# GENERALIZED MODELS FOR SPATIAL REGRESSION WITH DIFFERENTIAL PENALIZATION

Matthieu Wilhelm[1,2] and Laura M. Sangalli[3]

[1] Section de Mathématiques
Ecole Polytechnique Fédérale de Lausanne
Station 8, Bâtiment MA, Lausanne 1015, Switzerland
(e-mail: matthieu.wilhelm@epfl.ch)

[2] Institut de Statistique
Université de Neuchtel
Pierre â Mazel 7, 2000 Neuchâtel, Switzerland
(e-mail: matthieu.wilhelm@unine.ch)

[3] MOX - Dipartimento di Matematica
Politecnico di Milano
P.za Leonardo da Vinci 32, 20133 Milano, Italy
(e-mail: laura.sangalli@polimi.it)

ABSTRACT. We introduce a novel method for the analysis of spatially distributed data from an exponential family distribution, able to efficiently deal with data occurring over irregularly shaped domains. The proposed generalized additive framework can handle all distributions within the exponential family, including binomial, Poisson and gamma outcomes, hence leading to a very broad applicability of the model. Specifically, we maximize a penalized log-likelihood function where the roughness penalty term involves a suitable differential operator of the spatial field over the domain of interest. Space-varying covariate information can also be included in the model in a semi-parametric setting. The proposed model exploits advanced scientific computing techniques and specifically makes use of the Finite Element Method, that provide a basis for piecewise polynomial surfaces.

## 1 THE MODEL AND ESTIMATION PROBLEM

We develop a model that deals with spatially distributed realizations having a distribution within the exponential family. Consider a bounded regular domain $\Omega \in R^2$ and spatial locations $\mathbf{p}_1, \ldots, \mathbf{p}_n$ scattered over $\Omega$. Let $y_i$ be the variable of interest observed at $\mathbf{p}_i$ with an associated $q$-vector of covariates $\mathbf{x}_i^t$. Assume $y_1, \ldots, y_n$ have a distribution within the exponential family with canonical parameter $\theta_i = g(\mu_i)$, where $\mu_i = \mathbf{E}[y_i]$ and $g$ is the canonical link function associated to the distribution of interest (hence, the canonical and natural parameter coincide in this case). Assume the following semiparametric model:

$$\theta_i = g(\mu_i) = \mathbf{x}_i^t \beta + f(\mathbf{p}_i)$$

where $\beta \in R^q$ contains regression coefficients and the real valued smooth function $f$, defined over the domain $\Omega$, accounts for the spatial structure of the phenomenon.

We propose to estimate the regression coefficients $\beta$ and the spatial field $f$ by maximizing the following penalized log-likelihood function (see Wilhelm (2013)):

$$\mathcal{L}_p(\beta, f) = \mathcal{L}(\beta, f) - \frac{1}{2}\,\lambda \int_\Omega (\Delta f)^2 \tag{1}$$

where $\mathcal{L}$ denotes the log-likelihood of the considered distribution. Since the Laplacian of $f$, $\Delta f$, is a measure of local curvature, the second term in (1) penalizes the roughness of the estimated spatial field. In the special case where the considered distribution is the Gaussian, this estimation problem is equivalent to penalized least square error problem considered in Sangalli *et al.* (2013). For distributions other than normal, in order to maximize the penalized log-likelihood (1), we resort to the Penalized Iterative Reweighed Least Squares (PIRLS) algorithm (see, e.g., Wood, 2006). The choice of the smoothing parameter can be performed using the GCV criterion (see, e.g., Wahba, 1990).

Likewise in Ramsay (2002) and Sangalli *et al.* (2013), the function $f$ is approximated using a basis expansion provided by finite elements. This makes the model computationally highly efficient and allows to comply with complex domains and prescribed boundary conditions on $f$. The good performances of the proposed model are illustrated via simulation studies and application to real data.

**Acknowledgements.** L. Sangalli acknowledges funding by MIUR Ministero dell'Istruzione dell'Università e della Ricerca, *FIRB Futuro in Ricerca* research project "Advanced statistical and numerical methods for the analysis of high dimensional functional data in life sciences and engineering" (see http://mox.polimi.it/users/sangalli/firbSNAPLE.html), and by the program Dote Ricercatore Politecnico di Milano - Regione Lombardia, research project "Functional data analysis for life sciences".

## REFERENCES

RAMSAY, T.O. (2002): Spline smoothing over difficult regions. *Journal of the Royal Statistical Society Series B*, *64*, 307–319.

RAMSAY, J.O., SILVERMAN, B.W. (2005): *Functional data analysis*. Second edition. Springer Series in Statistics. Springer, New York.

SANGALLI, L.M, RAMSAY, J.O, RAMSAY, T.O. (2013): Spatial spline regression models. *Journal of the Royal Statistical Society Series B*, *75*, 4, 1–23.

WAHBA, G. (1990): *Splines models for observational data*. SIAM-SIMS Conference Series. Society for Industrial and Applied Mathematics, Philadelphia.

WILHELM, M. (2013): Generalized Spatial Regression with Differential Penalization. *Master thesis, EPFL*.

WOOD, S.N. (2006): *Generalized additive models: an introduction with application in R*. Springer Series in Statistics. Springer, New York.

WOOD, S.N.,BRAVINGTON, M. V., HEDLEY, S.L. (2008): Soap film smoothing. *Journal of the Royal Statistical Society Series B*, *70*, 1931–955.