Insurance
Data
Science

Insurance Data Science Conference

Bayes Business School, City, University of London
Northampton Square, London EC1V 0HB

15 - 16 June 2023

Regular Session
Telematics
& graphs

Estimating road accident risk

Research Team: Diego Zappa (*presenter*)[1] , Gabriele Cantaluppi [1], Gian Paolo Clemente[2], Francesco della Corte[2], Jianyi Lin[1], Nino Savelli[2]

[1]Department of Statistical Sciences
[2]Department of Mathematics for Economics, Finance and Actuarial Science

Università Cattolica del Sacro Cuore - Milan (Italy)

what are we presenting? (1)

- 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 policy makers: 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 onto 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 (blakboxes recordings are necessary)

what are we presenting? (2)

- **In particular we focus on “where the policyholder drives”**
  - 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., Gao. Meng, Wuthrich (2022), Huang, Meng (2019), Wuthrich and Buser (2019), Ayuso et al. (2018), …)
  - Driving habits (e.g. KM, daytime, weather conditions, etc.)

- **We follow a combined approach:**
  - **Penalized regression and spatially lagged models** 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.
- ☛ We will mainly focus on results and problems

Which “ingredients” did we use?

Road details (e.g. Open Street Map)

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

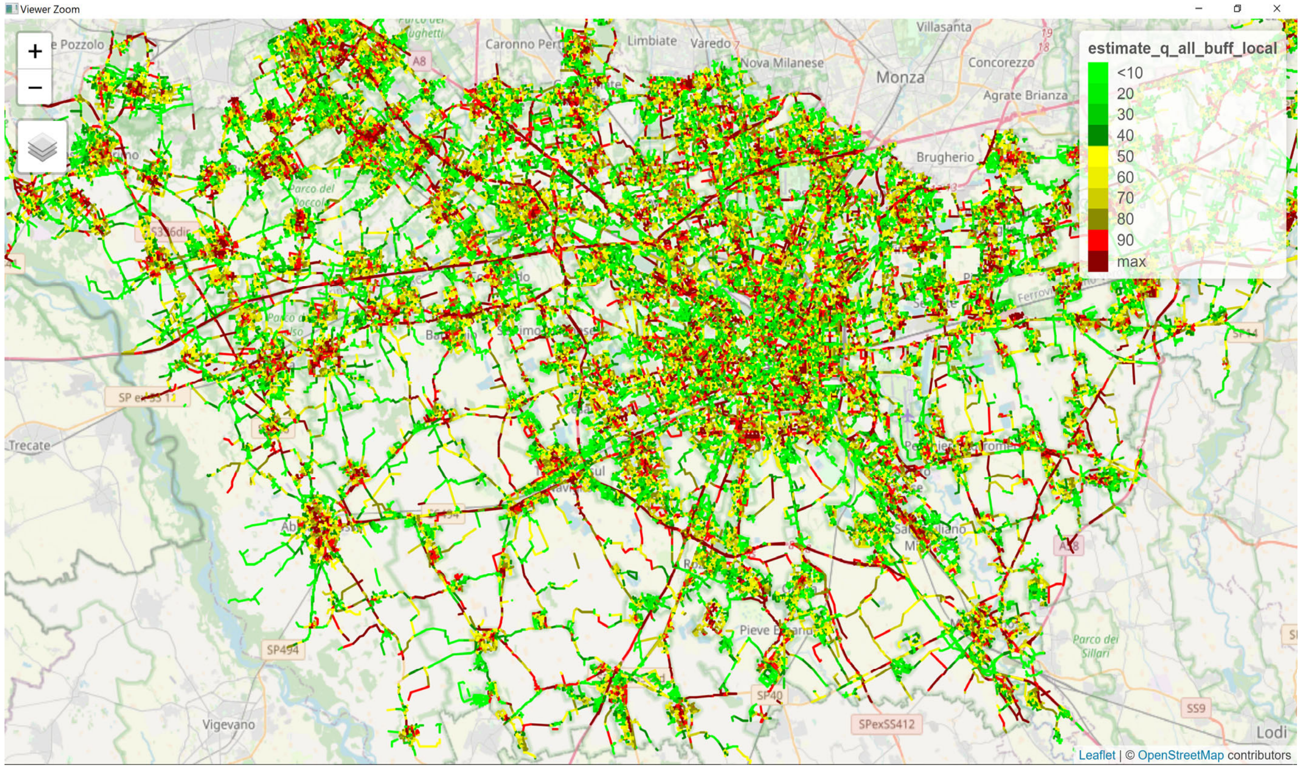
Demographic database (population density, building density, commuting people)

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

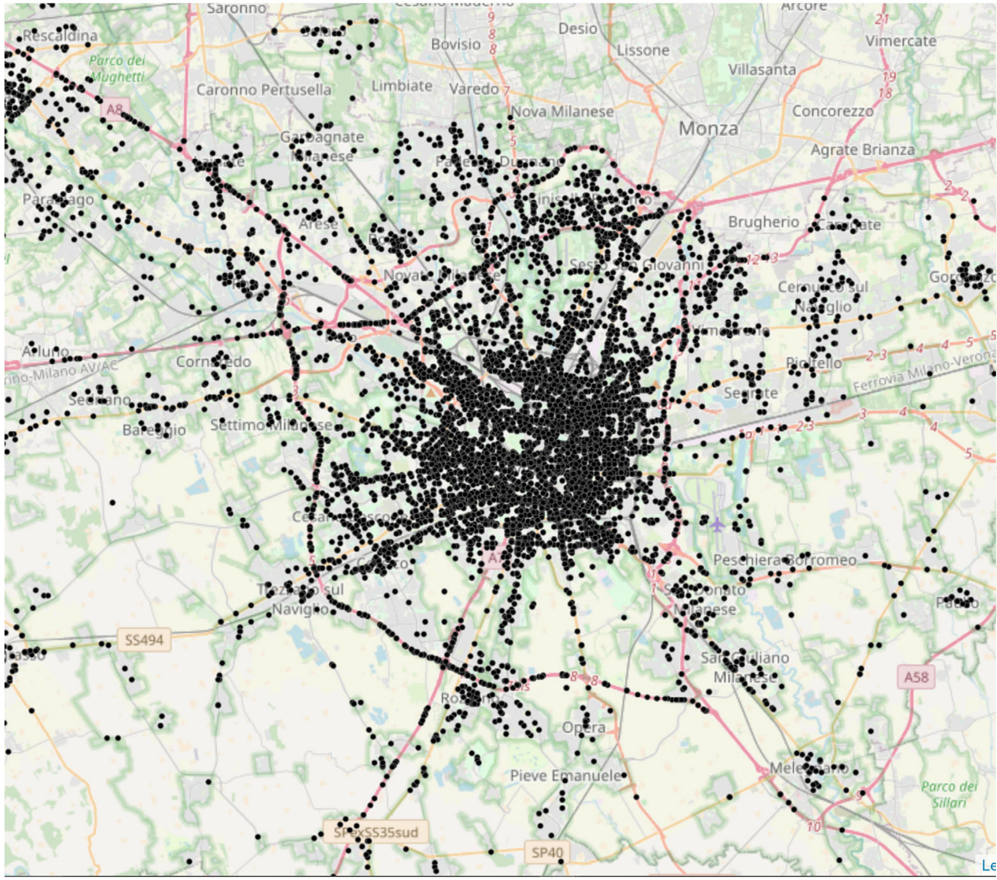
Location of accidents (e.g. from insurance companies, open data, ISTAT)

Weather conditions

$$E(\#Accidents) = g(features, \theta) + \text{offset}$$



The data



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
- Etc.

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 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 ISTAT)

- In general, datasets contains location of accidents.
- Reverse geocoding (i.e. lat/long coordinates) algorithms are necessary but othen 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),.Much more precise db allow to have knowledge of the road levels.

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

Some details about

$$E(\#Accidents) = g(features, \theta) + offset$$

#Accidents

- Data fitting involves (so far) only ***crashes that resulted in fatalities or injuries of at least one person***. We do not consider in this presentation accidents with damages only to vehicles or objects (data are often proprietary).
- **Time dependence** is somewhere measurable. It is not true for any subregion. At moment, data refers to quarters of 5 years from 2016 to 2020. Data of 2021 arrived 2w ago. It is in process.
- **Spatial dependence** is mostly retated to the network structure dependence.

g() (see next slide)

Features

- most of covariates are categorical or transformed into classes to allow estimate comparison among regions
- spatial dependences is considered
- distance matrix

Offset

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

The general model currently tested

Literature often makes use of hierarchical models. E.g.

At the first stage : $Y|\lambda \sim Poisson(\lambda)$
At the second stage: $\log(\lambda) = X\beta + \theta + \varphi$
At the third stage: the specification for the priors

Fit involves INLA methods

⇒To include spatial dependence of features we have used the Spatial Durbin model (SDM)

$$\log(\lambda) = (I_n - \rho W)^{-1}(X\beta + WX\gamma + \theta)$$

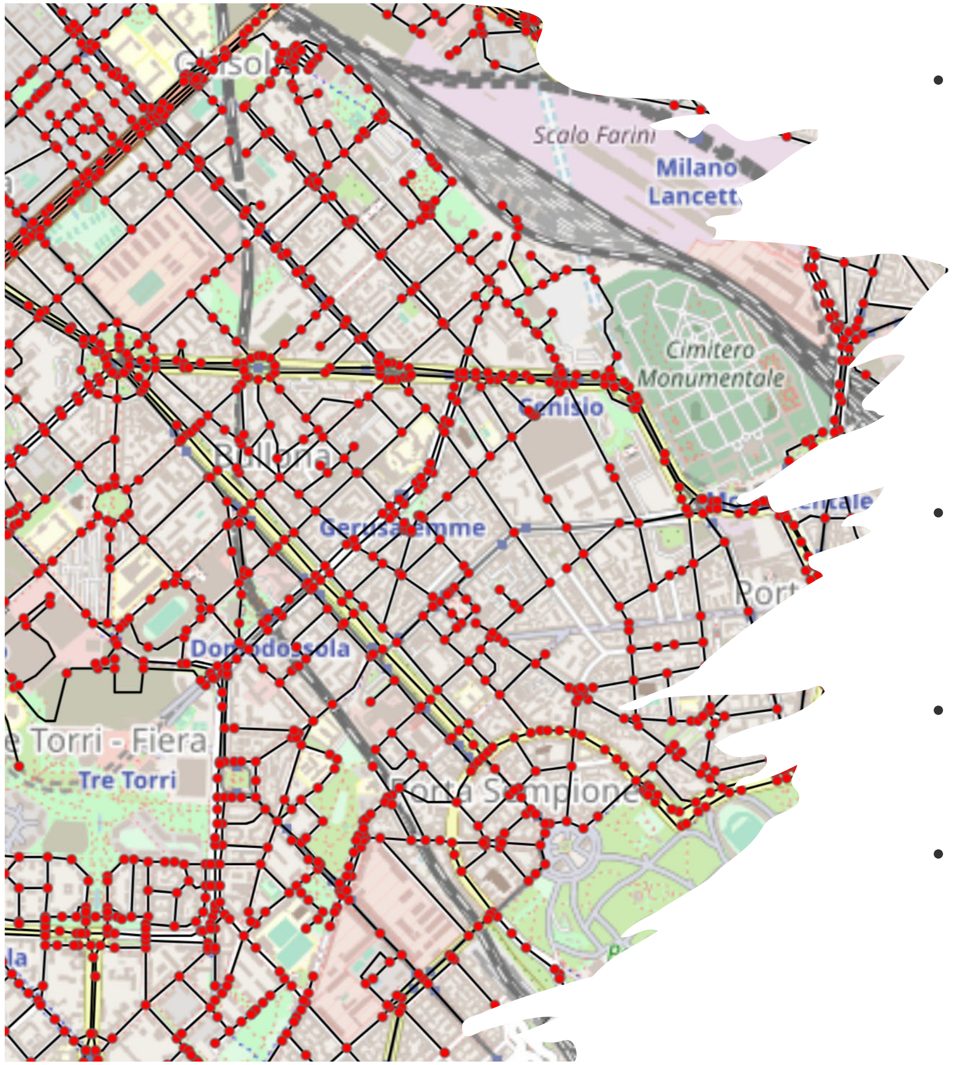
It allows for spatial correlation of the outcome as well as it considers the lagged matrix of covariates as regressors like the SLX model.

=> We have tested

- Spatial lagged models in a bayesian framework
- Elastic net to consider model flexibility in a frequentist framework
- Local graphical neural networks (*research in progress*)

Insurance
Data
Science

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).

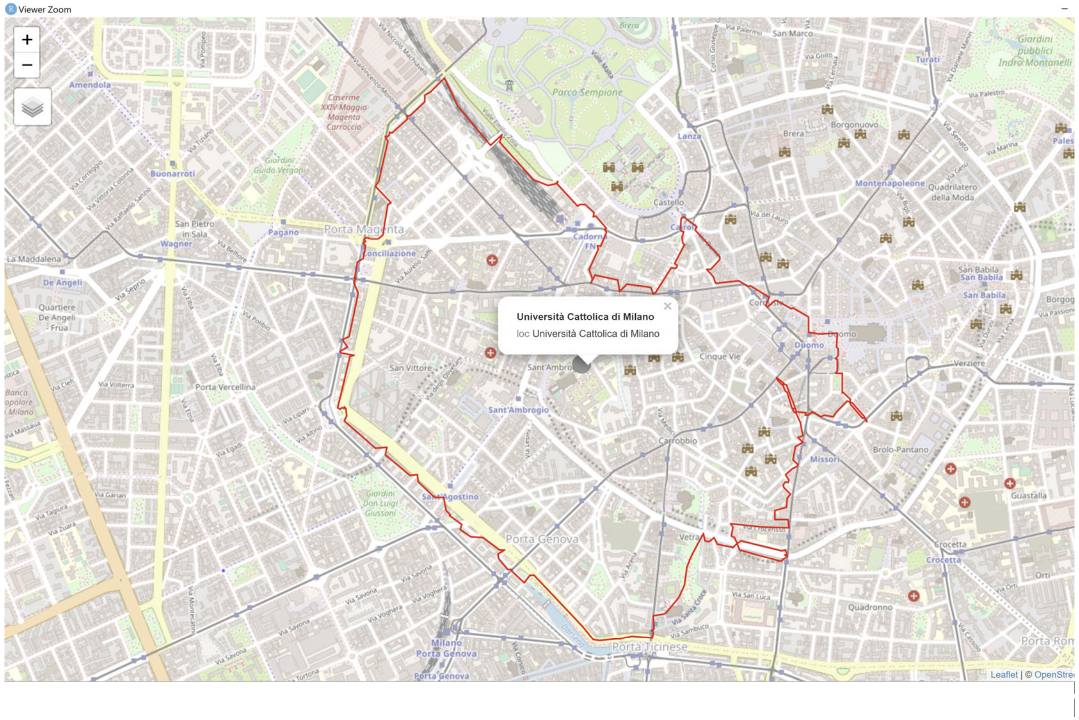
  Formally, given the street network, we build a graph $G = (V; E)$ where V and E are respectively the set of n vertices and m arcs. Two nodes are adjacent if there is an arc $(i, j) \epsilon E$ (i.e. a road segment) connecting them

- 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*.

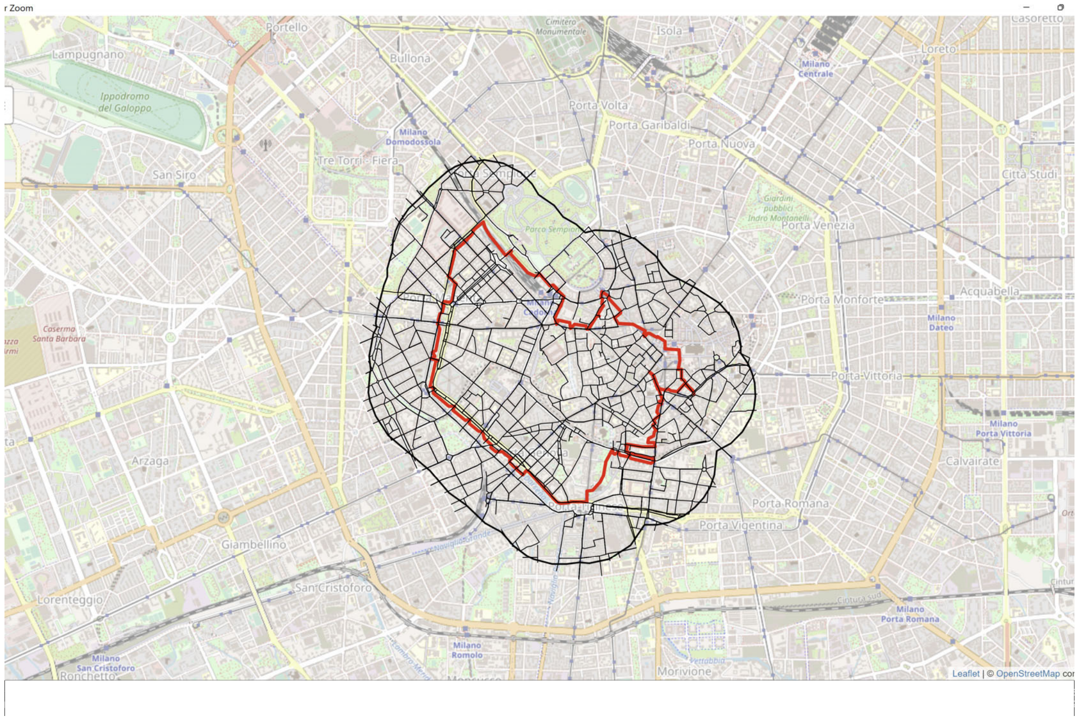
- Distances between two roads have been computed by adding centroid to each segment and by computing the directed weighted shortest path between two centroids.
- The **shortest path problem** is the problem of finding a path between two nodes in a graph such that the sum of the weights of its constituent edges is minimized.

Insurance
Data
Science

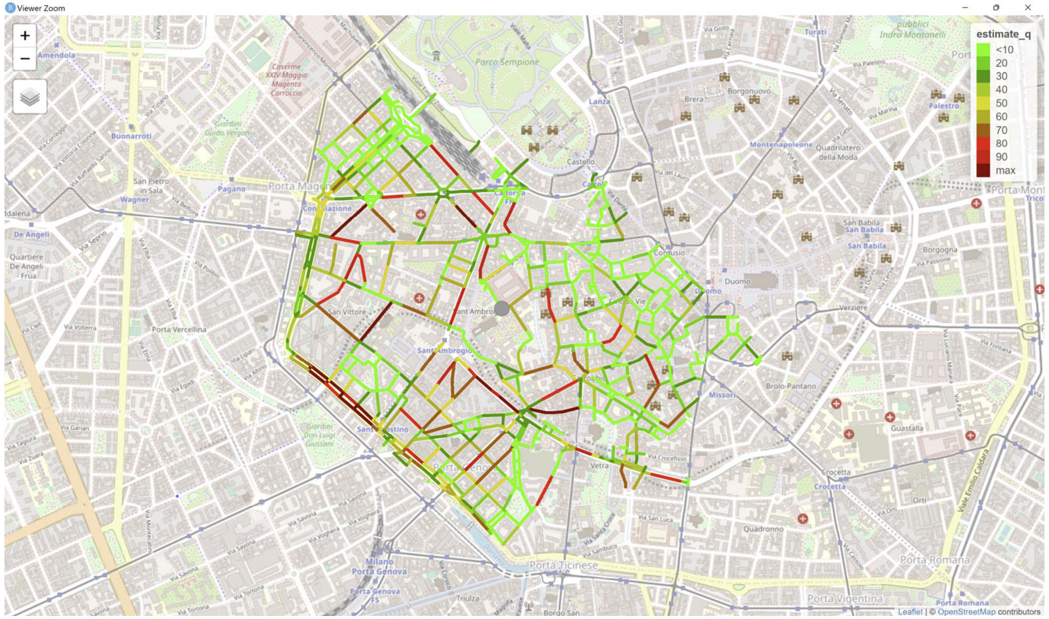
An example

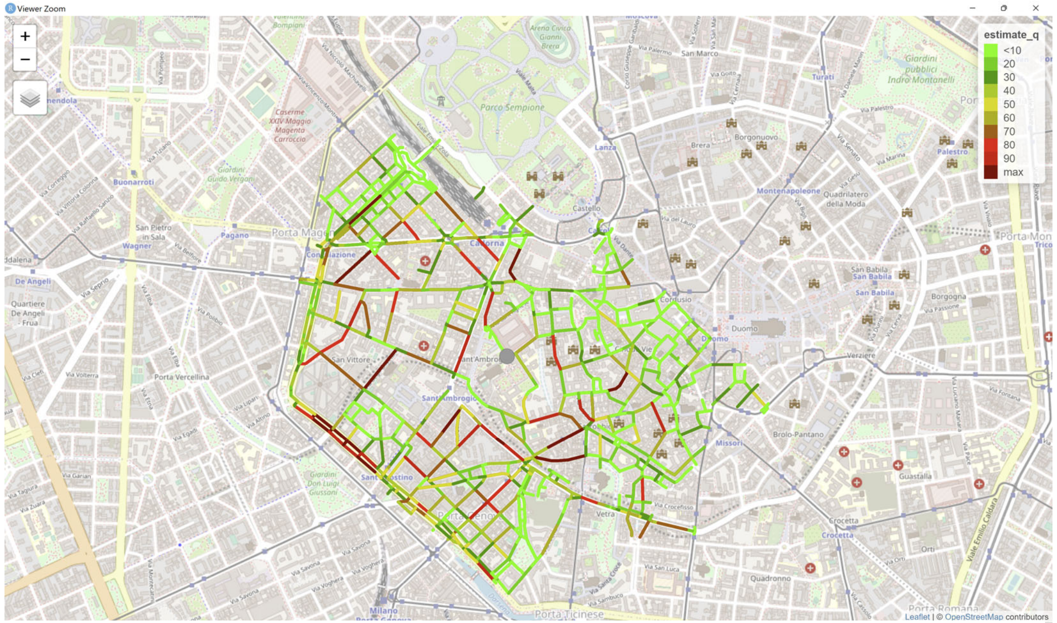


Università Cattolica di Milano
loc Università Cattolica di Milano



An example





Glmnet

Input: 38 predictors

Selected: 16

Execution time: 1’38"

SDM

Input: 38 predictors

Selected: 16

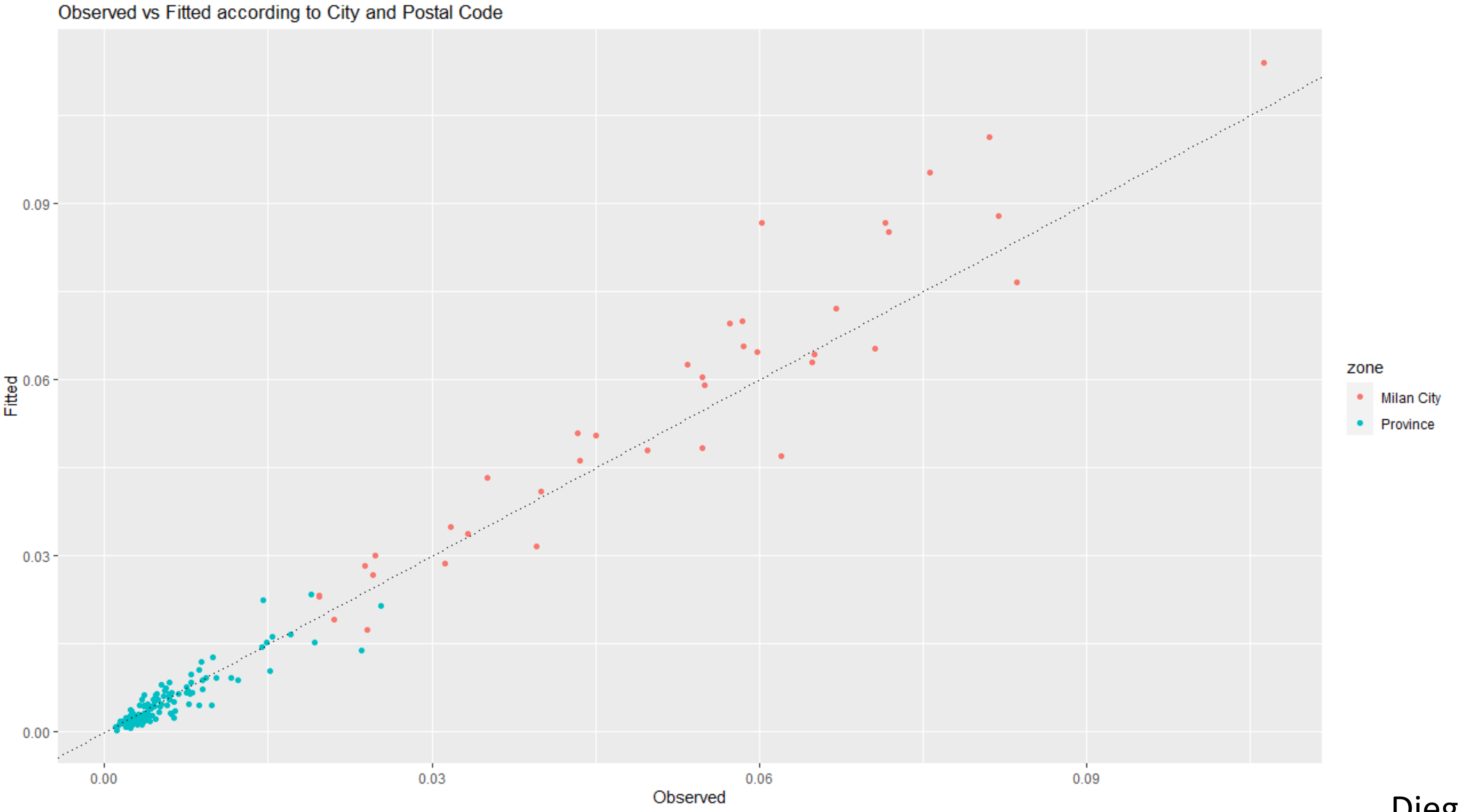
Execution time: 6’12"

Model hyperparameters:

```
                         mean    sd 0.025quant 0.5quant 0.975quant  mode
Precision for id_link 0.417 0.019      0.379    0.417      0.455 0.418
Rho for id_link       0.502 0.062      0.372    0.506      0.612 0.521
```

Insurance
Data
Science

Fit vs observed (at an aggregate level)

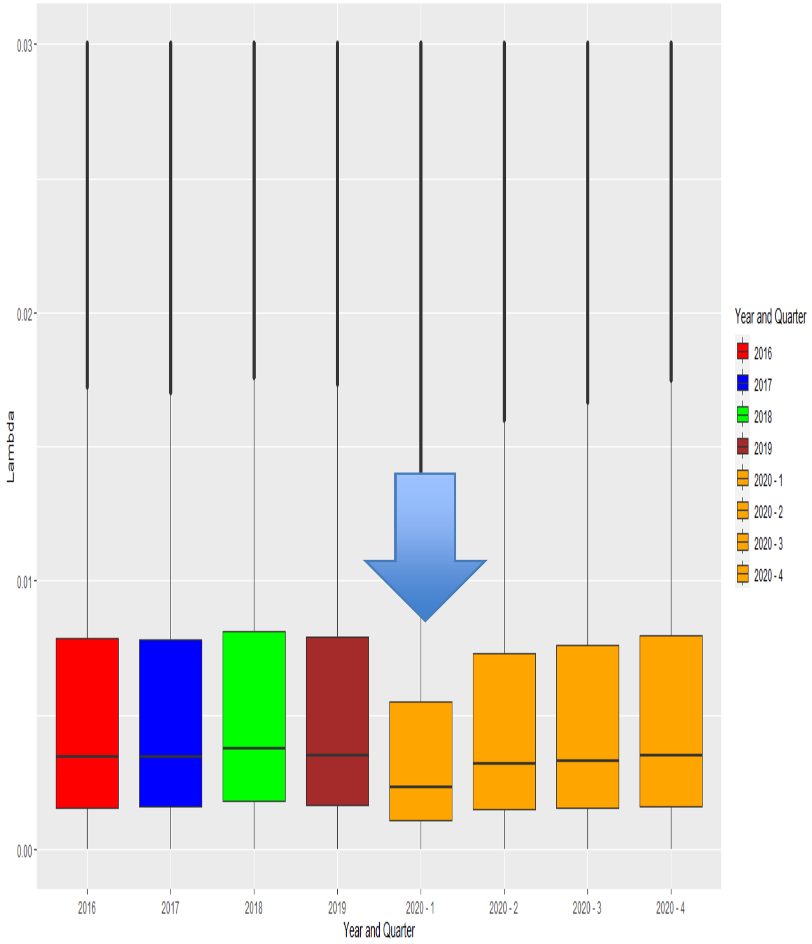


Diego Zappa et al.

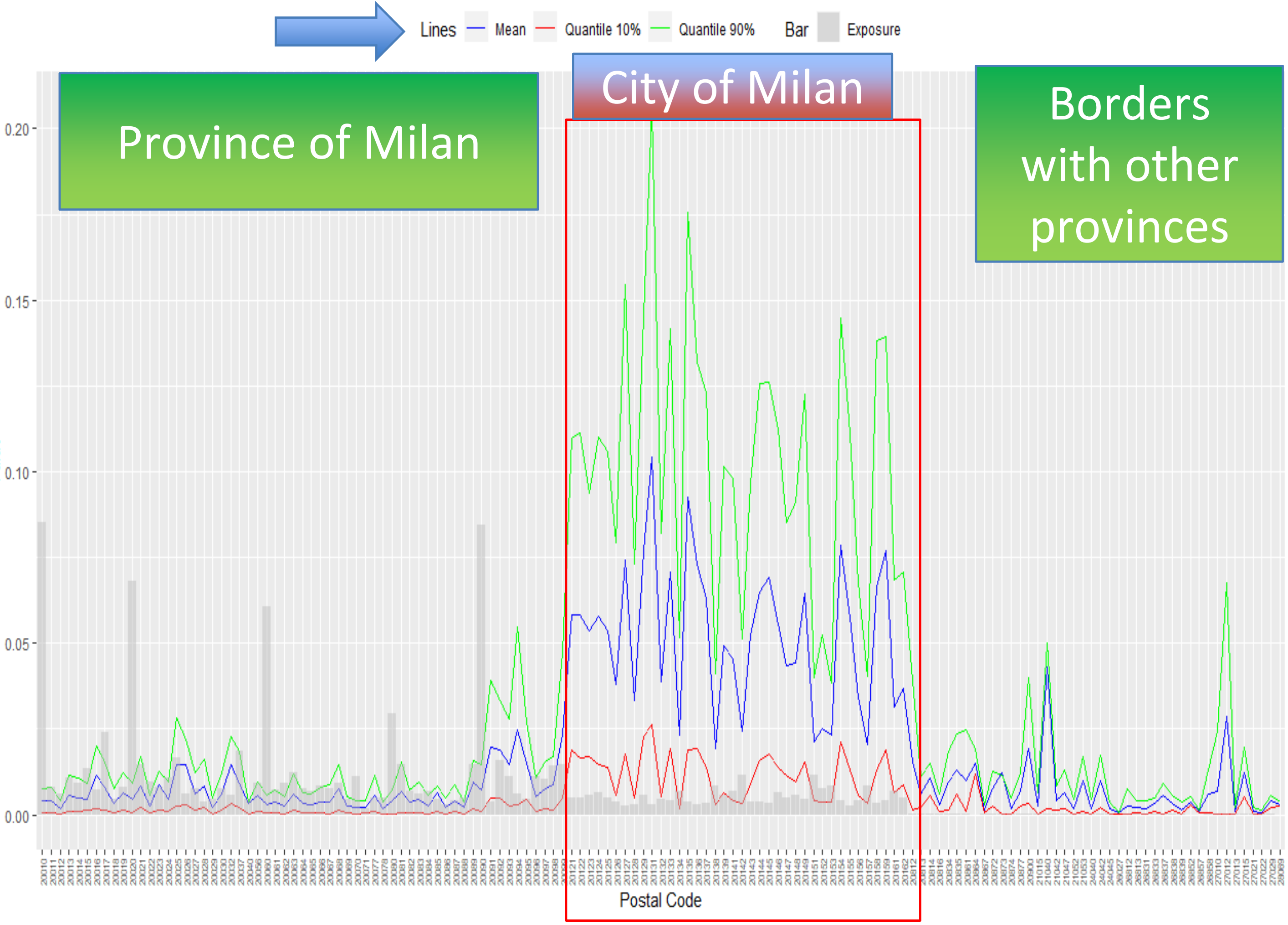
Insurance
Data
Science

Risk based on time and area

Risk wrt to time



Mean, 90% and 10% quantiles by postal code



Province of Milan

City of Milan

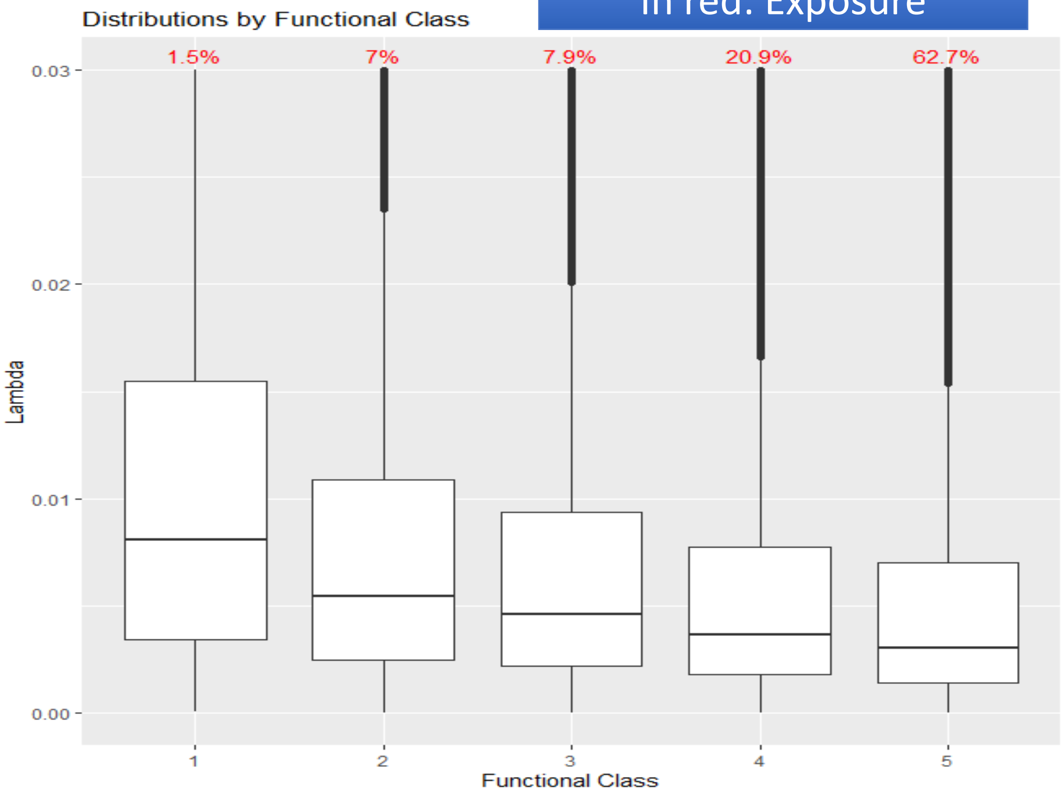
Borders with other provinces

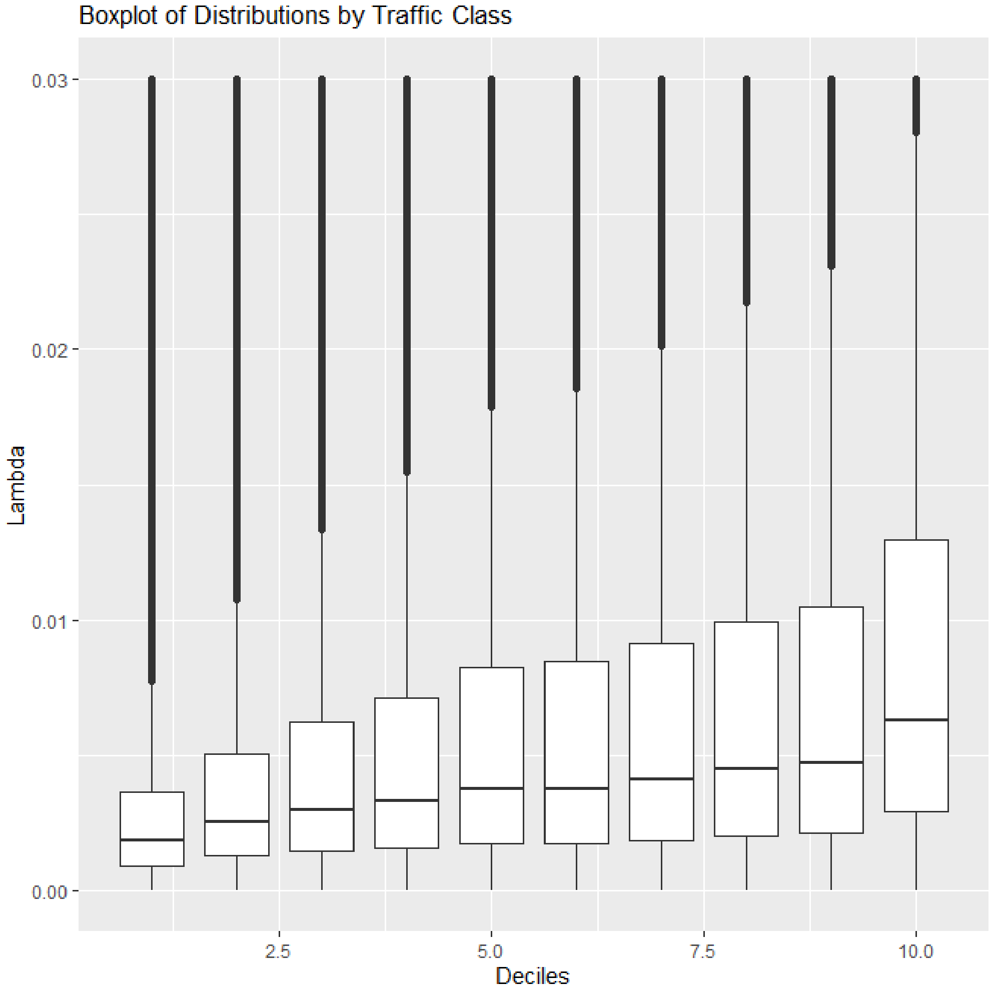
Insurance
Data
Science

The effect of traffic

Functional Class =1: roads allow for high volume, maximum speed traffic movement between and through major metropolitan areas.
Functional Class = 5 is applied to roads whose volume and traffic movement are below the level of any functional class. In addition, walkways, truck only roads, bus only roads, and emergency vehicle only roads.

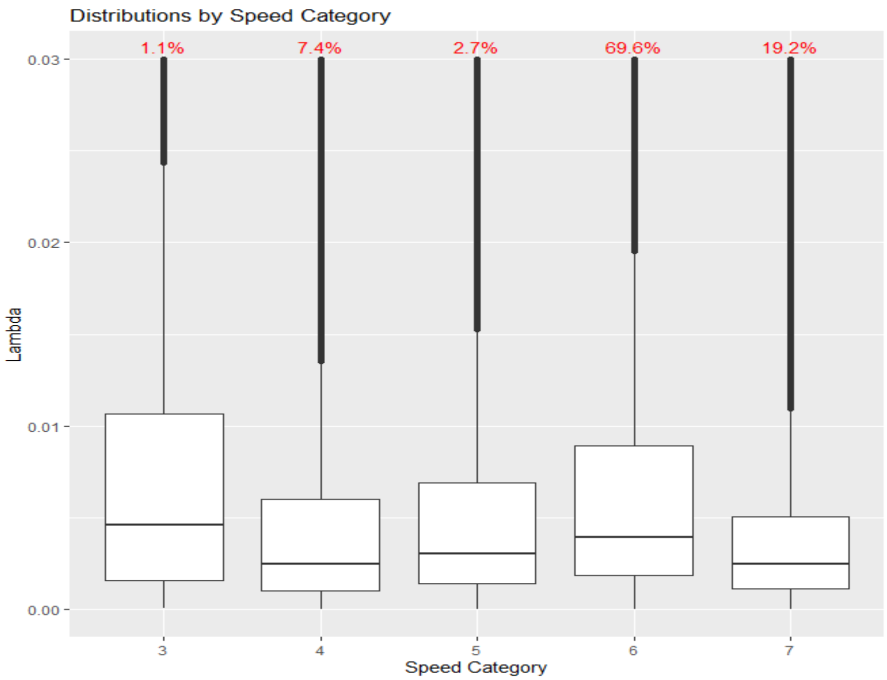
In red: Exposure



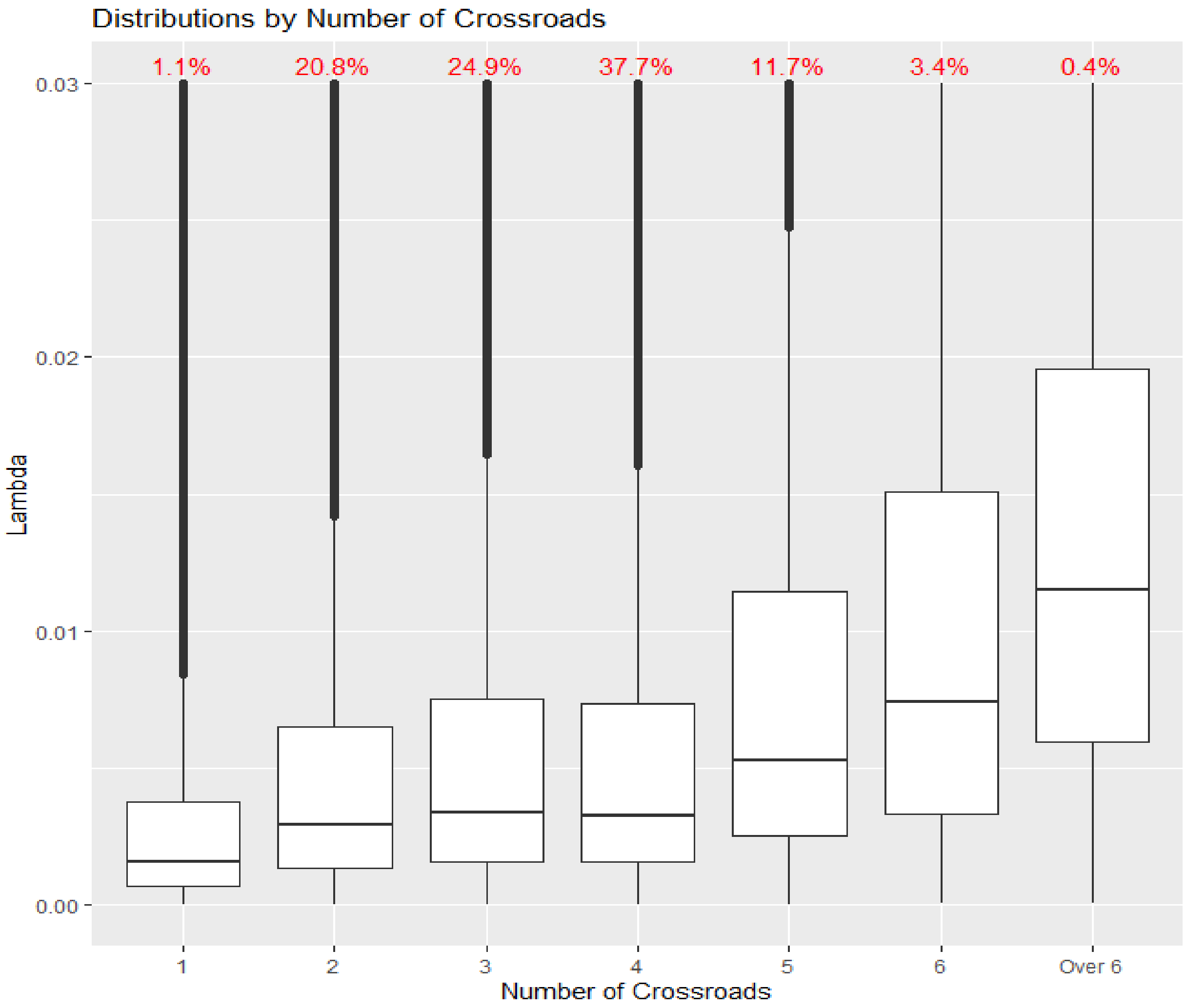


Insurance
Data
Science

Speed and Number of crossroads

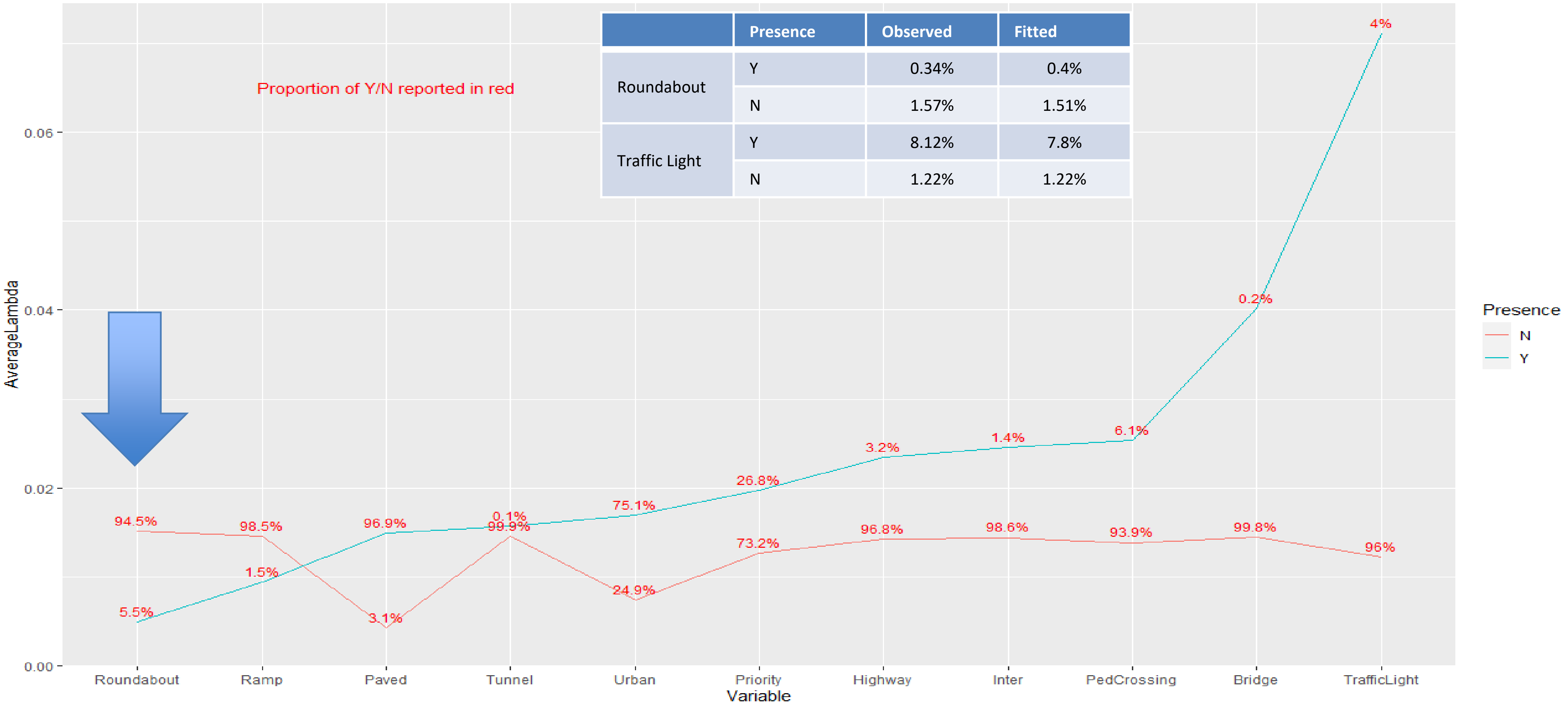


Speed Category	3	4	5	6	7
KPH	>90	71-90	51-70	31-50	<= 30
MPH	>54	41-54	31-40	21-30	<= 20



Insurance
Data
Science

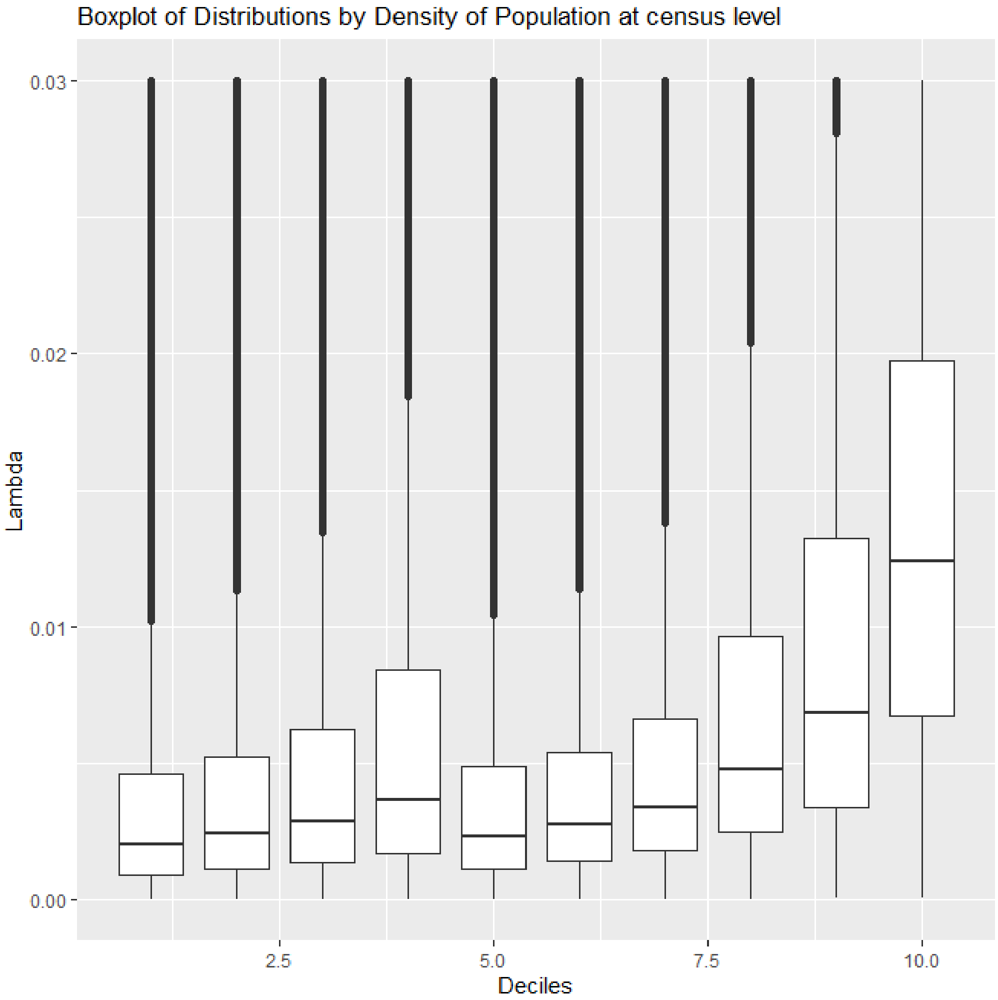
Other characteristics

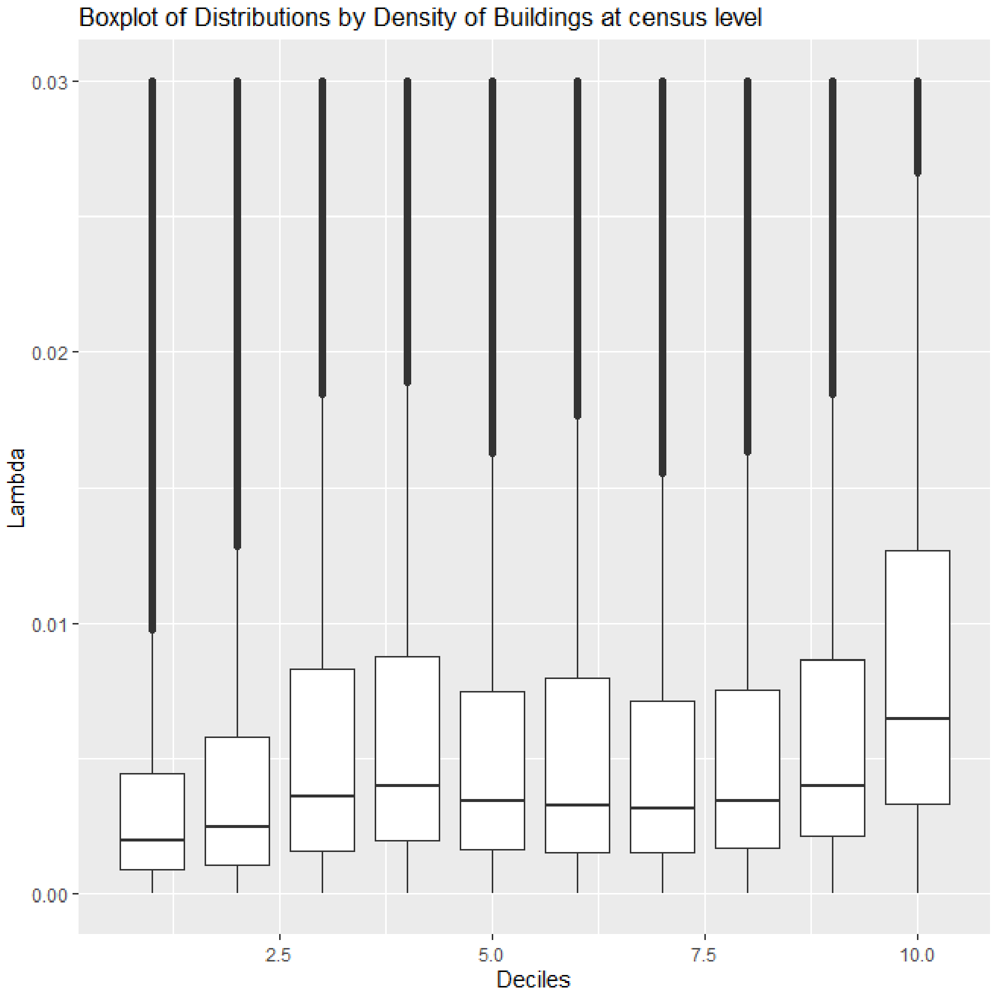


	Presence	Observed	Fitted
Roundabout	Y	0.34%	0.4%
	N	1.57%	1.51%
Traffic Light	Y	8.12%	7.8%
	N	1.22%	1.22%

Insurance
Data
Science

Density
Population and Buildings





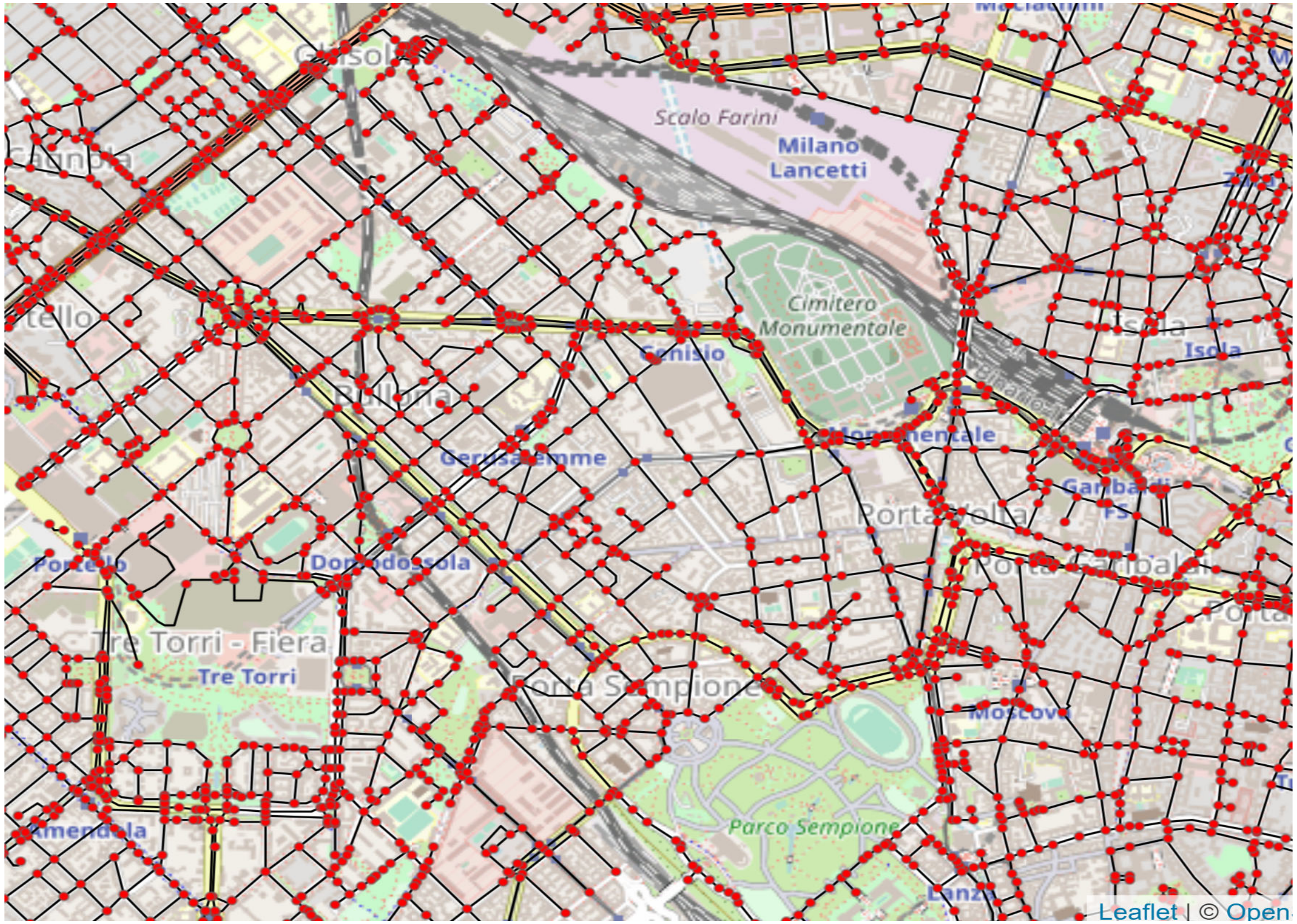
Additional aspects based on network theory

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.

Considering only city of Milan and province, we have an unweighted directed network with the following characteristics:

- roughly 138 thousand of nodes and 3.3 milion of arcs.
- very sparse (density is close to zero)
- assortativity and transitivity coefficients are also very low



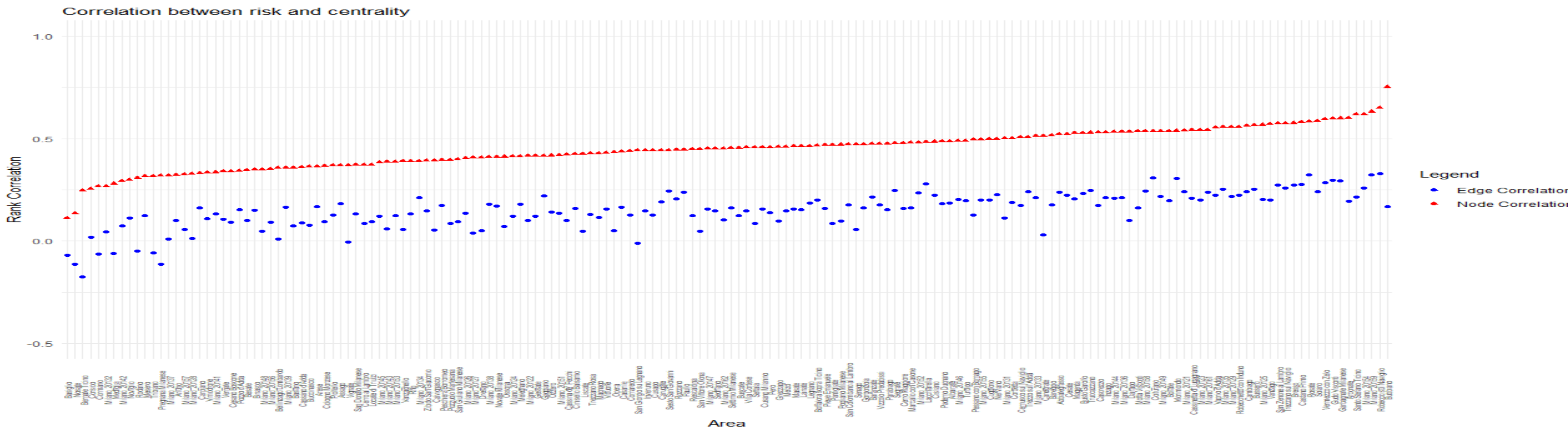
Risk vs centrality

- 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):

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

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.

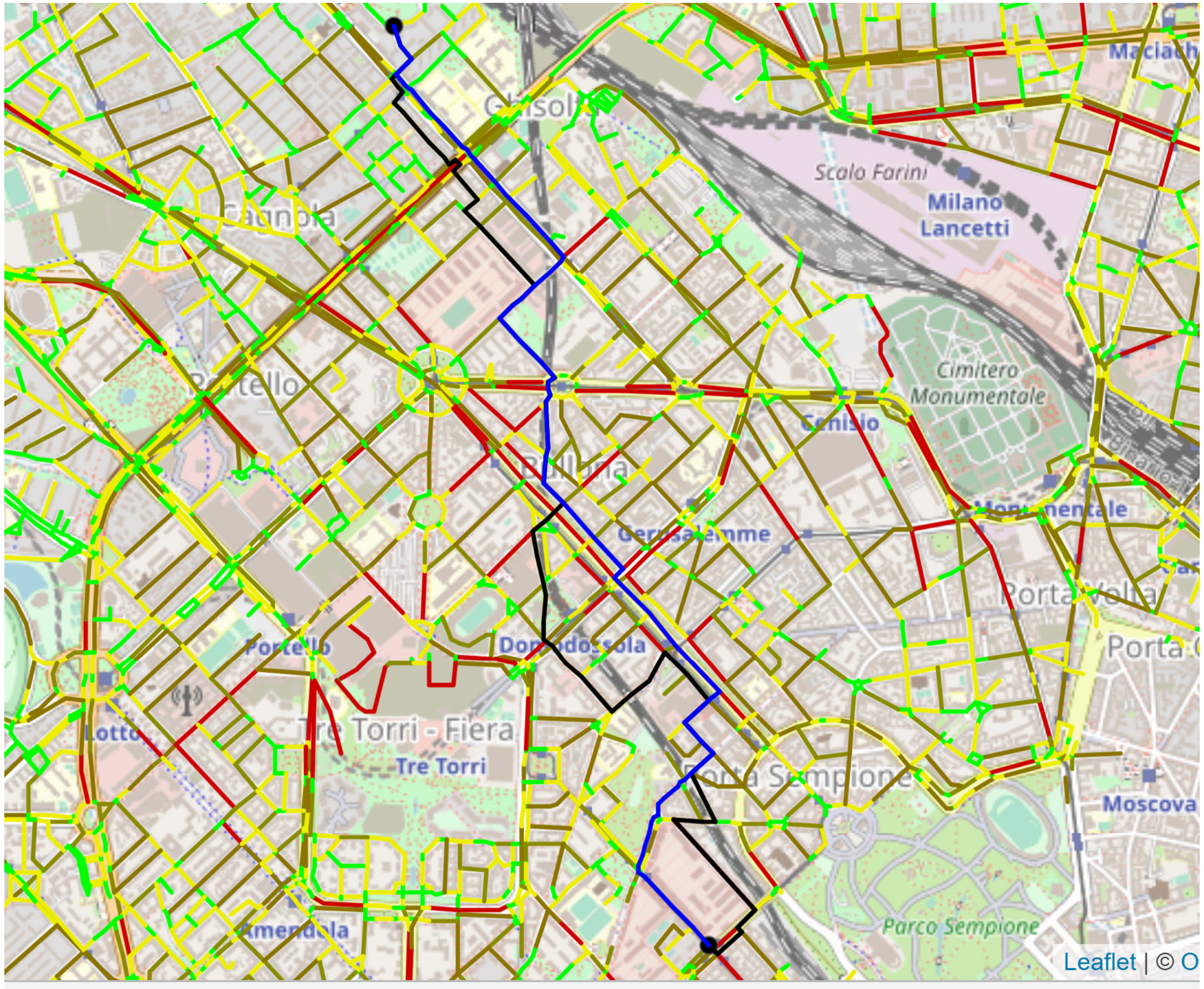
- Since the computation on the whole network G is really time consuming and does not provide significant value added, we considered separately nodes in the sub-graphs G_z based on the splitting of the whole network according to cities and zip codes.



Minimum Length vs Minimum Risk

- We apply the shortest path between two points.

- **In the plot, the shortest path has been applied considering:**
  - **The minimum length (in blue)**
  - **The minimum risk (in black)**



Insurance
Data
Science

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. In particular, at moment, we are evaluating the possibility of:
  - Improving results merging average speed per link
  - Extending results (computational issues are present)
  - Consider time dependence (scarcity of data per time unit might be present)
  - Validating the model using training and testing
  - Testing Neural networks including spatial dependence
  - Evaluating which improvements these results can offer for insurance pricing.
  - Testing the model using data of other countries

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*
- 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 (2022), Data analytics for non-life insurance pricing, g. Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2870308