Rejoinder to the discussion of "Analysis of Spatio-Temporal Mobile Phone Data: a Case Study in the Metropolitan Area of Milan"

Piercesare Secchi[1] · **Simone Vantini**[1] · **Valeria Vitelli**[2]

Accepted: 15 May 2015 / Published online: 27 May 2015

We thank all discussants for their deep and stimulating comments on our analysis of the Erlang dataset describing the use over time of the mobile phone network in the urban area of Milan: we are glad that our paper has evoked so many new thoughts and ideas for the statistical exploration of non-stationary spatial fields of functional data. These encouraging reactions confirm our prior belief that this dataset is both rich and challenging, and that the research which motivated its analysis is worth being proposed to a scholarly community. Even if many discussants raise similar themes or methodological recommendations (e.g. spatio-temporal models, functional principal component analysis, wavelets, ...), the proposed approaches do not always share the same motivations, and even less frequently the same purposes. Far from being

✉ Valeria Vitelli
valeria.vitelli@medisin.uio.no

Piercesare Secchi
piercesare.secchi@polimi.it

Simone Vantini
simone.vantini@polimi.it

1 MOX - Dipartimento di Matematica, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133 Milan, Italy

2 Department of Biostatistics, Oslo Center for Biostatistics and Epidemiology, University of Oslo, Sognsvannsveien 9, Domus Medica, 0372 Oslo, Norway

surprised, we are convinced that such diverse and heterogeneous annotations can only make the discussion more interesting, and we are highly gratified to supplement the statistical community working on spatially dependent functional data with such a stimulating data problem.

First of all, we agree with Delicado (2015) that the Erlang dataset is a case of Big Data open to many non trivial statistical analyses, some of them bound to extract similar information. In the paper we do analyze two weeks' worth of data, but yearly data are in principle available. Moreover, as hinted by Nicolis and Mateu (2015), a more complex statistical investigation of the population dynamics on the dense *textured domain* (Ramsay 2015) of Milan metropolitan area, would correlate the analysis of the Erlang data with that of observed spatio-temporal covariates, such as those provided by modern GIS technologies or OpenData urban portals. This richer and broader piece of information on local urban features could then be used to lay out a deeper interpretation of the results. Hence, we do believe that together with Volume and Velocity, already mentioned in Delicado (2015), the third V defining Big Data– Variety— can also come into play for the analysis of the Erlang dataset.

Some discussants raise the question whether a spatio-temporal analysis would be more suited to our Erlang data, for instance by modelling the Erlang function $E_{\mathbf{x}}(t)$ as the sum of a deterministic drift and a random residual space-time component having stationary parametric spatial covariance (Nicolis and Mateu 2015). These proposals, albeit being interesting, might not be practical for the analysis of the Erlang data: any stationary covariance model is way too restrictive for the complexity of the spatial features which we aim at capturing, not to mention that parameter estimation of a spatio-temporal model tends to be challenging in practice (Nicolis and Mateu 2015). We still think that our non-parametric Bagging Voronoi approach is a good compromise between the need for a flexible way to handle the complex spatial features hidden in the Erlang data, and computational feasibility. Indeed more flexibility could be added to the Bagging Voronoi Treelet Analysis (BVTA) following the suggestions made in Sørensen et al. (2015), where a different use of the Total Average Variance (TAV) criterion is proposed to adapt the Voronoi tessellation to heterogeneous spatial features: studying the site-wise bootstrap variance could lead to the identification of regions where the Voronoi cells should be more or less dense. Moreover, local representatives could be computed using a kernel smoother on all sites in the vicinity of the Voronoi centres. Modifying the BVTA in these directions would require virtually no effort. We believe the entire analysis might benefit from these proposals, and we will thus keep them in mind for future research. A different intriguing alternative is formulated by Antoniadis and Poggi (2015), who propose to model the number $N_{\mathbf{x}}(t)$ of mobile phones using the network within site $\mathbf{x}$ at time t as an inhomogeneous Poisson process. Count data are then transformed and analyzed through a general linear model. This modeling approach is worth investigating, even in combination with a Bagging Voronoi strategy for capturing non-stationarity in spatial dependence. Moreover the fitting of the linear model for the transformed data could be improved by introducing a roughness penalty described by a differential equation, as illustrated in Ramsay (2015).

Another possible modeling strategy consists in taking spatial dependence into account when reducing the dimensionality of the data through a representation obtained by projection on a suitable functional basis $\{\psi_k(t),\, k \geq 1\}$ (Delicado 2015;

Kokoszka 2015; Sørensen et al. 2015), for instance by performing functional principal component analysis (FPCA) for spatially dependent functional data as proposed in Horvath and Kokoszka (2012). These approaches are indeed conceivable, and proved to have good performances at least in the context of our simulation studies, as dis-cussed by Delicado (2015). We believe the main drawback of FPCA is that of being a global method subject to an orthogonality constraint, which is not really justified in the present application. Modifying the FPCA criterion to enforce sparsity and smoothness of the basis functions (Sørensen et al. 2015), or even trying local FPCA (Antoniadis and Poggi 2015), are other possible strategies to catch local features, which however do not remove the orthogonality constraint. The choice of using treelets goes in this same direction: treelets are also orthogonal, and they share with wavelets, mentioned by many discussants, the property of being a multi-resolution basis. However, differently from wavelets, treelets are also data driven, something that we deem very important for representing the internal structure of the Erlang data, as pointed out also by Antoniadis and Poggi (2015) and Nicolis and Mateu (2015). Inspired by treelets and by Independent Component Analysis, we are developing a data driven, sparse and multi-resolution functional basis, Hierarchical Independent Component Analysis (HICA). This novel functional basis is not constrained by orthogonality, and it looks very promising for obtaining a low dimensional and informative representation of Erlang data, and more generally for performing the blind source separation of functional data (Secchi et al. 2014).

On the other hand, different strategies to handle spatial dependence when finding a good representation of data in a space of reduced dimension have already been proposed (Sangalli et al. 2013), and this framework is particularly useful when the application at hand can provide a strong physical insight into the phenomenon under study (Azzimonti et al. 2014). Since a non-linear Darcy's filtration law for diffusion-advection in a porous medium would look reasonable as a physical model for traffic flows within the dense urban texture of Milan (Della Rossa et al. 2010), we have included this possible approach among our future research directions. These considerations bring us to the preprocessing part of the analysis: when smoothing temporal data, having a differential operator in mind could help in choosing a roughness parameter so that typical variation is not much penalized by the penalty. This approach has been strongly advocated by Jim Ramsay, who is to be congratulated for the suggestive "eyeballed" construction of a roughness penalty suited to the modeling of the aggregated Erlang data curve (Ramsay 2015). This is an insightful suggestion, which we should keep in mind in further explorations of the Erlang dataset. Indeed we agree with Nicolis and Mateu (2015) that smoothing Erlang data with a Fourier basis might be sub-optimal, and that denoising and/or missing data imputation are valid alternatives, even though the computational burden would be much increased. Other approaches suggested by the discussants, such as functional regression for incomplete curves (Kokoszka 2015), could also be appropriate.

Changing the role of space and time has been evoked in many comments (Antoniadis and Poggi 2015; Gonzáles-Manteiga and Crujeiras 2015; Sørensen et al. 2015), and pursuing this change of perspective might reveal different interesting data features. However, as pointed out by the discussants, representing the spatio-temporal process as a linear combination of few spatial basis functions (e.g. 2D wavelets, tensor splines)

with uncorrelated random coefficients $\psi_k(t)$ gives a different meaning to the model components. This change of roles is undoubtedly worth trying, even though it seems more suited to other applications, such as brain imaging, where the focus is on spatial activation of brain regions rather than temporal activation of mobile users.

Finally, the use of the proportion of explained total variance as a criterion for tuning the number K of treelets has raised concerns among some of the discussants, who proposed to use a variant of the best basis algorithm (Antoniadis and Poggi 2015), or other non-parametric approaches such as cross-validation (Grané and Romera 2015). These attempts are conceivable, and we might explore them further in the future. In the present paper we followed a parsimonious Goldilocks approach when selecting a "just right" number of treelets leading to interpretable components.

There are many other insightful comments made by the discussants that we did not have the time or the knowledge to touch upon. We deeply thank all of them for rendering our analysis of the Erlang dataset so much richer and thought provoking.

References

Antoniadis A, Poggi G (2015) Discussion of "Analysis of spatio-temporal mobile phone data: a case study in the metropolitan area of Milan". Stat Methods Appl. doi:10.1007/s10260-015-0309-8

Azzimonti L, Sangalli L, Secchi P, Domanin M, Nobile F (2014) Blood flow velocity field estimation via spatial regression with PDE penalization. J Am Stat Assoc. doi:10.1080/01621459.2014.946036 Delicado P (2015) Discussion of "Analysis of spatio-temporal mobile phone data: a case study in the metropolitan area of Milan" by P. Secchi, S. Vantini, and V. Vitelli. Stat Methods Appl. doi:10.1007/s10260-015-0320-0

Della Rossa F, D'Angelo C, Quarteroni A (2010) A distributed model of traffic flows on extended regions. Netw Heterog Media 5(3):525–544

Gonzáles-Manteiga V, Crujeiras R (2015) Discussion on the paper: "Analysis of spatio-temporal mobile phone data: a case study in the metropolitan area of Milan" by P. Secchi, S. Vantini, and V. Vitelli. Stat Methods Appl. doi:10.1007/s10260-015-0318-7

Grané A, Romera R (2015) Piercesare Secchi, Simone Vantini and Valeria Vitelli: analysis of spatio-temporal mobile phone data: a case study in the metropolitan area of Milan. Stat Methods Appl. doi:10.1007/s10260-015-0310-2

Horvath L, Kokoszka P (2012) Inference for functional data with applications. Springer series in statistics

Kokoszka P, 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/s10260-015-0300-4

Nicolis O, Mateu G (2015) Discussion of the paper "analysis of spatio-temporal mobile phone data: a case study in the metropolitan area of Milan". Stat Methods Appl. doi:10.1007/s10260-015-0311-1

Ramsay J (2015) Discussion of Secchi, Vantini and Vitelli paper. Stat Methods Appl. doi:10.1007/s10260-015-0312-0

Sangalli L, Ramsay J, Ramsay T (2013) Spatial spline regression models. J R Stat Soc Ser B 75(4):681–703

Secchi P, Vantini S, Zanini P (2014) Hierarchical independent component analysis: a multi-resolution non-orthogonal data-driven basis. Tech. Rep. 01, MOX, Dipartimento di Matematica, Politecnico di Milano

Sørensen H, Markussen B, Tolver A (2015) Discussion of "Analysis of spatio-temporal mobile phone data: a case study in the metropolitan area of Milan" by P. Secchi, S. Vantini, and V. Vitelli. Stat Methods Appl. doi:10.1007/s10260-015-0317-8