Stat Methods Appl (2015) 24:315–319
DOI 10.1007/s10260-015-0311-1

DISCUSSION

Discussion of the paper "analysis of spatio-temporal mobile phone data: a case study in the metropolitan area of Milan"

Orietta Nicolis[1] · Jorge Mateu[2]

Accepted: 13 April 2015 / Published online: 25 April 2015

Abstract The authors are to be congratulated on a valuable and thought-provoking contribution on the analysis of geo-referenced high-dimensional data describing the use over time of the mobile-phone network in the urban area of Milan, Italy. This is a timely and world-wide problem that opens wide avenues for new methodological contributions. The authors develop a Bagging Voronoi Treelet Analysis which is a non-parametric method for the analysis of spatially dependent functional data. This approach integrates the treelet decomposition with a proper treatment of spatial dependence, obtained through a Bagging Voronoi strategy. In our discussion, we focus on the following points: (i) a mobre general form of the spatio-temporal model proposed in Secchi et al. (Stat Methods Appl, 2015), (ii) alternative methods to approach the smooth temporal functions, (iii) additional methods to reduce the problem of dimension for spatial dependence data, and (iv) comments on the pros and cons of the proposed pre-processing methodology.

Keywords Basis functions · Dimension reduction · Gaussian random fields · Spatially dependent functional data · Spatio-temporal stochastic models

1 The spatio-temporal model

The approach used in Secchi et al. (2015) for analyzing the spatial and temporal variability of Erlang data bears strong resemblance to some spatio-temporal models

✉ Jorge Mateu
mateu@uji.es

Orietta Nicolis
orietta.nicolis@uv.cl

[1] Institute of Statistics, University of Valparaiso, Valparaiso, Chile

[2] Department of Mathematics, University Jaume I of Castellón, Castellón, Spain

proposed by Olives et al. (2014) and Lindström et al. (2014) for studying air pollution data. Similarly to air pollution data, the spatio-temporal distribution of Erlang quantities is characterized by both a strong temporal seasonality and a strong spatial dependence.

Using the approach introduced by Lindström et al. (2014), the spatio-temporal Erlang data $E_{\mathbf{x}(t)}$ can be modeled in a more general form as

$$E_{\mathbf{x}}(t) = y(\mathbf{x}, t) = \mu(\mathbf{x}, t) + \epsilon(\mathbf{x}, t) \tag{1}$$

where

$$\mu(\mathbf{x}, t) = \sum_{k=1}^{K} \beta_k(\mathbf{x})\psi_k(t). \tag{2}$$

The $\{\psi_k(t)\}_{k=1}^{K}$ is a set of (smooth) temporal basis functions with $\psi_1(t) = 1$ that can be estimated by the modified singular value decomposition method (see Fuentes et al. 2006; Szpiro et al. 2010). The terms $\beta_k(\mathbf{x})$ are spatially varying coefficients for the temporal functions that can be estimated using universal kriging (Matheron 1969), where the trend is a linear regression on geographical covariates and the spatial dependence structure is provided by a set of covariance matrices given by $\Sigma_{\beta_k}(\theta_k)$, parameterized by an unknown parameter vector θ_k. In the particular case of Erlang data, the trend of the regression kriging could contain information about land use (i.e, university, residential, or industrial areas).

The residual space-time component $\epsilon = \epsilon(\mathbf{x}, t)$ is assumed to be independent in time with stationary, parametric spatial covariance $\Sigma_{\epsilon}^{t}(\theta_{\epsilon})$, for $t = 1, \ldots, T$. In particular, the residual $\epsilon(\mathbf{x}, t)$ consists of a correlated component $\epsilon^*(\mathbf{x}, t)$ and a nugget-effect $\epsilon_{nugget}(\mathbf{x}, t)$ including small-scale variability and measurement errors, that is,

$$\epsilon(\mathbf{x}, t) = \epsilon^*(\mathbf{x}, t) + \epsilon_{nugget}(\mathbf{x}, t). \tag{3}$$

Assuming the independence of the components of (3), the spatial covariance of $\epsilon(\mathbf{x}, t)$ can be written as $\Sigma_{\epsilon} = \Sigma_{\epsilon}^{*} + \Sigma_{\epsilon,\text{nugget}}$, where $\Sigma_{\epsilon,\text{nugget}}$ is a diagonal matrix.

Then note that the model proposed in Secchi et al. (2015) could be considered a particular case of (1) where the components β_k (denoted in Secchi et al. (2015) by $D_1, \ldots, D_K$) do not take into account the spatial dependence, and the residual component ϵ is a random error variable, independent in time and space.

In addition, note that an alternative definition of the mean component $\mu(\mathbf{x}, t)$ in (1) has been recently proposed by Olives et al. (2014) as

$$\mu(\mathbf{x}, t) = \sum_{k=1}^{K} \{\beta_k(\mathbf{x}) + \gamma_k(\mathbf{x})\}\psi_k(t),$$

where $\beta_k(\mathbf{x})$ are Gaussian spatial random fields distributed as $\beta_k(\mathbf{x}) \sim N(0, \Sigma_{\beta_k}(\theta_k))$ as in Lindström et al. (2014), and $\gamma_k(\mathbf{x})$ are i.i.d. random effects distributed as $\gamma_k \sim N(0, \sigma_k^2\mathbf{I})$. They can be considered the nugget effect of the $\beta_k(\mathbf{x})-$fields.

1.1 Smooth temporal functions

The objective of the smooth temporal basis functions $\psi_k(t)$ is to capture the temporal variability in the data using deterministic functions, or functions obtained as smoothed singular vectors (see Fuentes et al. 2006). Nicolis and Nychka (2012) and Matsuo et al. (2011) suggest to use the non-orthogonal wavelet basis (such as the W-transform) for their ability to fit a variety of standard covariance models. The mayor drawback of these approaches is that one needs to specify the functional form of the basis ψ. The treelets used in Secchi et al. (2015) provide an interesting tool to the analysis of the temporal behavior of Erlang data, especially for their feature of being 'data-driven' basis.

However, other methods take different approaches to construct data-driven basis. The Tree-Based Wavelet (Gavish et al. 2010) and the lifting scheme are some examples. While the Tree-Based Wavelet transform is defined via a hierarchical tree (built through and adaptive Haar-like orthonormal basis) which is assumed to capture the geometry and structure of the input data, the lifting schemes provide a simple and general construction of second generation wavelets, where the choice for primal and/or dual lifting is fully determined by the values of the data (Jansen and Ooníncx 2005; Sweldens 1997).

In particular, the lifting scheme allows one to custom design the filters needed in the transform algorithms to the situation at hand, that is, the filters generate functions whose form depends on each particular case. Finally, lifting scheme leads to a fast, fully in-place implementation of the wavelet transforms (Sweldens 1997). We think that these methods could provide alternative tools to the analysis of temporal Erlang data.

1.2 Spatial dependence and dimension reduction

Parameter estimation of a spatio-temporal model tends to be challenging in practice. Methods for reducing the computational burden are becoming more common when the data set is very large. Some recent methods are based on 'low-rank' (or 'reduced rank') approaches which aim is to reduce the spatial process to a dimensional subspace of a lower dimension in order to increase the computational efficiency (see Banerjee et al. 2008; Nicolis and Nychka 2012; Olives et al. 2014). The idea of the low rank approach proposed by Olives et al. (2014) is to replace the covariance $\Sigma = \{||C(\mathbf{x}_i - \mathbf{x}_j)||\}_{i,j \in S}$ where S is the observed space of spatial locations $\mathbf{x}$, with a low rank covariance $Z\tilde{\Sigma}^{-1}Z$ where $Z = \{||C(\mathbf{x}_i - \kappa_j)||\}_{i \in \mathcal{S}, j \in \mathcal{K}}$, $\tilde{\Sigma} = \{||C(\kappa_i - \kappa_j)||\}_{i,j \in \mathcal{K}}$, and $\mathcal{K}$ is a set of spatial locations κ, of cardinality $n \ll N$ (N is the number of observations in the space $\mathcal{S}$). Then, the β_k-fields can be approximated by a vector with dimension $n \times 1$. A similar approach has been used by Nicolis and Nychka (2012) where the authors use a multiresolution approach based on non-orthogonal wavelet functions to reduce the dimension of the original space. The conditional simulation is then used for estimating the process over all the original space. Banerjee et al. (2008) use 'knots' and predictive processes for reducing the dimension.

In the approach proposed by Secchi et al. (2015) the spatial dependence has been estimated empirically over several subsets of data using the Voronoi tessellation, and 'low rank' matrices are produced. Then summary statistics on the simulation results (using the bootstrap technique) provides the estimation of the process on the complete space. We think that the low-rank above mentioned approaches can be considered an important contribution for the estimation of spatial dependence of non-stationary processes. For all these approaches the optimal choice of the subset of data n remains an open problem.

2 Pre-processing of data: denoising with missing data

If some data are missing, several denoising methods cannot be directly implemented. When the data are spatially and temporally correlated many methods have been proposed for infilling missing data and smoothing irregular curves (Glasbey 1995; Haworth and Cheng 2012; Olives et al. 2014; Onorati et al. 2013; Smith et al. 1996, 2003). For example, Smith et al. (1996) use spatial patterns from EOF for reconstructing the data in a given temporal period, and (Olives et al. 2014; Onorati et al. 2013) apply cubic smoothing splines to some of the left singular vectors of the singular vector decomposition. A comparison of methods for smoothing and gap filling in time series has been proposed by Kandasamy et al. (2013).

Similarly, Erlang data show strong spatial dependence in the principal surfaces that can be used for infilling temporal data. The Fourier-based technique used by Secchi et al. (2015) for denoising and infilling missing data have the following advantages: (i) it transforms discrete data into functionals that can be used in the space-time model; (ii) it is considered a denoising technique; and (iii) it resolves the problem of missing data.

However, the main drawbacks of the proposed pre-processing methodology are that: (i) it needs to choose a basis of very high dimension, in order to be sure to catch up all relevant localized features, with a consequent increase of the computational burden, and (ii) it does not consider the spatial dependence among sites.

We think that including denoising and imputation of missing data in the estimation procedure of the space-time model could improve the computational efficiency of the algorithm and the precision of the estimates.

Acknowledgments The present work has been partially supported by Fondecyt grant 1131147 from the Chilean government, and research funds of the University of Palermo with reference 2012-ATE-0332. Additionally, partial funding comes from grants P1-1B2012-52, and MTM2013-43917-P from Ministery of Economy and Competitivity.

References

Banerjee S, Gelfand AE, Finley AO, Sang H (2008) Gaussian predictive process models for large spatial data sets. J R Stat Soc B 70:825–848

Fuentes M, Guttorp P, Sampson PD (2006) Using transforms to analyze space-time processes. In: Finkenstadt B, Held L, Isham V (eds) Statistical methods for spatio-temporal systems. CRC/Chapman and Hall, Boca Raton, pp 77–150

Gavish M, Nadler B, Coifman RR (2010) Multiscale wavelets on trees, graphs and high dimensional data: theory and applications to semi supervised learning. In: Proceedings of the 27th international conference on machine learning

Glasbey CA (1995) Imputation of missing values in spatio-temporal solar radiation data. Environmetrics 6(4):363–371

Haworth J, Cheng T (2012) Non-parametric regression for space-time forecasting under missing data. Comput Environ Urban Syst 36(6):538–550

Jansen M, Oonincx P (2005) Second generation wavelets and applications. Springer, London

Kandasamy S, Baret F, Verger A, Neveux P, Weiss M (2013) A comparison of methods for smoothing and gap filling time series of remote sensing observations-application to MODIS LAI products. Biogeosciences 10(6):4055–4071

Lindström J, Szpiro AA, Sampson PD, Oron A, Richards M, Larson T, Sheppard L (2014) A flexible spatio-temporal model for air pollution with spatial and spatio-temporal covariates. Environ Ecol Stat 21:411–433

Matheron G (1969) Part 1 of Cahiers du Centre de morphologie mathématique de Fontainebleau. Le krigeage universel. Ècole nationale supérieure des mines de Paris

Matsuo T, Nychka D, Paul D (2011) Nonstationary covariance modeling for incomplete data: Monte Carlo EM approach. Comput Stat Data Anal 55:2059–2073

Nicolis O, Nychka D (2012) Reduced rank covariances for the analysis of environmental data. In: Ciacco Di A, Coli M, Ibanez JMA (eds) Advanced statistical methods for the analysis of large data-sets. Series: Studies in Theoretical and Apllied Statistics, Springer

Olives C, Kaufman JD, Sheppard L, Szpiro AA, Lindström J, Sampson PD (2014) Reduced-rank spatio-temporal modeling of air pollution concentrations in the multi-ethnic study of atherosclerosis and air pollution. Ann Appl Stat 8(4):2509–2537

Onorati R, Sampson P, Guttorp P (2013) A spatio-temporal model based on the SVD to analyze daily average temperature across the Sicily region. J Environ Stat 5(2)

Secchi P, Vantini S, Vitelli V (2015) Analysis of spatio-temporal mobile phone data: a case study in the metropolitan area of Milan. Stat Methods Appl. doi:10.1007/s10260-014-0294-3

Smith RL, Kolenikov S, Cox LH (2003) Spatio-temporal modeling of PM2.5 data with missing values. J Geophys Res Atmos 108:D24, 9004

Smith TM, Reynolds RW, Livezey RE, Stokes DC (1996) Reconstruction of historical sea surface temperatures using empirical orthogonal functions. J Clim 9:1403–1420

Sweldens W (1997) The lifting scheme: a construction of second-generation wavelets. SIAM J Math Anal 29(2):511–546

Szpiro AA, Sampson PD, Sheppard L, Lumley T, Adar SD, Kaufman JD (2010) Predicting intra-urban variation in air pollution concentrations with complex spatio-temporal dependencies. Environmetrics 21:606–631