# Spatial Logistic Regression for Events Lying on a Network: Car Crashes in Milan

## *Regressione logistica per eventi su network: gli incidenti automobilistici nel comune di Milano*

Andrea Gilardi, Riccardo Borgoni and Diego Zappa

**Abstract** In this paper we propose a methodology to estimate the probability that a car accident occurs in urban roads. Our approach is based on logistic regression and takes into account the particular nature of the data which conforms to a spatial point pattern on a network. Using the open data on street networks provided within the OpenStreetMap project, we estimate the probability of car accidents for every street in the municipality of Milan.

**Abstract** *In questo lavoro viene presentato un approccio basato sulla regressione logistica per stimare la probabilità che avvenga un incidente automobilistico sui tratti di strada urbani. La metodologia proposta tiene conto della natura dei dati disponibili, assimilabili ad eventi casuali su un supporto spaziale di network. Utilizzando gli open data disponibili dal progetto OpenStreetMap, la stima è ottenuta per la totalità delle strade presenti sul territorio comunale di Milano.*

**Key words:** urban geography, car accidents, open data

## 1 Introduction

A precise definition of Smart City is extremely difficult. Nevertheless, one of the key point that several authors agree upon is that a smart city should use modern technologies to improve urban traffic and street safety [8]. During the last years the European commission launched a new campaign to improve the transportation

Andrea Gilardi
University of Milano - Bicocca, e-mail: a.gilardi5@campus.unimib.it

Riccardo Borgoni
University of Milano - Bicocca

Diego Zappa
Catholic University of Milan

system, promote sustainable mobility and avoid road congestion, accidents and air pollution [2].

In this paper we estimate the probability that at least one car accident occurs for every street in Milan considering a dataset reporting the spatial location of the accidents.

Hereafter, it is assumed that the occurrence of accidents in the city area conforms to a spatial point process. Since the likelihood of a planar Poisson point process can be approximated by the likelihood of logistic regression for the discretised process, pixel-based logistic regression is now commonplace to analyse the spatial dynamic of a point pattern in applied statistics and GIS literature. The connections between the two approaches are thoroughly discussed in [1]. In addition, resorting to a GLM model is particularly convenient when one is interested in the potential effect of a number of covariates.

Car accidents, however, are a classical example of a point pattern occurring on a linear network [4]. In these circumstances, the usual statistical techniques designed for point patterns occurring on the plane can not work since it is necessary to take into account the fact that the events are constrained to lie on the network.

In this paper, we propose a spatial logistic regression to model network binary data akin to the pixel-based logistic regression approach. However, given the nature of the process, we abandon the idea of dividing the space into pixels. Instead, we split the street network of Milan into smaller segments recording the presence of car crashes into each segments and fit a logistic regression model.

We underline the fact that, whenever it was possible, we chose to use open access data. They are becoming more and more popular in the road safety research methods (e.g. [6] and [3]) but they are still of very limited use in the Municipality of Milan. Nevertheless, we strongly believe in their importance for a digital, sustainable and safe development of a Smart City, as it was suggested by [7].

## 2 Data and methods

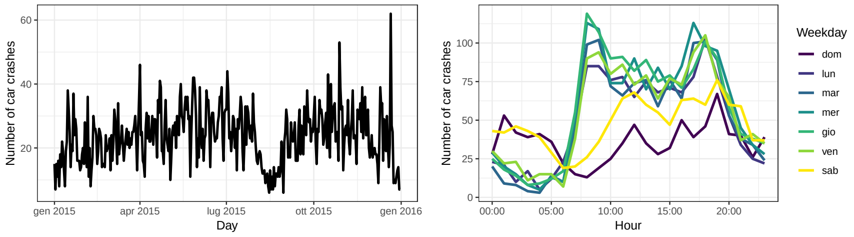
We consider accidents occurred between the 1st of January and 31st of December 2015, that require an ambulance intervention and for which the exact time and location, geocoded in UTM coordinates, were recoded. The sample includes 8601 events. Their annual temporal distribution is reported in Figure 1(a) and their hourly temporal distribution is reported in Figure 1(b). Data show some seasonality. Firstly, the number of interventions per day exhibits some peaks in winter (in particular during the Christmas holidays) and a drastic reduction in the last days of July and in the first days of August (at the beginning of summer holidays). Secondly, the hourly temporal distribution of car crashes is deeply different between working days and weekends.

The road network was built using data from OpenStreetMap [5], which is a project that aims to build a free and editable map of the world with an open-content license. The basic components of OpenStreetMap data are called *elements* and they consist of: *nodes* (related to points on the earth surface), *ways* (which is a list of nodes and the most important structure for our model) and *relations* (describing the interactions between nodes and ways). Every physical object in the landscape is represented by these three elements and its attributes are stored using a `tag`, which is simply a pair of items which identify a category, a `key`, and the corresponding `value`[1], for instance `street=motorway` or `name="Viale Sarca"`. It is important to point out that almost all streets are internally stored as the union of a set of *segments* and the logistic regression model presented later on in the paper is based on this particular segmentation.

We downloaded data for all *trunk*, *primary*, *secondary*, *tertiary*, *unclassified*, *residential* and *service* highways for a total of 34085 segments with different length. Although there is no exact relationship between OpenStreetMaps tags and Italian Administrative road classification, the selected levels range from *Strada statale* to *Strada vicinale*. In order to apply a spatial logistic regression model, we projected all car crashes to the nearest point belonging to the linear network and the final result is reported in Figure 2.

For each segment of a street, a binary response variable $Y$ is defined taking value 1 if at least one car accident occurred in the segment and 0 otherwise. Given the high granularity of the network, this procedure resulted in a slightly imbalanced response variable with 30149 *zeros* and 3936 *ones*. Notice that, in the rare cases (approximately 1 every 200 interventions) when a car accident got projected exactly to the common boundary point of two touching segments, then it was assigned to both of them.

The covariates included in the logistic regression are: the segment length, its OpenStreetMap classification, and the total number of crossed segments, the latter used as a proxy of urban traffic. A logarithm transformation is applied to the segment length, given the strong skewness of its distribution.



(a)

(b)

Fig. 1: Temporal distribution of ambulance interventions in case of car crash per day of the year (a) and per hour and day of the week (b) during 2015

[1] A complete list of all the keys and the corresponding values can be found at `https://wiki.openstreetmap.org/wiki/Map_Features`

Using the iteratively reweighted least squares (IWLS) algorithm implemented in the R function `glm`, we fitted the following logistic regression model

$$logit(\mathbb{P}(Y = 1|X_1, X_2, X_3)) = \beta_0 + \beta_1 \log(X_1) + \beta_2 X_2 + \beta_3 X_3, \qquad (1)$$

where $logit(p) = \frac{p}{1-p}$, $X_1$ denotes the segment length, $X_2$ is a categorical variable representing its OpenStreetMap classification and $X_3$ is the number of touching segments.

## 3 Results and conclusions

The main results obtained applying the methodology presented above to the car accident data of Milan can be summarised in the following points:

1. the length of the segment is found to be the most important determinant to evaluate its risk with longer segments having an higher probability of car crashes;
2. *Secondary* and *Tertiary* segments (partially corresponding to *Strada Statale* and *Strada Provinciale*) are classified as the most dangerous roads regarding car crashes.
3. the number of touching segments does not influence the probability of car crashes, which probably means we are just using a bad proxy for urban traffic.

A graphical representation of the estimated probability of car crashes is reported in Figure 3. The segments far from the city centre exhibit higher probability of car

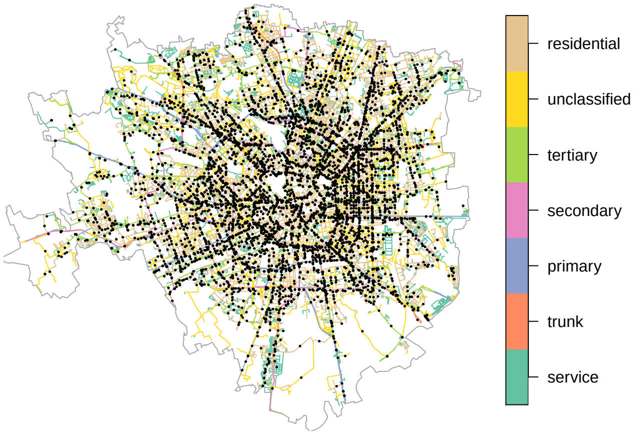


Fig. 2: Spatial representation of all car crashes in Milan during 2015 that required the intervention of an ambulance. Every street is colored according to its OpenStreetMap classification.

accident. The goal of this study is to propose a procedure to estimate a car accident risk index for every street in the Municipality of Milan taking into account the network nature of the spatial support of the data appropriately. Nevertheless this is just a preliminary result towards the development of a safety index and several enhancements are possible, in particular regarding the data and modelling perspective. Firstly it may be relevant to add demographic (e.g. population density), economic (e.g. number of vehicles per household) and traffic (e.g. total network volume and average travel speed) variables to the logistic regression model to improve its predictive performances. Secondly, alternative specifications for the link function could be considered. Finally, the classic independence assumption of GLM models can be somewhat questionable for this particular application. Hence, a reasonable extension of the model proposed in this paper is to take into account the potential spatial autocorrelation of adjacent segments.

**Acknowledgements** Map data copyrighted OpenStreetMap contributors and available from `https://www.openstreetmap.org`.

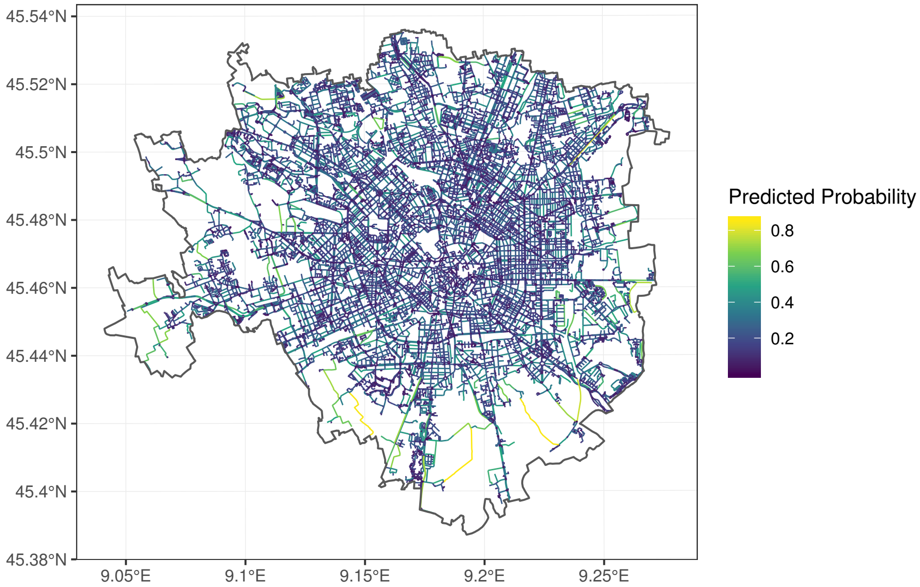


Fig. 3: Estimated probability of accident occurrence of the street segments displayed in Figure 2 based on a logistic regression model. Segments far from the city centre exhibits higher probability of car crashes.

## References

1. Baddeley, A.; Berman, M.; Fisher, N.I.; Hardegen, A.; Milne, R.K.; Schuhmacher, D.; Shah, R.; Turner, R. Spatial logistic regression and change-of-support in Poisson point processes. Electron. J. Statist. 4 (2010).
2. European Commission: European Initiative on Smart Cities, 2010–2020. `http://setis.ec.europa.eu/set-plan-implementation/technology-roadmaps/european-initiative-smart-cities`
3. Lovelace et al., (2019). stats19: A package for working with open road crash data. Journal of Open Source Software, 4(33), 1181.
4. Okabe A, Sugihara K (2012) Spatial Analysis along Networks: Statistical and Computational Methods. Wiley, Chichester.
5. OpenStreetMap contributors: Planet dump retrieved from https://planet.openstreetmap.org (2017). `https://planet.openstreetmap.org`
6. Chinmoy, S.; Chris, W. & Sarika, K. (2018) Street morphology and severity of road casualties: A 5-year study of Greater London, International Journal of Sustainable Transportation, 12:7, 510-525.
7. Tuba, B.; Esteve, A.; Jonathan, W. (2013), A Smart City Initiative: the Case of Barcelona, Journal of the Knowledge Economy, 4, (2), 135-148.
8. Vito, A.; Berardi, U.; Dangelico, R. (2015) Smart Cities: Definitions, Dimensions, Performance, and Initiatives, Journal of Urban Technology, 22:1, 3-21.