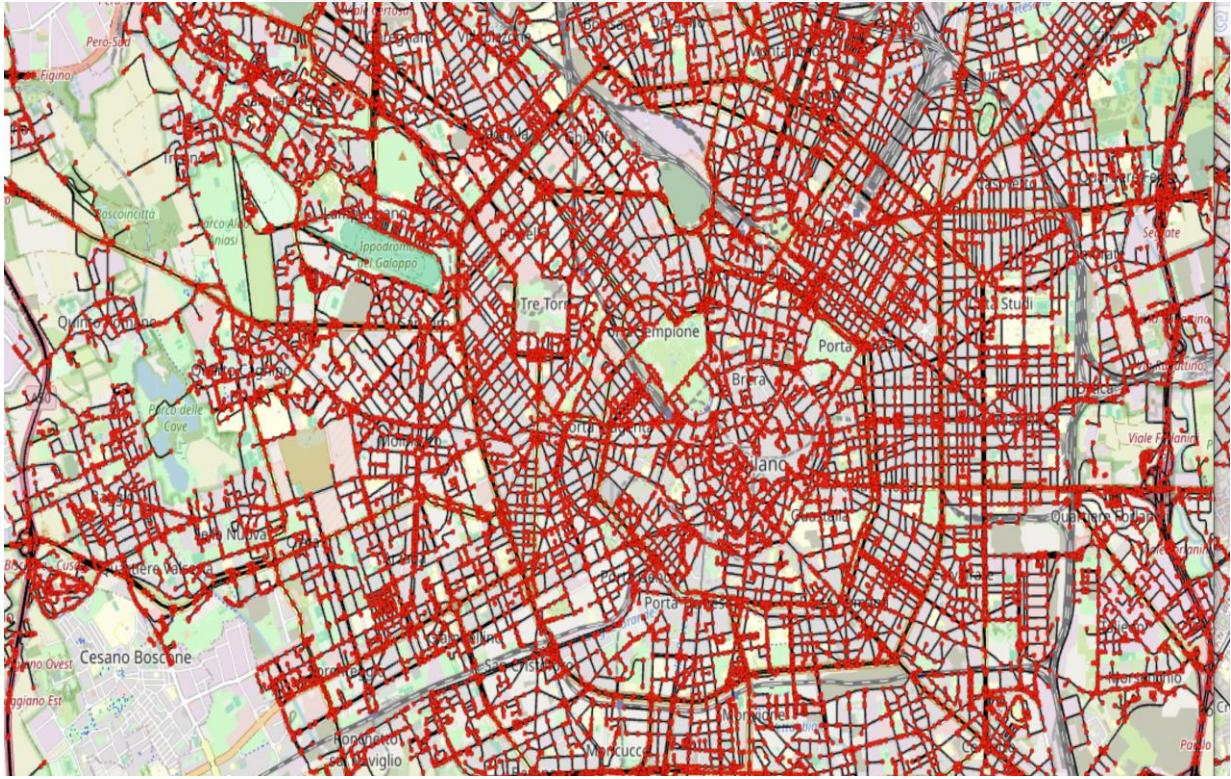
We deal now with two types of network:

- $G = (V; E)$ ***an unweighted network*** with n nodes (junctions/road terminations) and m arcs (road segments)
- $G_w = (V; E; \boldsymbol{W})$ ***a weighted network*** equal to G, where each arc ***is weighted according to the risk of the segment*** detected at previous step.

The unweighted directed network have the following characteristics:

- very sparse (density is close to zero)
- assortativity and transitivity coefficients are also very low

Example of network for city of Milan



Risk vs centrality

- We focus here on the topology of the network, ***assessing the global importance of network elements***.

- In particular, focusing on road segments and junctions, the node and edge betweenness appears as key indicators for this context. The node betweenness is a function of the number of shortest paths between pairs of nodes that pass through that node (see Newman, Girvan, 2004):

$$b_i = \sum_{\substack{h,k \in V \\ h \neq k \neq i}} \frac{n_{h,k}(i)}{n_{h,k}}$$

where $n_{h,k}$is the number of shortest paths between h and k and $n_{h,k}(i)$ is the number of shortest paths between h and k that passes through the node i. A similar definition can be provided in case of edges.

- Since the computation on the whole network G is really time consuming and does not provide significant value added, we considered separately nodes in the sub-graphs G_z based on the splitting of the whole network according to cities and zip codes.

Risk vs centrality