Spatial outlier detection in the air quality monitoring network of Normandy (France)

M. Bobbia[1], M. Misiti[2,*], Y. Misiti[2,*], J.-M.. Poggi[3,*] and B. Portier[4]

[1] *Air Normand; michel.bobbia@airnormand.fr*
[2] *Univ. Paris-Sud Orsay; michel.misiti@gmail.com, yves.misiti@math.u-psud.fr*
[3] *Univ. Paris-Sud Orsay and Univ. Paris Descartes; jean-michel.poggi@parisdescartes.fr*
[4] *Normandie Université, INSA Rouen; bruno.portier@insa-rouen.fr* *Corresponding author*

Abstract. *We consider hourly PM10 measurements from 22 monitoring stations located in Basse-Normandie and Haute-Normandie regions (France) and also in the neighboring regions. All considered monitoring stations are either urban background stations or rural ones. The paper focused on the outlier statistical detection of the hourly PM10 concentrations from a spatial point of view.*
The general strategy uses a jackknife type approach and is based on the comparison of the actual measure with some robust prediction. Two ways to handle spatial prediction are considered: the first one is based on the nearest neighbors weighted median which directly consider concentrations while the second one is based on kriging increments, instead of more traditional pseudo-innovations.
The two methods are applied to the PM10 monitoring network in Normandy and are fully implemented by Air Normand (the official association for air quality monitoring in Haute-Normandie) in the Measurements Quality Control process. Some numerical results are provided on recent data from January 1, 2013 to May 31, 2013 to illustrate and compare the two methods.

Keywords. *Air quality, Kriging, Nearest neighbors, Particulate matter, Spatial outlier detection.*

1 Pollution context

In France, the air quality is supported in each region by an official association (approved by the Ministry of Ecology). In the Normandy, consisting of two regions, Air Normand, based in Rouen (Haute-Normandie), and Air C.O.M. (Air COM for short in the sequel), based in Caen (Basse-Normandie), monitor air quality. In addition to their supervisory functions, their role is to inform the public regarding the air he breathes. Thus, to fulfill their missions, Air Normand and Air COM measure air quality with automatic analyzers scattered throughout the region, and publicize these measures, including by using a website to inform the public on exposure to air pollution. Air Normand and Air COM work closely to publish their measurements on a common website. In particular, measurements are spatially interpolated

to produce maps of air quality. These maps are also published on the website.

However, it is important to prevent false maps due to outliers. Indeed, Air Normand provides a mapping of air quality on the Normandy region updated every hour. The air quality is derived hourly concentrations of four pollutants (O3, PM10, NO2 and SO2) and the maps provided by the numerical model outputs. Each pollutant is mapped by correcting the numerical model outputs by the concentrations measured in the network stations, using assimilation methods (see for example de Fouquet et al. 2011). Thus undiagnosed measurement errors can seriously affect the quality of the spatial reconstruction of concentrations leading to give bad information to the public.

2 Statistical context

Thus, the aim of this work is twofold. First, to build tools in real time to detect the presence of outliers generating an incorrect map, this would be valuable to validate obtained maps. Second, additionally tools for outlier detection could also help in the measure validation of a specific pollutant network (here the PM10 measure network). We consider in this work carried out with Air Normand, the problem of spatial outlier detection in the context of particulate matter and more precisely the one of PM10, which is the more crucial pollutant in Normandy.

A short survey of the literature about outliers among a large number of references can be quickly performed. For example, we can first highlight the classical book of Barnett (2004), which contains a chapter especially dedicated to this topic as well as some survey papers Planchon (2005), Ben-Gal (2005) or more recently Chandola et al. (2009). However, these references are mainly concerned by univariate or multivariate outliers but not specifically dedicated to the spatial nature of the data. Haslett et al. (1991) as well as Laurent et al. (2006) explore spatial data analytic tools and deduce some outlier detection procedures (see for example Filzmoser et al. 2012). In the case of spatial data, a classical distinction is of interest: between a "global" atypical value, which consists in reasoning starting from the behavior of the majority of data and a "local" atypical value which consists in reasoning from the behavior of the observations that are geographical neighbors. Then four classes of observations can be defined: typical, global atypical only (distinguished by standard tools), local atypical only and the last one: local and global atypical. In this paper, we are interested in detecting local atypical observations. Let us remark that a local atypical observation is often defined as an observation that differs from the closer observations, so it is implicitly assumed that the data exhibit a positive spatial autocorrelation. Of course it is important to check that it is realistic in each specific application. Some references are particularly useful in our case: Cerioli et al. (1999) define a procedure based on kriging schemes and dedicated to multiple outliers while Kou. et al., (2006), Shekhar et al. (2003) and Lu et al., (2003) develop some simple, intuitive and robust ways to detect spatial outliers. Let us finally mention the very recent paper of Li (2013).

3 Spatial outlier detection procedures

The basic idea of the detection algorithm we proposed is based on comparing the measured concentration to some robust spatial prediction, following a classical jackknife type approach. The decision rule is then based on thresholds coming from the distributions of prediction residuals along time. We consider two ways to handle spatial outlier detection depending on the prediction method. The first one, inspired

by Kou et al. (2006) and by Lu et al. (2003), is a nonstationary spatial way by comparing directly the concentration of a given site with the weighted median of the concentrations of its neighbors with respect to a pre specified neighborhood system. The second one is based on kriging the innovations, namely the difference between the current observations and a reference set of past observations or numerical model outputs, and not directly on concentrations.

In addition to an effective implementation of real-world solution and an example of successful collaboration between academics and experts in air quality, the originality of our statistical contribution is, to our knowledge, double. First, about the nearest neighbors method, our idea is to introduce the historical data, that is to say the time, both in the choice of limits of detection and the definition of the neighborhoods by station with the weights associated to neighbors. Second, about the kriging method of pseudo-innovations, our idea is to introduce time for the calculation of detection bounds and to work on kriging increments, which is equivalent to an autoregressive modeling of the map of concentrations.

Our talk will be organized as follows. We describe the PM10 monitoring network, the Measurements Quality Control process, the PM10 data and finally we recall some basics about kriging methodology. Then we give the general principle of the outlier detection procedure and the two methods for spatial outlier detection: one based on the nearest neighbors weighted median and one based on kriging increments instead of pseudo-innovations. Finally we present the outlier detection results on some recent database and then we propose a discussion and some concluding remarks.

4 Conclusion

We have developed a general statistical outlier detection strategy from a spatial point of view. It uses a jackknife type approach and is based on the comparison of the actual measure with some robust prediction. Two ways to handle spatial prediction are considered: the first one is based on the nearest neighbors weighted median which directly considers concentrations while the second one is based on kriging increments, instead of more traditional pseudo-innovations, with two different variants. The proposed methods have been applied to the hourly PM10 concentrations of the monitoring network in Normandy (France). As a result, the actual situation is to compute all the three indicators coming from the different variants and to alert the technicians, which take the final decision to keep or not the data, when it is possible. In off-line mode, the provided methods are currently used as new tools for data validation. In a fully automatic mode for the construction of maps during the night especially, the problem of the decision rule is under consideration. An adaptive solution could be to follow the best method up to the current hour for each station.

Acknowledgments. This work comes from a scientific collaboration between Air Normand (see the website http://www.airnormand.fr) from the applied side and Orsay University and INSA Rouen from the academic side. We would like to thank Véronique Delmas, from Air Normand, for providing the problem, the data as well as for supporting the statistical study.

References

[1] Barnett V. (2004). *Environmental Statistics: Methods and Applications*. Wiley Series in Probability and Statistics.

[2] Ben-Gal I. (2005). Outlier detection, In: Maimon O. and Rockach L. (Eds.) Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers, Kluwer Academic Publishers.

[3] Cerioli A., Riani M. (1999). The Ordering of Spatial Data and the Detection of Multiple Outliers, Journal of Computational and Graphical Statistics, 8(2), 239-258.

[4] Chandola V., Banerjee A., Kumar V. (2009). Anomaly Detection: A Survey, ACM Computing Surveys, Vol. 41, 3, Article 15.

[5] Cressie, N. (1993). Statistics for Spatial Data, Revised Edition, John Wiley and Sons, New York.

[6] Filzmoser, P., Ruiz-Gazen, A., Thomas-Agnan, C. (2012). Identification of local multivariate outliers. Statistical Papers, 1-19.

[7] de Fouquet, C., Malherbe, L., Ung, A. (2011). Geostatistical analysis of the temporal variability of ozone concentrations. Comparison between CHIMERE model and surface observations. Atmospheric Environment, 45(20), 3434-3446.

[8] Genuer, R., Poggi, J.-M., Tuleau C. (2010). Variable selection using Random Forests. Pattern Recognition Letters, 31(14), p. 2225-2236.

[9] Haslett, J., Bradley, R., Craig, P., Unwin, A., Wills, G. (1991). Dynamic graphics for exploring spatial data with applications to locating global and local anomalies. The American Statistician 45(3), 234-242.

[10] Kou. Y., Lu C.-T., Chen D. (2006). Spatial Weighted Outlier Detection. In Proceedings of SIAM Conference on Data Mining.

[11] Laurent, T., A. Ruiz-Gazen, Thomas-Agnan, C. (2012). GeoXp: An R Package for Exploratory Spatial Data Analysis, Journal of Statistical Software, Vol. 47, Issue 2.

[12] Li Y., Nitinawarat S., Veeravalli V. V. (2013). Universal Outlier Detection. arXiv:1302.4776 (February 2013)

[13] Lu C.-T., Chen D., Kou. Y. (2003). Algorithms for Spatial Outlier Detection. Proceedings of the Third IEEE International Conference on Data Mining (ICDM03).

[14] Pebesma, E. J. (2004). Multivariable geostatistics in S: the gstat package. Computers & Geosciences, 30(7), 683-691.

[15] Planchon V. (2005). Traitement des valeurs aberrantes : concepts actuels et tendances générales. Biotechnol. Agron. Soc. Environ., 9(1), 19-34.

[16] R Core Team (2013). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/

[17] Ribeiro Jr, P. J., Diggle, P. J. (2001). geoR: A package for geostatistical analysis. R news, 1(2), 14-18.

[18] Rousseeuw, P. J., Leroy, A. M. (2005). Robust regression and outlier detection (Vol. 589). Wiley.

[19] Shekhar, S., Lu, C. T., Zhang, P. (2003). A unified approach to detecting spatial outliers. GeoInformatica, 7(2), 139-166.