European Journal of Operational Research 000 (2016) 1–10

Contents lists available at ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Invited Review

Statistical analysis of complex and spatially dependent data: A review of Object Oriented Spatial Statistics

Alessandra Menafoglio, Piercesare Secchi*

MOX-Department of Mathematics, Politecnico di Milano, Piazza Leonardo da Vinci 32, Italy

A R T I C L E I N F O

Article history:
Received 26 January 2016
Accepted 29 September 2016
Available online xxx

Keywords:
Object oriented data analysis
Kriging for object data
Bagging Voronoi algorithm
Spatial regression models with differential regularization

A B S T R A C T

We review recent advances in Object Oriented Spatial Statistics, a system of ideas, algorithms and methods that allows the analysis of high dimensional and complex data when their spatial dependence is an important issue. At the intersection of different disciplines – including mathematics, statistics, computer science and engineering – Object Oriented Spatial Statistics provides the right perspective to address key problems in varied contexts, from Earth and life sciences to urban planning. We illustrate a few paradigmatic methods applied to problems of prediction, classification and smoothing, giving emphasis to the key ideas Object Oriented Spatial Statistics relies upon.

1. Introduction: a few paradigmatic problems

Object Oriented Spatial Statistics (O2S2) is a new branch of statistics, that aims to provide a unifying viewpoint to address a variety of application-oriented challenges of modern spatial statistics. In O2S2 the “atom” of the analysis is typically a complex data point (such as a curve or a surface), spatially distributed. The object oriented viewpoint to the statistical treatment of these types of data consists in considering as building block of the analysis the whole data point, regardless of its complexity. The datum is interpreted as a point within an appropriate space of objects (called *feature* space), which should properly represent all the salient data features through its geometry. Implicit or explicit formalization of the notion that nearby data provide similar information then enables one to develop statistical procedures that take full advantage of the information content embedded within the data for the purpose of modeling, classifying or predicting spatial data. As such O2S2 is part of Object Oriented Data Analysis, the seminal name chosen by Wang and Marron (2007) to baptize a system of ideas according to which new statistical tools are developed urged by the analysis of populations of complex objects. Object Oriented Data Analysis requires a strong interplay of statistics with other scientific disciplines, including maths (analysis, scientific computing, geometry and algebra, operational research), engineering, scientific communication design, computer graphics, computer science, and information technology (see Marron & Alonso, 2014 for a recent review on Object Oriented Data Analysis).

Nowadays, the analysis of complex spatial objects plays a key role in a variety of data-driven engineering and geoscience applications. Even though most methods in OODA rely on the key assumption that the observations are independent and identically distributed, collections of spatially distributed data are increasingly available in field studies. O2S2 then meets the need of analysing populations of spatially dependent object data. Here, we propose a first paradigmatic example in O2S2, to engage the reader in the object oriented viewpoint to spatial statistics. This example represents the *fil rouge* of the present review, since it fully expresses our viewpoint on the way the research on O2S2 has been motivated, interpreted and developed in recent years.

Example 1. Fig. 1 displays a representation of a heterogenous aquifer system at the Lauswiesen site, in the Neckar river valley, near the city of Tübingen (Germany). Data at the site consist of particle-size densities (PSDs) measured at 406 locations along 12 boreholes. PSDs describe the local distributions of grain sizes within the aquifer system. From the mathematical viewpoint, a particle-size density is a probability density function, associated with the distribution of particle sizes within a given soil sample. As such, available data consist of a set of constrained curves, spatially distributed. The statistical characterization of PSDs plays a key role in the classification of soil types, for inferring hydraulic parameters (e.g., porosity and hydraulic conductivity), and reconstructing the internal architecture of the groundwater system. In this vein, the analysis of PSDs may be concerned with, e.g., (a) the classification of PSDs to identify geomaterials at the site, and (b)

* Corresponding author.
E-mail addresses: piercesare.secchi@polimi.it, piercesare.secchi@gmail.com (P. Secchi).

http://dx.doi.org/10.1016/j.ejor.2016.09.061

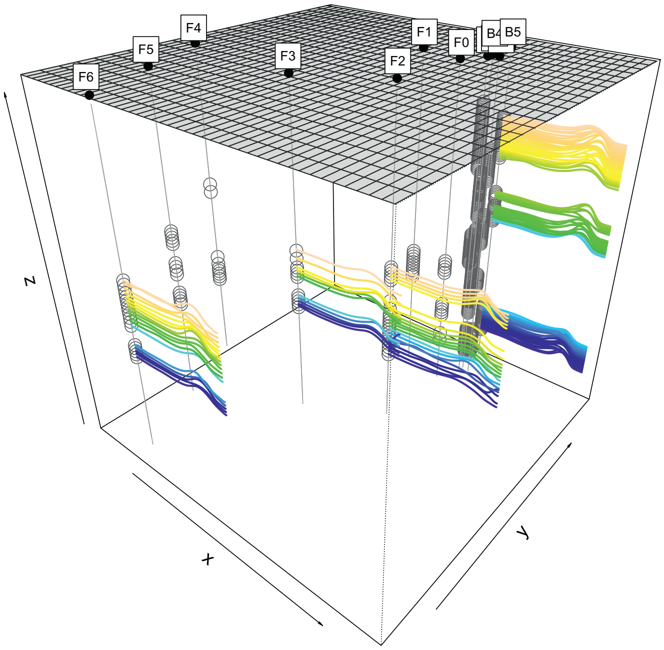

Fig. 1. Three-dimensional representation of particle-size densities at the Lauswiesen site. Grey points represent measurement locations, colored curves represent a subset of the dataset of PSDs. Colors indicate the ordering of the curves along the borehole.

the spatial prediction and stochastic simulation of the PSDs at unsampled locations of the system (Menafoglio, Guadagnini, & Secchi, 2014; 2016b; Menafoglio, Secchi, & Guadagnini, 2016c).

In Example 1, each data point is a complex object: here, critical elements of complexity are the data dimensionality and constraints. Indeed, PSDs are curves, thus infinite-dimensional data: they need an infinity of point evaluations to be fully characterized. Furthermore, PSDs are distributional data, hence constrained to be positive and integrate to unity. This cannot be just neglected: capturing the complexity of the objects is key to precisely model and express the information content embedded within data, which will then be used for exploration, inference and prediction with the final aim of accruing knowledge.

Nevertheless, the spatial domain of Example 1 is relatively simple: the aquifer can be thought of as a three-dimensional Euclidean parallelepiped, of moderate size. In this setting, one can imagine to extend to object data the ensemble of model-based methodologies that are widely-employed in classical geostatistics to deal with the spatial dependence (e.g., Kriging, Cressie, 1993).

In fact, one can readily envision situations of complex objects observed over a more complex spatial domain. Here, complexity may be associated with the size or the texture of the area of interest. For instance, data observed over very large regions pose challenges related to the practical impossibility of employing approaches based on global models for spatial dependence. Similarly, local models need to be employed to account for a very strong, irregular or sudden variability induced by, e.g., the texture of the domain. An instance of this is found in the study of vehicular traffic in urban areas: here, the road system induces a very fine texture. The spatial variability can be then expressed through local models of diffusion, governed by spatially varying tensors (see, e.g., Della Rossa, D'Angelo, & Quarteroni, 2010). In some cases, the process could develop over a complex and evolving system such as a spatial network (Reggiani, Nijkamp, & Sabella, 2001). In all these cases, local models or fully non-parametric approaches appear to be more appropriate than those inspired by classical geostatistics.

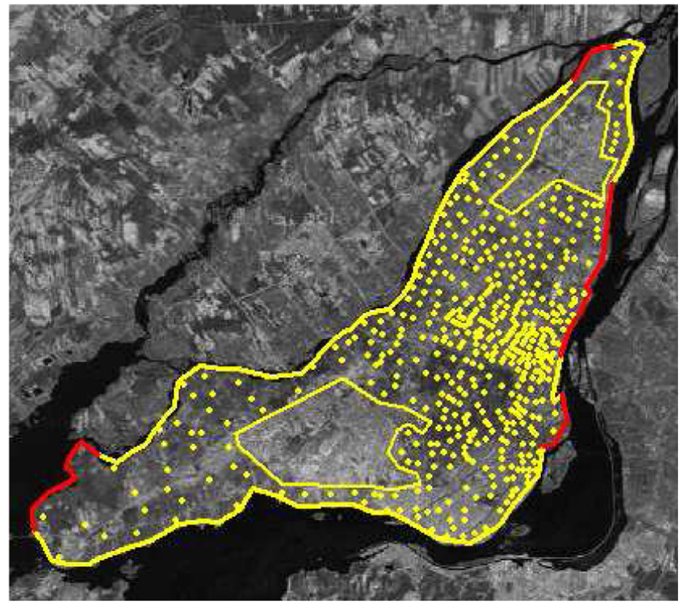

Fig. 2. Island of Montréal and sampling sites for the census data of Example 2. Red lines along the boundary of the region denote places where population density needs to be zero. (Figure modified from Sangalli et al., 2013, courtesy of Laura Sangalli).

An example in this class, based on a computationally intensive, explorative approach, will be illustrated in Section 5.

A different kind of complexity is that pertaining to domains with holes or highly irregular borders determined by, e.g., geographical constraints. For instance, in the setting of Example 1, one may think at the Neckar river as a peculiar boundary of the aquifer domain, and particular geological formations within the aquifer as holes in the domain. Part of the modern literature on spatial statistics – which in our view partially shares the approach of O2S2 – has focused on developing methods for dealing with these issues. To get closer to the topic, we now present an example where relatively simple data over a complex domain are concerned.

Example 2. Sangalli, Ramsay, and Ramsay (2013) analyze census data in the Island of Montréal. The data consists of population densities (scalar observations) collected at a limited number of spatial locations over the region, depicted in Fig. 2. The Island of Montréal is located between two rivers – Saint Lawrence and Rivière de Prairis – which form natural geographical constraints. Additionally, the harbour and the public parks are areas where people cannot live. Hence the spatial domain is here defined by irregular boundaries and it also has holes. In this case, the goal of the analysis is to reconstruct the target surface (density over the spatial domain) by (a) including a set of covariates and possible prior knowledge on the phenomenon (e.g., a set of known differential equations governing the phenomenon) and (b) properly accounting for the topology of the spatial domain and the shape of its boundary.

Note that highly textured spatial regions, characterized by irregular boundaries and holes, are inevitably associated with boundary conditions: for instance, the population density of Example 2 needs to be zero in correspondence of the uninhabited areas (e.g., the harbour or some parts of the river banks marked with red in Fig. 2), as noted by Sangalli et al. (2013). Furthermore, observations of the phenomenon might be also repeated in time, or other kinds of more complex objects might be observed at the measurement locations. This projects us to the very cutting-edge of O2S2, where methods which jointly treat both the object and the domain complexities are yet to be developed.

Please cite this article as: A. Menafoglio, P. Secchi, Statistical analysis of complex and spatially dependent data: A review of Object Oriented Spatial Statistics, European Journal of Operational Research (2016), http://dx.doi.org/10.1016/j.ejor.2016.09.061