Stat Methods Appl (2015) 24:313–314
DOI 10.1007/s10260-015-0310-2

DISCUSSION

Piercesare Secchi, Simone Vantini and Valeria Vitelli: Analysis of spatio-temporal mobile phone data: a case study in the metropolitan area of Milan

Aurea Grané[1] · Rosario Romera[1]

Accepted: 13 April 2015 / Published online: 29 April 2015

We thank Editor Cerioli for giving us the opportunity to read such an interesting contribution to SMAP. We congratulate the authors for the topic chosen for this research which can enlighten researchers in many different areas related to statistical methods and their application.

This paper presents a novel spatio-temporal modelization of the mobile network data provided by the Telecom Italia database that is connected to the interdisciplinary Green Move project financed by Regione Lombardia. The authors offer a new perspective to exploit the Erlang data, which are progressively arousing the enthusiasm of the urban planners' community.

The authors propose a dimensional reduction of spatially dependent signals, by integrating a treelet analysis for dimensional reduction based on Lee et al. (2008) with a Bagging Voronoi strategy for the exploration of spatial dependence proposed by Secchi et al. (2013). The paper is well-written, easy to follow and the provided additional material clarifies the new BVTA algorithm. We think that the proposed methodology is a valuable contribution of great interest for the Functional Data, Data Mining and Big Data Communities.

Sections 2 and 3 are the core of the paper and they are devoted to the proposed methodology and the data preprocessing process, respectively. Along those sections the authors design a complex and powerful instrument named BVTA algorithm. We have to say that at a first glance some results of Section 4 seem somehow unexpected. Honestly, from a statistical point of view we were surprised by the poor explanatory power, in the sense of proportion of explained total variance, exhibited by the estimated treelets with the exception of ψ_1. In fact, according to Figure 6, the sum of the total

✉ Rosario Romera
mrromera@est-econ.uc3m.es

[1] Statistics Department, Universidad Carlos III de Madrid, Madrid, Spain

variance explained by $\{\psi_j\}_{j\geq 2}$ is less than 1 %. One reason for this to happen may be the specificity of this particular Erlang dataset but we will not go further in this regard. We have the impression that our comments in other directions can help more effectively the readers of this paper.

We make a first consideration to authors. According to Lee and Nadler (2007), for dimensional reduction purposes the selection of treelets instead of wavelets represents a suitable choice for unstructured and very noisy data. A natural question arises: if a wavelet basis was used in the present case, would the results shown by Figure 6 be substantially different in terms of the number of significant explanatory elements of the basis and their proportion of explained total variance?

We understand that the selection of both the L^1 distance and the functional median in the estimation step of the BVTA algorithm seeks for the statistical robustification of the proposed algorithm. However, according to the Data Mining Community, other distance measures may be more suitable. Some interesting references are Ding et al. (2008) and Batista et al. (2014).

Concerning the number of treelets to be retained, the paper lacks in introducing a statistical criterion for the selection of parameter K, different from that based on the explained total variance. Can the authors provide any nonparametric technique to evaluate the significance of the estimated elements of the basis?

We think that the authors should consider the possibility of validating the BVTA algorithm through a two-step procedure: calibration and prediction, for instance, by means of crossvalidation methods. If possible, this will provide objective measures to evaluate, for example, the performance of the algorithm when considering alternative basis selection, or alternative functional distances, or even different values for K.

Finally, we have a couple of minor comments. In Section 2.3, the authors should avoid the use of the same symbol $\tilde{d}$ for denoting different things, i.e., the distance between two elements of the treelet basis and the collection of surfaces. In page 11, line 39, the ψ's elements in the estimation step formula should be φ's.

References

Batista GE, Keogh EJ, Tataw OM, Souza VM (2014) CID: an efficient complexity-invariant distance for time series. Data Min Knowl Discov 28(3):634–669

Ding H, Trajcevski G, Scheuermann P, Wang X, Keogh E (2008) Querying and mining of time series data: experimental comparison of representations and distance measures. Proc Very Large Data Bases Endow VLDB 1(2):1542–1552

Lee AB, Nadler B (2007) Treelets, A tool for dimensionality reduction and multi-scale analysis of unstructured data. J Mach Learn W&P 2:259–266

Lee AB, Nadler B, Wassermann L (2008) Treelets—an adaptive multi-scale basis for sparse unordered data. Ann Appl Stat 2(2):435–471

Secchi P, Vantini S, Vitelli V (2013) Bagging Voronoi classifiers for clustering spatial functional data. Int J Appl Earth Obs Geoinf 22:53–64