A novel method for spatial smoothing

Laura M. Sangalli and James O. Ramsay

Abstract A novel method for surface estimation and spatial smoothing is presented, that exploits advanced numerical techniques. The proposed model handles accurately data distributed over irregularly shaped domains, characterized by complex boundaries and holes, and also has the capacity to comply with different conditions at the boundaries of the domain. Thanks to the flexible framework considered, the model can be generalized to include a priori information about the spatial structure of the phenomenon, and to deal with data distributed over non-planar domains.

Key words: Functional Data Analysis, penalized smoothing, finite elements.

1 Introduction

We briefly describe the Spatial Spline Regression (SSR) models proposed in [4] for the estimation of surfaces and spatial fields. These models, developed following an approach typical of Functional Data Analysis, make also use of advanced numerical techniques, specifically of finite elements, that provide a computationally highly efficient system of local basis for piecewise polynomial surfaces. SSR are semi-parametric models, or generalized additive models, and allow for inclusion of spatially distributed covariate information. How shown in [3] and [4], SSR outperform classical methods such us thin-plate splines and filtered kriging, as well as more recent techniques for spatial data analysis. In particular, SSR models are able to accurately handle data distributed over irregularly shaped domains, featur-

Laura M. Sangalli
MOX - Dipartimento di Matematica, Politecnico di Milano, Piazza L. da Vinci 32, 20133 Milano; e-mail: laura.sangalli@polimi.it

James O. Ramsay
Dep. of Psychology, McGill University, Montréal, Canada; e-mail: ramsay@psych.mcgill.ca

ing complex boundaries, strong concavities and even interior holes; moreover, they can comply with specific conditions at the boundaries of the domain.

Applied illustrative problem: Island of Montréal census data. To illustrate the issue of spatial smoothing over irregularly shaped domains and with boundary conditions, consider the problem of estimating population density over the Island of Montréal (Québec, Canada), starting from census data (1996 Canadian census). Figure 1, left panel, displays census tract locations over the Island of Montréal; population density and other census information are available at each census tract, together with a binary covariate indicating whether a tract is predominantly residential or industrial/commercial. The figure highlights two parts of the island without data: the airport and rail yards in the south and an industrial park with an oil refinery tank farm in the north-east; these two areas are not part of the domain of interest when studying population density, since people cannot live there. Notice that census quantities can vary sharply across these uninhabited parts of the city; for instance, in the south of the industrial park there is a densely populated area with medium-low income, but north-east of it there is a wealthy neighborhood with low population density. Hence, whilst it seems reasonable to assume that population density features a smooth spatial variation over the inhabited parts of the island, there is no reason to assume smooth spatial variation across uninhabited areas. The figure also shows the island coasts as boundaries of the domain of interest; those parts of the boundary that are highlighted in red correspond respectively to the harbor, in the east shore, and to two public parks, in the south-west and north-west shore; and no people live by the river banks in these boundary intervals. We thus want to study population density, taking into account covariate information, being careful not to artificially link data across areas where people cannot live, and also efficiently including prior information concerning those stretches of coast where the population density should drop to zero.

2 Spatial Spline Regression models

Consider a set of n points $\{\mathbf{p}_i = (x_i, y_i);\, i = 1\ldots, n\}$ on a polygonal domain $\Omega \subset R^2$, where Ω can be quite complex (for instance, it can have strong concavities and interior holes). Let z_i be the value of a real valued variable of interest, observed at point $\mathbf{p}_i$, and let $\mathbf{w}_i = (w_{i1}, \ldots, w_{iq})^t$ be a q-vector of covariates associated to observation z_i at $\mathbf{p}_i$. Assume the semi-parametric model

$$z_i = \mathbf{w}_i^t \beta + f(\mathbf{p}_i) + \varepsilon_i \qquad\qquad i = 1, \ldots, n$$

where ε_i, $i = 1, \ldots, n$, are i.i.d. errors with zero mean and constant variance, and where $\beta \in R^q$ is a vector of coefficients and f is a twice differentiable real-valued function on Ω. In [3] and [4] we propose to estimate β and f by minimizing the following penalized sum-of-square-error functional

$$J_\lambda(\beta, f) = \sum_{i=1}^{n} \left(z_i - \mathbf{w}_i^t \beta - f(\mathbf{p}_i)\right)^2 + \lambda \int_\Omega (\Delta f)^2 d\Omega \tag{1}$$

where the roughness term is the integral, over the domain of interest Ω, of the square of the Laplacian of f, $\Delta f = \partial^2 f/\partial x^2 + \partial^2 f/\partial y^2$. Recall that the Laplacian is a measure of the local curvature of f that is invariant with respect to Euclidean transformations of spatial coordinates, therefore ensuring that the concept of smoothness does not depend on the orientation of the coordinate system. In [4] we hence show how the estimation problem (1) can be efficiently solved resorting to finite elements, that provide a system of local bases for piece-wise polynomial surfaces, associated to a domain triangulation. Finite elements naturally enforce the computation of distances within the domain, and make use of local coordinates, thus providing accurate surface estimation also in the case of irregularly spaced domains. The obtained estimators of β and f turn out to be linear in the observed data values, so that classical inferential tools may be readily derived.

Application to Island of Montréal census data. Figure 1, center panel, displays a triangulation of the domain of interest for Montréal census data application. The right panel of the same figure shows the estimated spatial structure of population density, measured as 1000 inhabitants per km^2, using as covariate the binary variable that indicates whether a tract is predominantly residential (1) or commercial/industrial (0). Notice that the estimate complies with the required boundary conditions, dropping to zero along the stretches of coast corresponding to the harbor and public parks (highlighted in red in the left and center panels). Also, the estimate has not artificially linked data points on either side of the uninhabited parts; see for instance the densely populated areas just in the south and in the west of the industrial park, with respect to the low population density neighborhood north-east of it. The β coefficient that corresponds to the binary covariate is estimated to be 1.300; this means that census tracts that are predominantly residential are in average ex-

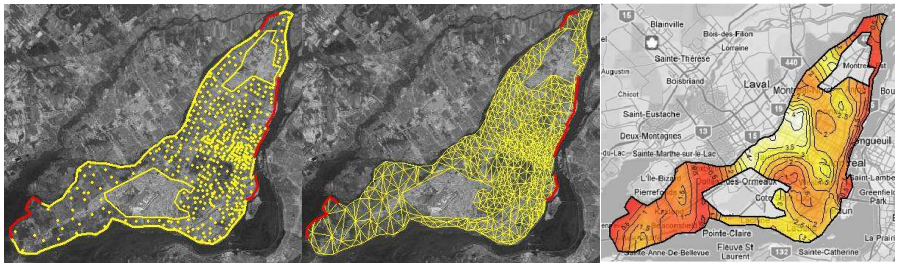

Fig. 1 Left: Island of Montréal census data. The yellow dots are the centroids of census enumeration areas, for which population density and other census information are known. The two parts of the island where there are no data, encircled by yellow lines, are areas where people cannot live; the hole in the south-wester part of the island is the Dorval airport and the one in the north-eastern end represents an industrial park containing oil refineries and a water purification plant. The island coast is also evidenced with a yellow line; the stretches highlighted in red are the harbour, in the east coast, and two parks, in the west coast. Center: Triangulation of the Island of Montréal. Right: SSR estimate of spatial structure for population density over the Island of Montréal.

pected to have 1300 more inhabitants per km^2, with respect to those predominantly commercial; the approximate 95% confidence interval is given by $[0.755; 1.845]$.

Notice that classical techniques for spatial data analysis are not able to efficiently deal with this problem. These techniques in fact smooth across internal and external boundaries, leading for instance to inaccurate estimates of population density around the uninhabited areas; moreover they cannot comply with specific boundary conditions. On the contrary SSR models have accurately handled the problem.

3 Model extensions

SSR models may be generalized to roughness terms penalizing more complex partial differential operators, instead of the simple Laplacian; this is particularly interesting for applications where some a priori knowledge of the problem (physical, chemical, mechanical, morphological) suggests the choice of a partial differential operator modeling to some extent the phenomenon under study. [1] investigates this research direction, and shows an application to the estimation of the blood-flow velocity field in a section of a carotid artery, using data provided by eco-color dopplers.

Moreover, the method can also be extended to deal with data distributed over non-planar domains. This research front is explored in [2], and the proposed model extension is applied to the study of hemodynamical data, such as wall shear stress and pressure, observed over the wall of a carotid artery (data coming from computational fluid dynamics and image reconstructions of 3-dimensional angiographies).

The code implementing SSR models has been fully integrated with the `fda` sotware available in R and Matlab, and will be shortly released within these packages.

Acknowledgements Funding by MIUR *FIRB Futuro in Ricerca* research project "Advanced statistical and numerical methods for the analysis of high dimensional functional data in life sciences and engineering", and by the program Dote Ricercatore Politecnico di Milano - Regione Lombardia, research project "Functional data analysis for life sciences".

References

1. Azzimonti, L., Sangalli, L. M., Secchi, P., Domanin, M. (2011), "Surface estimation via spatial spline models with PDE penalization," Proceedings of SCo 2011, available at http://sco2011.stat.unipd.it/index.php/sco2011/SCo2011.
2. Ettinger, B., Perotto, S., Sangalli, L. M. (2012), "Spatial smoothing over non-planar domains," 46th Scientific Meeting of the Italian Statistical Society, submitted.
3. Ramsay, J. O., Ramsay, T. O., Sangalli, L. M. (2011), "Spatial Functional Data Analysis," in Recent Advances in Functional Data Analysis and Related Topics, Contributions to Statistics, Physica-Vergal Springer, pp. 269–276.
4. Sangalli, L. M., Ramsay, J. O., Ramsay, T. O. (2012), "Spatial spline regression models," Tech. Rep. MOX N. 08/2012, Dipartimento di Matematica, Politecnico di Milano, available at http://mox.polimi.it/it/progetti/pubblicazioni.