Women in Big Data Workshop
Renaissance Zurich Tower Hotel, 14-15 June 2018

Big Data in Environmental and Health Sciences

Modelling Challenges for Large Spatio-Temporal Data Structures

Raquel Menezes
Department of Mathematics and Applications and CBMA, University of Minho, Portugal
CEAUL, Center of Statistics and Applications, University of Lisbon
rmenezes@math.uminho.pt

Keywords: space-time data, high temporal resolution, stochastic processes.

Introduction

Spatial and temporal data are quite common in environmental science, and occasionally in health studies. Different inference methodologies are in use for spatio-temporal data, and the complexity and nature of each problem dictates the nature of the methodology to be adopted. We present two motivating examples: the former related to air quality monitoring (using geostatistical inference methods); and the other related to disease modelling, when data are collected along time and aggregated in space per administrative areas (involving generalized linear mixed effects models).

The high dimensionality of spatio-temporal data often leads to inference methods offering model simplifications, approximations of estimators and likelihood functions, as well as efficient computational and algorithmic techniques. However, an understanding of the physical properties of the spatio-temporal process in question frequently enables many simplifying assumptions to be made which allows for convenient mathematical properties (e.g. stationarity, isotropy and various forms of conditional independence).

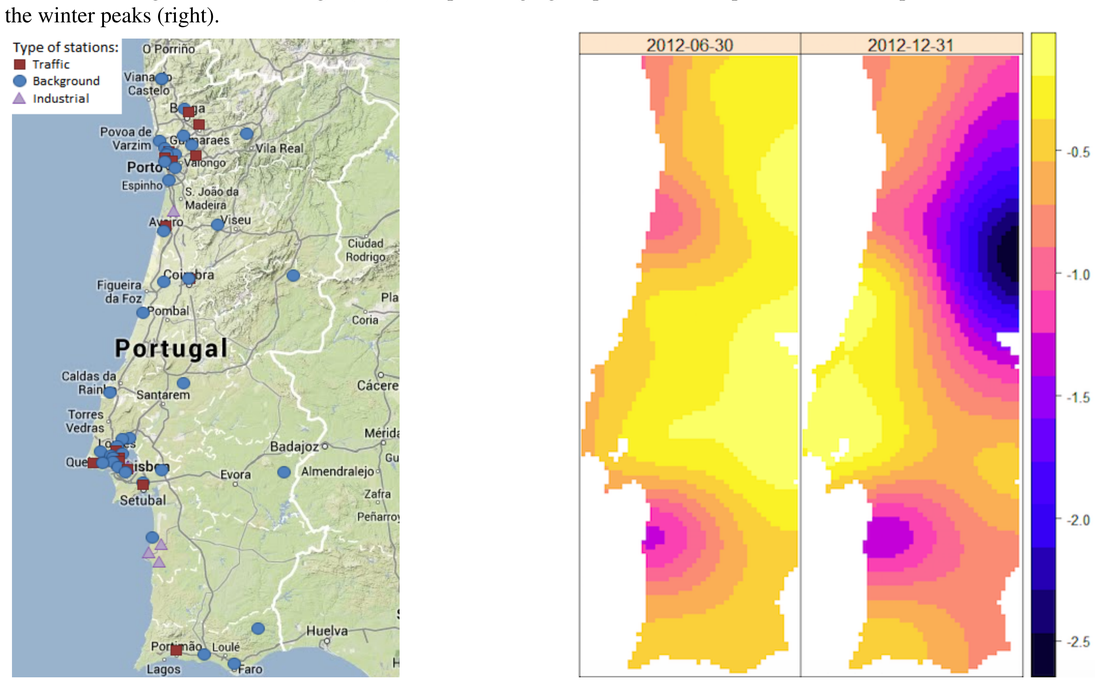
Specific concerns in modelling space-time 'big data":

- To **characterize the spatial and temporal dynamics** of spatio-temporal data sets (e.g. through the estimation of a space-time variogram in Figure 2), often characterized by **high resolution in temporal dimension**. The latter is becoming the norm rather than the exception in many application areas, namely in environmental modelling.
- In the particular case of modelling air pollution data, to **consider multiple recurring patterns in time** imposed by social habits, anthropogenic activities and meteorological conditions (e.g. through an harmonic regression).
- To **do model assessment** and obtain accuracy measures of parameter estimates (e.g through Bootstrap methods, as given in Table 1), which might strongly depend on the representativeness of sampled data among all "big data".
- To deal with **preferential sampling issues**, e.g. *smart cities* (data collected through sensors that are located in points where levels of pollution are expected to be higher). This originates a stochastic dependence between the point process (given by sensor's location) and the measurement process, which must be taken into account in inference methods (Diggle, Menezes and Su, 2010).

Motivating Example I: *Air Pollution Monitoring in Portugal*

The nitrogen dioxide is a primary pollutant, regarded for the estimation of the air quality index, whose excessive presence may cause significant environmental and health problems. The European Environment Agency, EEA (2015) considers air pollution the single largest environmental health risk in Europe. Thus the **need for accurate assessment of air pollution arises not only to investigate the linkage between ambient exposure and health effects, but also with regard to compliance with legislated regulatory standards to control levels of environmental exposure**. This requires statistical models aimed at characterizing and predicting air quality events and assessing policies over specified areas.

Figure 1: Monitoring network in Portugal (left) and Space kriging-maps for two time points chosen as representative of the summer and the winter peaks (right).



Main Goals:

- Analyze **hourly NO2 concentrations**, collected in **49 stations stations** located over Portugal (mainland) from October 1st to December 31st in 2014, in a total of 108192 observations;
- Characterize the evolution of NO2 levels in Portugal, by using geostatistical approaches that deal with both the space and time coordinates;
- Take into account that environmental data often incorporate distinct recurring patterns in time, resulting from social habits (e.g. traffic rush hours) and from the influence of meteorological variables.
- Be able to model multiple seasonalities and **capture intra- and inter-day periodicities**, typically underlying monitoring data collected in populated areas;
- Consider a **block bootstrap procedure to correctly assess uncertainty in parameters estimates** and produces reliable confidence regions for the space-time phenomenon under study.

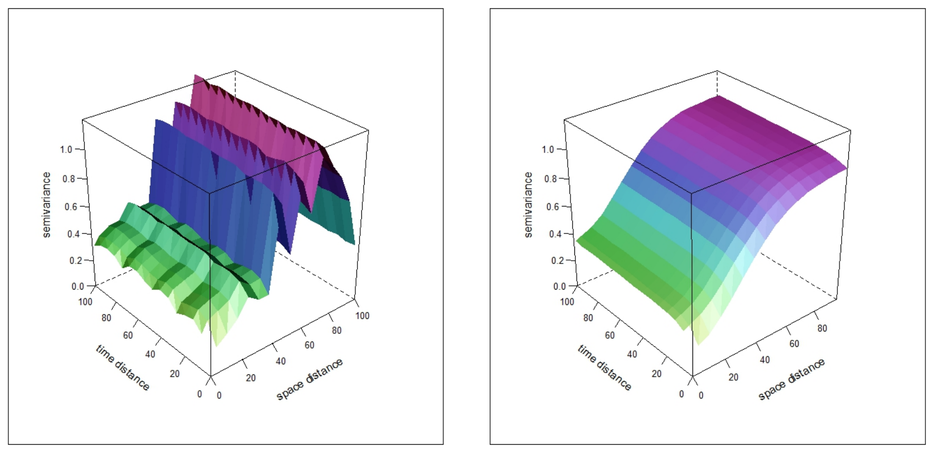


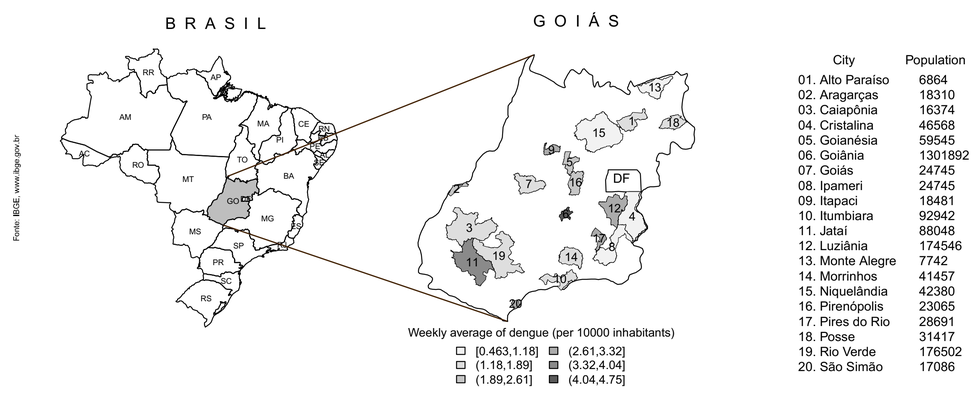
Figure 2: Plots of the experimental estimator (left) and the fitted model (right) for the space-time variogram.

Table 1: Parameters estimates, and corresponding bootstrap std.errors (replicates with 5 weeks, obtained by sliding 8 hours), for the spatial, temporal and spatio-temporal variograms

Variogram	Model	τ^2	σ^2	ϕ	α
Spatial	Gaussian	0.015 (0.022)	0.66 (0.13)	**40km** (1.373)	
Temporal	Exponential	0.010 (0.018)	0.07 (0.02)	**100hours** (0.003)	
Joint	Gaussian	0.172 (0.018)	0.13 (0.03)	70km (0.024)	13.0 (0.1)

Motivating Example II: *Dengue Counts in the State of Goiás*

Dengue fever is a rapidly spread viral disease. In Brazil, all suspected cases must be notified to the epidemiological surveillance of each municipality by the local health units, via the Notifiable Diseases Information System. There is **a worrying increase of dengue cases mostly in young people and children**. The most populous state of center-west region of Brazil is Goiás (Figure 3), with around 6 million inhabitants, distributed among a total of 246 cities, being Goiânia the capital with 1.5 million of people.



City	Population
01. Alto Paraíso	6864
02. Aragarças	18310
03. Caiapônia	16374
04. Cristalina	46568
05. Goianésia	59545
06. Goiânia	1301892
07. Goiás	24745
08. Ipameri	24745
09. Itapaci	18481
10. Itumbiara	92942
11. Jataí	88048
12. Luziânia	174546
13. Monte Alegre	7742
14. Morrinhos	41457
15. Niquelândia	42380
16. Pirenópolis	23065
17. Pires do Rio	28691
18. Posse	31417
19. Rio Verde	176502
20. São Simão	17086

Figure 3: State of Goiás, Brazil. The 20 cities considered in our study.

Main Goals:

- Study the relation between the number of dengues notifications in the state of Goiás and its climate variations. We have available **weekly data**, collected across **20 cities** well-representative of Goiás (Figure 3), starting from January 2008;
- Analyse the influence of climate conditions, which suggests the presence of temporal and spatial correlations among dengue data. So, we intend to investigate generalised linear **mixed models**, capable of **incorporating temporal and spatial random effects** with distinct correlation structures;
- Investigate alternative covariates, including cumulative **rainfall**, **minimum and maximum temperatures**, **relative humidity** and **wind speed** (Table 2), at different time-lags;
- Model dengue counts, considering a **Poisson** distribution, and a **negative binomial** distribution due to data overdispersion (Table 2), with a population offset.

Table 2: Summary statistics for the responde (Y) and possible explanatory (X) variables.

Variables	Minimum	Mean	Maximum	Std.Dev.	Relative Std.Dev.
dengue counts - Y	0	45.4	5145	248.4	5.5
precipitation – $prec$ (mm)	0	25.1	251.8	36.3	1.5
min.temperatures – $tmin$ (^{o}C)	1.6	16.7	26.3	3.3	0.19
max.temperatures – $tmax$ (^{o}C)	23.3	32.4	41.7	2.6	0.08
rel.humidity – $humid$ (%)	18.1	66.2	91.0	14.8	0.22
wind speed – $wind$ (m/s)	0	1.7	4.8	0.7	0.42

References

- European Environment Agency, EEA (2015). Air quality in europe - 2015 report. URL: http://www.eea.europa.eu/publications/ air-quality-in-europe-2015.
- Diggle P., Menezes R. and Ting-Li Su (2010). Geostatistical Inference under Preferential Sampling. Journal of Royal Statistics Society (with discussion), series C, vol 59, part 2, 191-232.
- Menezes R., Piairo H., Garcia-Soidán P. and Sousa I. (2016). Spatial–temporal modellization of the NO2 concentration data through geostatistical tools. Journal of Statistical Methods & Applications, 25: 107–124.
- Monteiro A., Menezes R. and Silva M.E. (2017). Modelling spatio-temporal data with multiple seasonalities: the NO2 Portuguese case. Spatial Statistics journal, Volume 22, Part 2: 371–387.