Spatial modelling road accidents in Besançon (France) using log-Gaussian Cox processes

Cécile Spychala [a,1], Camelia Goga [a,2] & Clément Dombry [a,3]

[a] *Laboratoire de Mathématiques de Besançon, 16 route de Gray, 25000, Besançon*
[1] *cecile.spychala@univ-fcomte.fr*
[2] *camelia.goga@univ-fcomte.fr*
[3] *clement.dombry@univ-fcomte.fr*

Résumé. Dans un objectif de prévention et/ou d'anticipation des accidents routiers, la modélisation statistique de la dépendance spatiale et des facteurs de risque potentiels représente un atout majeur. L'intérêt de cette étude se porte sur la localisation géoréférencée des accidents. Nous avons croisé ces événements avec des co-variables caractérisant la zone géographique d'étude (socio-démographiques et infrastructures par exemple). Après une sélection de variables (méthodes de pénalisation, random forest, ...), la survenue des accidents a été modélisée par un processus de Cox log-Gaussien spatial. Les résultats de cette analyse permettent l'identification des principaux facteurs de risques d'accident et l'identification des zones critiques. Les données mises en application sont les accidents routiers s'étant produits entre 2017 et 2019 dans la CAGB (communauté urbaine de Besançon).

Mots-clés. accidents routiers, données spatiales, processus de Cox log-Gaussien

Abstract. In order to prevent and/or forecast road accidents, the statistical modelling of spatial dependence and potential risk factors is a major asset. The focus of this study is on the georeferenced location of accidents. We crossed these events with covariates characterizing the study geographical area such as sociodemographic and infrastructure measures. After a variable selection (penalization methods, random forest, ...), the occurrence of accidents was modelled by using a spatial log-Gaussian Cox process. The results of this analysis enable us to identify principal risk factors of road accidents and critical areas. The data used are road accidents that occurred between 2017 and 2019 in the CAGB (urban community of Besançon).

Keywords. road accidents, spatial data, log-Gaussian Cox process

1 Introduction

Road accidents being one of the main causes of death in the world are therefore the subject of serious concern. Statistical analyses of road accident data are major tools for law enforcement agencies who can act to prevent and predict these events. Depending on the main goal of the analysis, several methods can be undertaken. The focus of this study here is to identify the areas of the CABG (urban community of Besançon) with high occurrence rates of road accidents. Taking into account the spatial characteristics of these accidents therefore seems essential. Road accident data of this study are georeferenced points, called point pattern, and the main motivation is to understand how spatially these points are located. In the field of spatial statistics, point process are stochastic mechanism that generates a countable set of events in the plane (point pattern).

As the locations of road accidents are at the heart of the subject, these events have been modelled by a spatial point process named log-Gaussian Cox process. Moreover in order to identify accident risk factors, additional information was brought as covariates in the model.

2 Material and methods

2.1 Case study

The study is focused on road accidents that occurred between 2017 and 2019 in the CAGB (*Communauté d'Agglomération du Grand Besançon*), headquarters of the region Bourgogne-Franche-Comté. This urban community is composed of 68 cities including Besançon which is the central municipality. These road accidents are georeferenced and represent hence our point pattern to be analysed.

2.2 Log Gaussian Cox Processes

Data in the form of a set of points $\mathbf{x} = \{x_1, \ldots, x_n\}$ of $\mathbb{R}^2$, irregularly distributed over a study area, is called a spatial point pattern and referred to the locations as events (Diggle, 2013). This set of points is considered to be the results from the realization of an underlying stochastic mechanism called point process. As the analysis is not about the points themselves, but about the way the points were generated, fit a spatial point process is generally the approach undertaken to model such a situation. Dealing with point processes lies on the study of the intensity of the point process and interaction between points. The analysis of the intensity is usually the most important, it is the average number of random points per unit area. Then, interaction between points is investigate whether the points appear to have been placed independently of each other or whether they exhibit some kind of interpoint dependence.

We suggest in this paper to model our point pattern by a Cox process and more particularly by a Log-Gaussian Cox Process (LGCP). An LGCP is a Cox process whose log-intensity is modelled by a latent Gaussian process. For more details see for example Moller et al. (1998), Diggle et al. (2013) or Taylor et al. (2015). In practice, the adjustment is carried out on a calculation grid (considered regular). The spatial LGCP is noted as follows:

$$X(u) \sim Poisson(\Lambda(u))$$
$$\Lambda(u) = C_A \exp(Z(u)\beta + Y(u))$$

where $X(u)$ denotes the number of events in the computational grid cell containing u, C_A is cell area, $Z(u)$ is a vector of covariate values, with associated effects β and $Y = \{Y(u) : u \in \mathbb{R}^2\}$ is a Gaussian process (which has variance σ^2 and exponential correlation function $r(t) = \exp(-t/\phi)$).

Bayesian inference for the parameters Y, β, σ and ϕ is performed using Markov chain Monte Carlo methods. The reader may find a comprehensive description in Robert (1996), Robert and Casella (2004) or Taylor et al. (2015).

2.3 Covariate description

As seen before, covariate can be included in the definition of the intensity of the LGCP. Hence, information from data sources such as INSEE (*Institut National de la Statistique et des Etudes Economiques*), data.gouv or Open Street Map Data, has been collected. These measures have been merged into one spatial dataset (interpolation methods were used) and onto a grid, composed of regular cells.

The covariates can be classified in three groups: sociodemographic, road infrastructures and global infrastructures. As sociodemographic covariate we have the population, that is, the number of individuals per cell. As road infrastructure covariates we have, per cell, the length of municipal roads and national roads (departmental, national or highway), the number of traffic lights, the number of stop sign, the number of passage way, the number of give way and the number of radars. Finally, as global infrastructure covariates we have, per cell, the number of shops, the number of food shops (such as supermarkets or grocery stores for example), the number of restaurants, the number of leisure activities (such as cinemas or tennis courts for example), the number of schools, the number of colleges, the number of gas stations and the number of stations (such as train stations or taxis for example). The total number of covariate is sixteen.

Some of these measures were chosen as they showed relevant characteristics in previous accident studies (Ramírez and Valencia, 2021; Spychala et al., 2021) and others were taken as they could probably show significant results.

2.4 Variable Selection

As the number of covariates is quite important and the point process model chosen to be fitted is computational challenging, a variable selection is intended. To do this, supervised learning methods are used in order to analyse dependencies between the response variable, which gives the number of accidents, and covariates mentioned before. Several methods were used such as Random Forest, Elastic Net or Poisson models.

3 Results

After fitting a LGCP, we can interpret the results and thus meet the objectives set. To quantify the risks of potential factors, focus is on the covariate effects. Estimations of the parameters β make it possible to conclude on the significant or non significant effect of the covariate and hence, identify the most dangerous factors.

Finally, in order to identify the critical geographical areas of low or high incidence, the following quantity $\mathbb{P}(\exp(Y) > k|X)$ is plotted on the study window as in Taylor et al. (2015). The map obtained represents the probability that the average number of accidents per cell exceeds the threshold k and enable to visualize the critical zones.

References

Diggle, P. (2013). *Statistical Analysis of Spatial and Spatio-Temporal Point Patterns.* Chapman & Hall/CRC, third edition.

Diggle, P., Moraga, P., Rowlingson, B., and Taylor, B. (2013). Spatial and spatio-temporal log-gaussian cox processes: Extending the geostatistical paradigm. *Statistical Science*, 28(4):105848.

Moller, J., Syversveen, A., and Waagepetersen, R. (1998). Log gaussian cox processes. *Scandinavian Journal of Statistics*, 25(3):451–482.

Ramírez, A. and Valencia, C. (2021). Spatiotemporal correlation study of traffic accidents with fatalities and injuries in bogota (colombia). *Accident Analysis & Prevention*, 149:105848.

Robert, C. (1996). *Méthodes de Monte Carlo par chaînes de Markov.* Economica.

Robert, C. and Casella, G. (2004). *Monte Carlo Statistical Methods.* Springer.

Spychala, C., Armand, J., Dombry, C., and Goga, C. (2021). Multivariate statistical analysis for exploring road crash-related factors in the franche-comté region of france. *Communications in Statistics: Case Studies, Data Analysis and Applications*, 7(3):442–474.

Taylor, B. M., Davies, T. M., Rowlingson, B. S., and Diggle, P. J. (2015). Bayesian inference and data augmentation schemes for spatial, spatiotemporal and multivariate log-gaussian cox processes in r. *Journal of Statistical Software*, 63(7).