Analysis of Telecom Italia Mobile phone data by space-time regression with differential regularization

Publisher:

Springer

Published version:

DOI:10.1007/978-3-030-47756-1_2

MOX-Report No. 54/2020

Analysis of Telecom Italia mobile phone data by space-time regression with differential regularization

Arnone, E.; Bernardi, M. S.; Sangalli, L. M.; Secchi, P.

MOX, Dipartimento di Matematica
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

mox-dmat@polimi.it http://mox.polimi.it

Published in Functional and High-Dimensional Statistics and Related Fields, Springer, 2020, 5–10
Journal version available at https://doi.org/10.1007/978-3-030-47756-1_2

Analysis of Telecom Italia mobile phone data by space-time regression with differential regularization

Eleonora Arnone, Mara S. Bernardi, Laura M. Sangalli, Piercesare Secchi

MOX– Dipartimento di Matematica, Politecnico di Milano

Abstract

We apply spatio-temporal regression with partial differential equation regularization to the Telecom Italia mobile phone data. The technique proposed allows to include specific information on the phenomenon under study through a definition of the non-stationary anisotropy characterizing the spatial regularization based on the texture of the domain on which the data are observed.

1 Space-Time regression with differential regularization

The analysis of functional data with spatial dependence has been of great interest in the last years and various methods have been recently proposed to deal with this kind of data [10]. In this work, we consider spatial regression methods with Partial Differential Equation (PDE) regularization [12, 13, 4, 5]. In particular, we consider the Space-Time regression with PDE penalization method (ST-PDE) introduced in [7] and extend it to deal with observations featuring complex spatial dependency.

ST-PDE is a penalized regression method that models separately the spatial and the temporal regularization by considering two roughness penalties, which account separately for the regularity of the field in space and in time by using a tensor product, following the approach used also by [1, 3, 9]; while, in the generalization of the technique proposed by [2], a single roughness penalty is used to jointly model the spatial and temporal dimensions. Therefore, in the ST-PDE model, the field is estimated minimizing a functional composed by three parts: a data-fitting part, a penalization for the spatial regularity, and a penalization for the temporal regularity. In [7], the spatial penalization involves a simple differential operator that imposes smoothness to the solution. Instead, in this work, we consider a spatial penalization involving a more general PDE, that allows to impose non-stationary anisotropy to the solution, thus modeling more complex spatial dependencies. Moreover, the PDE can model problem-specific knowledge on the phenomenon under study. For example, if the PDE governing the physical phenomenon generating the data is available, it can be exploited in the spatial regularization term of the ST-PDE functional, thus driving the estimation towards a physically sound solution. In the context of the analysis of mobile phone data, where no physical knowledge on the phenomenon under study is available, we use the PDE to include in the model information about the texture of the spatial domain; in particular, we here characterize the PDE using the road network, which highly influences the data. This application highlights the high flexibility of the definition of spatial dependence imposed by the ST-PDE model.

Section 2 describes the Telecom Italia mobile phone data. Section 3 presents the model and how the texture of the domain can be used to estimate the non-stationary anisotropy characterizing the regularization.

2 Telecom Italia mobile phone data

We consider the Telecom Italia database, provided by Convenzione di Ricerca DiAP–Politecnico di Milano and Telecom Italia. This dataset concerns the usage of mobile phone data in the metropolitan area of Milan. It collects the measurements of the Erlang, a dimensionless unit calculated by adding up the length of all the calls made by mobile phones within a region of the spatial domain in a time interval, and dividing the sum by the length of the time interval. In the case of the Telecom Italia database, Erlang data are collected over time intervals of 15 minutes from Wednesday, March 18th 2009, 00:15 to Tuesday, March 31st 2009, 23:45 on a uniform lattice of 97×109 sites with dimension $232\text{m} \times 309\text{m}$ covering the metropolitan area of Milan. In Figure 1, the top panel shows the map of the metropolitan area of Milan on which the data are observed, the central panel shows the Erlang data for a fixed time instant, the bottom panel shows the data in a fixed spatial location.

Since the Erlang is a measurement of the average number of active mobile phones, these data can be considered as an approximation of the number of people present in the considered sites during the sampling time windows. Therefore, the goal of the analysis of these data is the study of the population distribution and dynamics. Indeed, this dataset has been used in the context of the Green Move Project, an an interdisciplinary research project financed by Regione Lombardia and focused on the development of a vehicle sharing system. Some works on this dataset are [8, 14, 17, 11, 15].

The data can be interpreted as a sampling of temporal curves with spatial

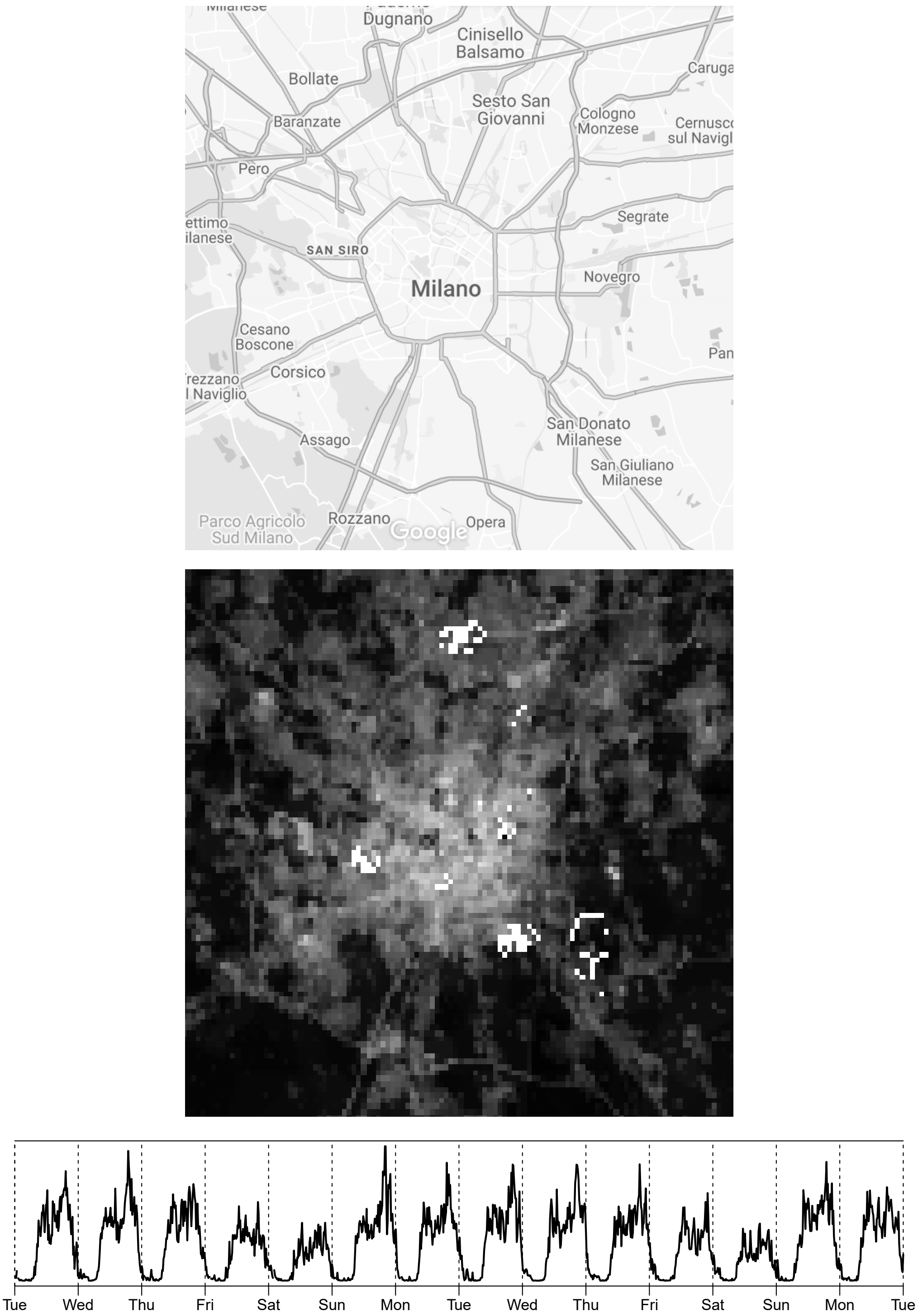

Figure 1: Telecom Italia mobile phone data. Top panel: the metropolitan area of Milan, the spatial domain of the dataset. Central panel: data for a fixed time instant (white corresponds to missing data). Bottom panel: evolution in time of the data for a fixed spatial location.

dependencies; equivalently, they can also be interpreted as a sampling of spatial surfaces with temporal dependencies. In both interpretations, the data are functional in nature and using the functional data analysis framework allows us to properly characterize the complex dependencies and extract meaningful results.

Furthermore, the data are integrals over both time and space of the quantity of interest, since each Erlang datum is a cumulative measurement over a 15-minutes time interval and a 232m×309m site. Therefore, the analysis should properly take into account the fact that the data are areal in space and integral in time.

Moreover, as Figure 1 shows, the spatial distribution of the data is strongly influenced by the characteristics of the urban area considered. Therefore, it is of paramount importance to take into consideration the spatial dependence driven by physical phenomenon generating the data, i.e. the population dynamics in the metropolitan area of Milan, and to adapt the estimation technique to properly take into account the characteristics of the specific urban configuration under study.

Next Section deals with the characterization of the spatial dependence of the data through the definition of a penalization term involving a non-stationary anisotropic diffusion operator which represents the structure of the underlying spatial domain.

3 ST-PDE model and estimating the non-stationary anisotropy

In the ST-PDE functional, the classical square L^2-norm of the second derivative is employed for the temporal penalty, while we need a term which allows us to model non-stationary anisotropy for the spatial penalty. This is obtained by penalizing the misfit from a diffusion PDE $-\mathrm{div}(K(\mathbf{p})\nabla f) = 0$, where $K(\mathbf{p})$ is a function defined on the spatial domain, taking values in the space of symmetric and positive definite 2×2-matrices. When K is a constant function equal to the identity matrix all over the spatial domain, the smoothing is isotropic in space (which is the case considered in [7]); otherwise, the smoothing is anisotropic. If, moreover, K is non-constant as a function of the spatial location $\mathbf{p}$, the smoothing is non-stationary. In our work, we exploit the texture of the spatial domain to estimate the symmetric tensor $K(\mathbf{p})$.

We can observe, from Figure 1, that the number of active phones presents localized strongly anisotropic features in correspondence of the main roads. Thus, we want to use the information about the morphology of the road network of the city to include non-stationary anisotropy in the ST-PDE model. The motivation for our choice is that, when we deal with cars moving on highways, we know that it is more probable that these cars will stay in the highway then that they will exit. Thus, for the spatial locations corresponding to main roads, we want to impose anisotropic smoothing that smooths more in the direction tangential to the road, and less in the other directions.

We use data form Regione Lombardia about the road network of the metropolitan area of Milan (see Figure 2, left panel), in order to estimate $K(\mathbf{p})$ from the city texture. In particular, we select the main roads and highways (see Figure 2, right panel) and exploit the orientation of the roads to identify the direction of the major axis of $K(\mathbf{p})$, i.e. the eigenvector corresponding to the larger eigenvalue. Indeed, for each spatial location $\mathbf{p}$, the direction of the major axis of $K(\mathbf{p})$ can be defined by looking at the road map at a small scale that allows to consider one road at a time. Where no roads are present, the isotropic diffusion operator is used.

The intensity of the anisotropy can be set either exploiting prior knowledge on the phenomenon (for example, the speed limits of the roads) or extracting information from the data using an approach similar to [6], which proposes to estimate the anisotropy directly from the data.

The use of the ST-PDE model with a spatial regularization involving a non-stationary and anisotropic diffusion differential operator carrying information about the road network is particularly useful in the analysis of Telecom Italia mobile phone data, since this technique is able to suitably capture the non-trivial spatial dependencies of the observed data.

Figure 2: Road network in the metropolitan area of Milan. Left panel: a view of the area from Google maps which includes main roads, secondary roads, highways and railways. Right panel: main roads and highways from www.geoportale.regione.lombardia.it used to estimate $K(\mathbf{p})$.

References

[1] Aguilera-Morillo M. C., Durbán M., Aguilera A. M.: Prediction of functional data with spatial dependence: a penalized approach. Stochastic Environ Res Risk Assess 1–16 (2016) doi: 10.1007/s00477-016-1216-8

[2] Arnone E., Azzimonti L., Nobile F., Sangalli L. M.: Modeling spatially dependent functional data via regression with differential regularization. Journal of Multivariate Analysis, 170, 275–295 (2019)

[3] Augustin N. H., Trenkel V. M., Wood S. N., Lorance P.: Space-time modelling of blue ling for fisheries stock management. Environmetrics **24(2)** 109–119 (2013)

[4] Azzimonti L., Nobile F., Sangalli L. M., Secchi P.: Mixed Finite Elements for Spatial Regression with PDE Penalization. SIAM/ASA Journal on Uncertainty Quantification **2(1)**, 305–335 (2014)

[5] Azzimonti L., Sangalli L. M., Secchi P., Domanin M., Nobile F.: Blood flow velocity field estimation via spatial regression with PDE penalization. Journal of the American Statistical Association **110(511)**, 1057–1071 (2015)

[6] Bernardi M. S., Carey M., Ramsay J. O., Sangalli L. M.: Modeling spatial anisotropy via regression with partial differential regularization. Journal of Multivariate Analysis, 167, 15–30 (2018)

[7] Bernardi M. S., Sangalli L. M., Mazza G., Ramsay J. O.: A penalized regression model for spatial functional data with application to the analysis of the production of waste in Venice province. Stochastic Environ Res Risk Assess, 1–16, (2016)

[8] Manfredini F., Pucci P., Secchi P., Tagliolato P., Vantini S., Vitelli V.: Treelet decomposition of mobile phone data for deriving city usage and mobility pattern in the milan urban region. In: Advances in complex data modeling and computational methods in statistics, pp 133–147. Springer (2015)

[9] Marra G., Miller D. L., Zanin L.: Modelling the spatiotemporal distribution of the incidence of resident foreign population. Statistica Neerlandica **66(2)** 133–160 (2012)

[10] Mateu J., Romano E.: Advances in spatial functional statistics. Stochastic Environ Res Risk Assess (2016) doi: 10.1007/s00477-016-1346-z

[11] Passamonti F., Spatio-temporal mobile phone data in Milan: Bagging-Voronoi exploration and modeling through soil use and land cover data. Master's thesis, Politecnico di Milano, MOX - Dipartimento di Matematica (2016)

[12] Ramsay, T.: Spline smoothing over difficult regions. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **54(2)**, 307–319 (2002)

[13] Sangalli L. M. and Ramsay J. O. and Ramsay T. O.: Spatial spline regression models. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **75(4)**, 681–703 (2013)

[14] Secchi P., Vantini S., Vitelli V.: Analysis of spatio-temporal mobile phone data: a case study in the metropolitan area of Milan. Statistical Methods & Applications **24(2)**, 279–300 (2015)

[15] Secchi P., Vantini S., Zanini P.: Analysis of Mobile Phone Data for Deriving City Mobility Patterns. In Electric Vehicle Sharing Services for Smarter Cities (pp. 37-58). Springer, Cham (2017)

[16] Xun X., Cao J., Mallick B., Maity A., Carroll R. J.: Parameter Estimation of Partial Differential Equation Models. Journal of the American Statistical Association **108(503)**, 1009–1020 (2013)

[17] Zanini P., Shen H., Truong Y.: Understanding resident mobility in Milan through independent component analysis of Telecom Italia mobile usage data. The Annals of Applied Statistics **10(2)**, 812–833 (2016)

MOX Technical Reports, last issues

Dipartimento di Matematica
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

49/2020 Bonaventura,L.; Garres Diaz,J.
Flexible and efficient discretizations of multilayer models with variable density

50/2020 Bonaventura,L.; Gomez Marmol, M.
The TR-BDF2 method for second order problems in structural mechanics

51/2020 Ferraccioli, F.; Sangalli, L. M.; Arnone, E.,; Finos, L.
A functional data analysis approach to the estimation of densities over complex regions

52/2020 Arnone, E.; Kneip, A.; Nobile, F.; Sangalli, L. M.
Some first results on the consistency of spatial regression with partial differential equation regularization

53/2020 Arnone, E.; Kneip, A.; Nobile, F.; Sangalli, L. M.
Some numerical test on the convergence rates of regression with differential regularization

48/2020 Clerici, F.; Ferro, N.; Marconi, S.; Micheletti, S.; Negrello, E.; Perotto, S.
Anisotropic adapted meshes for image segmentation: application to 3D medical data

46/2020 Bucelli, M.; Salvador, M.; Dede', L.; Quarteroni, A.
Multipatch Isogeometric Analysis for Electrophysiology: Simulation in a Human Heart

45/2020 Gatti, F.; Menafoglio, A.; Togni, N.; Bonaventura, L.; Brambilla, D.; Papini, M; Longoni, L.
A novel dowscaling procedure for compositional data in the Aitchison geometry with application to soil texture data

44/2020 Masci, C.; Ieva, F.; Paganoni A.M.
EM algorithm for semiparametric multinomial mixed-effects models

47/2020 Sangalli, L.M.
A novel approach to the analysis of spatial and functional data over complex domains