Détection spatiale de données aberrantes. Application à la surveillance de la qualité de l'air.

Michel Bobbia[1] & Michel Misiti[2] & Yves Misiti[2] & Jean-Michel Poggi[3] & Bruno 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 et Univ. Paris Descartes, jean-michel.poggi@parisdescartes.fr
[4] Normandie Université, INSA Rouen, bruno.portier@insa-rouen.fr

Résumé. Nous considérons les mesures horaires des concentrations de poussières PM10 de 22 stations de surveillance situées dans les régions Basse-Normandie, Haute-Normandie et dans les régions voisines. Toutes les stations de surveillance considérées sont des stations urbaines de fond ou des stations rurales. Le travail porte sur la détection statistique des concentrations de PM10 horaires d'un point de vue spatial. Deux façons de gérer la détection spatiale des valeurs aberrantes sont considérées. La première basée sur la médiane pondérée des voisins les plus proches, considère directement les concentrations tandis que la seconde est basée sur le krigeage d'incréments de concentrations, au lieu des plus traditionnelles pseudo-innovations, et utilise une approche de type jackknife. Les deux méthodes sont appliquées au réseau de surveillance des PM10 en Normandie et sont pleinement mises en œuvre dans le processus de contrôle de la qualité des mesures. Quelques résultats numériques sont fournis sur des données récentes de janvier 2013 à janvier 2014 pour illustrer et comparer les deux méthodes.

Mots-clés. Qualité de l'air, krigeage, plus proches voisins, PM10, détection spatiale des données aberrantes

Abstract. We consider hourly PM10 measurements from 22 monitoring stations located in Basse-Normandie and Haute-Normandie regions and also in the neighboring regions. All considered monitoring stations are either urban background stations or rural ones. The paper focused on the statistical detection of the hourly PM10 concentrations from a spatial point of view. Two ways to handle spatial outlier detection are considered: the first one is based on the nearest neighbours weighted median which directly consider concentrations while the second one is based on kriging increments, instead of more traditional pseudo-innovations, and uses a jackknife type approach. The two methods are applied to the PM10 monitoring network in Normandie (France) and are fully implemented in the Measurements Quality Control process. Some numerical results are provided on recent data from January 2013 to January 2014 to illustrate and compare the two methods.

Keywords. Air quality, kriging, nearest neighbors, PM10, spatial outlier detection

1. Introduction

En France, la surveillance de la qualité de l'air est prise en charge dans chaque région par une association officielle, agréée par le Ministère de l'Ecologie. En Normandie, composée de deux régions administratives, Air Normand, basée à Rouen (Haute-Normandie), et Air COM, basée à Caen (Basse-Normandie), surveillent la qualité de l'air. En plus de leurs fonctions de contrôle, leur rôle est d'informer le public sur l'air qu'il respire. Ainsi, pour accomplir leurs missions, Air Normand et Air COM mesurent la qualité de l'air avec des analyseurs disséminés dans la région, et publient ces mesures, en particulier par l'utilisation du web pour informer le public sur son exposition à la pollution de l'air.
Air Normand et Air COM travaillent en étroite collaboration pour publier leurs mesures horaires sur un site Internet commun. Les mesures sont interpolées dans l'espace pour produire des cartes de qualité de l'air, cartes qui sont mises à jour toutes les heures sur le site web des associations.

Le problème est alors d'éviter des cartes erronées causées par des mesures aberrantes. La qualité de l'air est construite sur la base des mesures des concentrations et des cartes fournies par les sorties de modèles numériques physico-chimiques. La carte affichée est obtenue par correction, en utilisant des procédés d'assimilation, des simulations fournies par des modèles numériques par les concentrations mesurées dans les stations de réseau (voir, par exemple, de Fouquet et al. 2011). Les erreurs de mesures non diagnostiquées peuvent donc sérieusement affecter la qualité de la reconstruction spatiale des concentrations, comme on peut le voir sur la Figure 1 montrant une telle carte, qui contient une donnée aberrante à Lisieux.

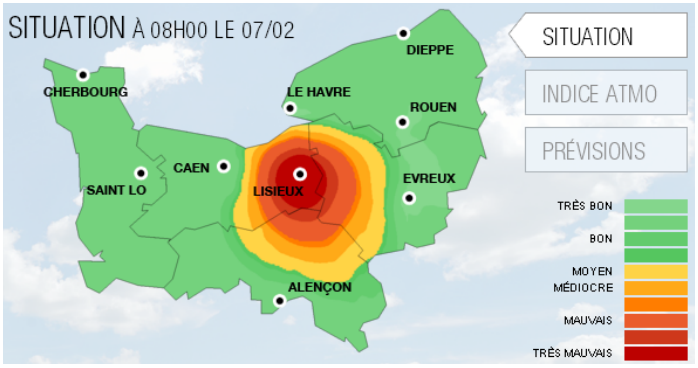

Figure 1: Une carte qui contient une donnée aberrante à Lisieux.

Ainsi, le but de ce travail est double : construire des outils de détection en temps réel des valeurs aberrantes générant une carte incorrecte, et contribuer à la validation d'une mesure dans chaque station de mesure du réseau.
Nous considérons dans ce travail effectué avec Air Normand, le problème de la détection spatiale des valeurs aberrantes dans le contexte des poussières et plus précisément celui des PM10, qui est le polluant le plus important en Normandie.

On peut esquisser un panorama rapide de l'abondante littérature sur les valeurs aberrantes. Par exemple, nous pouvons d'abord citer le livre classique Barnett 2004, qui contient un chapitre spécialement dédié à ce sujet, ainsi que quelques papiers de synthèse Planchon 2005, Ben-Gal 2005 ou plus récemment Chandola et al. 2009. Toutefois ces références concernent

principalement les valeurs aberrantes univariées ou multivariées, mais ne sont pas spécifiquement dédiées à la nature spatiale des données. Haslett et al. 1991 ainsi que Laurent et al. 2006 explorent des outils d'analyse de données spatiales et en déduisent des procédures de détection de valeurs aberrantes (voir par exemple Filzmoser et al. 2012). Dans le cas des données spatiales, une distinction classique est faite entre une valeur atypique "globale", qui consiste à raisonner à partir du comportement de la majorité des données et une valeur atypique "locale" qui consiste à raisonner à partir du comportement des observations voisines géographiquement. Ainsi, on peut définir quatre classes d'observations : typique, atypique globale uniquement (détecté par des outils standards), atypique locale seulement et la dernière : atypique locale et globale. Dans cet article, nous nous intéressons à la détection des observations atypiques locales. Notons qu'une observation atypique locale est souvent définie comme une observation qui diffère des mesures en des sites proches. Il est donc implicitement supposé que les données présentent une autocorrélation spatiale positive. Bien sûr, il est important de vérifier si cela est réaliste dans chaque application spécifique. Les travaux de Cerioli et al. 1999, basés sur le krigeage sont consacrés à la détection de valeurs aberrantes multiples tandis que Kou et al. 2006 et Lu et al. 2003, élaborent des moyens simples, intuitifs et robustes pour détecter les valeurs aberrantes spatiales. Signalons enfin la très récente référence Li 2013.

Dans ce papier, l'idée de base de l'algorithme de détection est fondée sur la comparaison de la concentration mesurée avec une certaine prédiction spatiale robuste, avec et sans le site suivant une approche classique de type jackknife. Nous considérons deux méthodes de détection des valeurs aberrantes différant par la méthode de prédiction spatiale. La première, inspirée de Kou et al., 2006 et de Lu et al. 2003, est une voie spatiale non stationnaire comparant directement la concentration en un site donné avec la médiane pondérée des concentrations de ses voisins par rapport à un système de voisinage pré-spécifié. La seconde est basée sur le krigeage des innovations, à savoir la différence entre les observations actuelles et un ensemble de référence d'observations passées ou de sorties de modèles numériques, et non pas directement sur les concentrations.

2. Détection spatiale des valeurs aberrantes

Le principe général de l'algorithme de détection est basé sur la comparaison de la concentration mesurée en un site à une prédiction spatiale robuste, supposée exempte de valeurs aberrantes. Cette stratégie est classique, voir par exemple Rousseeuw et al. (2005).
Plus précisément, on considère les résidus définis par :

$$R_t(s_k) = Z_t(s_k) - \hat{Z}_t^{(-k)}(s_k), 1 \leq k \leq K$$

où $Z_t(s_k)$ est la concentration mesurée à l'instant t et à la station de mesure s_k, et $\hat{Z}_t^{(-k)}(s_k)$ est une certaine prédiction spatiale basée sur les mesures des stations voisines en excluant bien sûr s_k, suivant ainsi une approche de type jackknife.

Ainsi, l'idée de base est, en partant de la suite des résidus $\left(R_t(s_k)\right)_{1\leq k\leq K}$, de comparer chaque résidu $R_t(s_k)$ à deux valeurs $L(s_k)$, limite inférieure généralement négative, et $U(s_k)$, limite supérieure généralement positive. Ensuite, trois cas peuvent se présenter :

- Tous les résidus sont à l'intérieur des limites, alors aucune valeur aberrante n'est détectée ;
- Un résidu unique est en dehors de ses limites, alors la mesure associée est aberrante ;
- Plusieurs résidus sont en dehors de leurs limites associées. Dans ce cas, la mesure correspondant au résidu le plus éloigné de sa limite en valeur absolue, est considérée comme une valeur aberrante ; la concentration correspondante est alors considérée comme donnée manquante et l'ensemble du processus est réitéré.

Une question naturelle est de savoir comment choisir les limites de rejet $L(s_k)$ et $U(s_k)$. Tout d'abord, remarquons que, comme pour toute procédure de test, les limites de rejet sont basées sur l'hypothèse nulle, correspondant donc ici aux données validées. Deuxièmement, puisque nous considérons directement les concentrations, les résidus ne sont pas homogènes dans l'espace. Par conséquent, pour chaque station, nous définissons les seuils comme les quantiles extrêmes de la distribution empirique des résidus sur l'historique des données validées, sans procéder à une quelconque normalisation spatiale. Ces limites dépendent alors de la station. Bien entendu, la qualité et la taille de l'historique des données validées peuvent affecter les performances de la procédure de détection. Dans la suite, nous prenons pour $L(s_k)$ et $U(s_k)$ les quantiles à 0,25% et 99,75% de la distribution empirique de $\left(R_t(s_k)\right)_{t\in T}$, T étant un sous-ensemble de données validées dans ce travail.

Bien entendu, la procédure implique des prédictions spatiales locales calculées à partir des mesures des stations voisines. Examinons les deux approches expérimentées.

3. Médiane pondérée des plus proches voisins

La première méthode de prédiction (conduisant à l'algorithme noté NN dans la suite) est basée sur la médiane pondérée des concentrations des plus proches voisins. La prédiction est définie par :

$$\hat{Z}_t^{(med)}(s_k) = \text{médiane pondérée}\left(Z_t(s_m) \text{ pour } s_m \in V(s_k)\right)$$

Bien entendu, étant donné que la corrélation spatiale des concentrations n'est pas homogène, les ensembles de voisins ainsi que les pondérations associées dépendent du site s considéré. Deux méthodes différentes ont été utilisées pour les fixer, la première n'est pas entièrement automatique, mais une référence doit être construite et validée par des experts. La seconde est une variante entièrement automatique basée sur les indices d'importance des variables des forêts aléatoires (voir Genuer et al. 2010).
La première méthode considère un sous-ensemble des données validées et procède en trois étapes pour déterminer les voisins de s_k. Un modèle linéaire expliquant la concentration $Z(s_k)$ par les autres concentrations $(Z(s_m)\ ;\ s_m \neq s_k)$ est construit, utilisant une pénalité lasso pour pousser les petits coefficients vers zéro. Puis, une élimination pas-à-pas descendante (au seuil de 5%) est réalisée sur le modèle linéaire sans terme constant. Enfin, une double interprétation experte (pollution, géographique) conduit à une liste définitive des voisins et des poids associés.
Notons que cette procédure de définition des voisinages implique que le nombre de voisins dépend du site et conduit à des voisinages non nécessairement symétriques, i.e. $s_m \in V(s_k)$ n'implique pas que $s_k \in V(s_m)$. Ceci est cohérent avec l'absence de stationnarité spatiale du phénomène de pollution.

4. Krigeage des incréments

La seconde méthode de prédiction est basée sur le krigeage. L'idée est d'utiliser pour prédire $Z_t(s_k)$, la sortie d'un modèle de krigeage basé uniquement sur les autres sites $(Z(s_m)\ ;\ s_m \neq s_k)$. Rappelons que pour appliquer le krigeage, la stationnarité est une propriété clé que nous avons besoin de vérifier a priori. Cela peut être contourné en travaillant sur des pseudo-innovations au lieu des concentrations, c'est à dire la différence entre les observations réelles et des quantités reflétant la valeur moyenne des concentrations locales.

Nous avons considéré deux variantes conduisant à deux algorithmes différents :

- la première considère les différences avec les mesures à l'instant $t-1$ si celles-ci sont toutes disponibles et dans le cas contraire avec les résultats d'un krigeage basé sur les mesures disponibles à l'instant $t-1$. L'algorithme sera noté KK.
- la seconde, conduisant à l'algorithme noté KM, considère les différences avec les valeurs prédites par le modèle numérique ESMERALDA le plus récent.

Pour être un peu plus explicite, en supposant que les données $\left(Z_{t-1}(s_k)\right)_{1\leq k\leq K}$ à l'instant $t-1$ sont disponibles et sans valeurs aberrantes, alors pour prédire $Z_t(s_k)$, on considère le sous-ensemble jackknife de mesures $(D(s_m) = Z_t(s_m) - Z_{t-1}(s_m)\ ;\ s_m \neq s_k)$, on estime le variogramme exponentiel et on construit les poids optimaux de krigeage pour prédire $D(s_k)$ par $\widehat{D}(s_k)$. Puis, en revenant aux concentrations, on en déduit une estimation $\hat{Z}_t(s_m) = \widehat{D}(s_k) + Z_{t-1}(s_k)$. Cette prédiction peut alors être comparée avec la mesure réelle $Z_t(s_k)$ en calculant le résidu.

Notons que l'utilisation du krigeage des pseudo-innovations élimine la majeure partie du champ moyen et une partie importante des aspects non stationnaires. Il reste néanmoins une certaine adaptation au site en raison du fait que pour chaque site une procédure de krigeage différente est nécessaire pour obtenir le résidu associé au site et par suite les diagnostics des valeurs aberrantes. Ce type d'idées ouvre des perspectives : plus généralement, nous pourrions aussi penser à produire des cartes par bootstrap (au lieu du jackknife) ou par des techniques de simulation conditionnelles pour obtenir des cartes de valeurs aberrantes. Mais dans notre cas, et à cette échelle, avec un trop petit nombre de sites, nous avons préféré limiter notre stratégie au jackknife.

5. Applications

Les deux méthodes sont appliquées au réseau de surveillance des PM10 en Normandie et sont pleinement mises en œuvre dans le processus de contrôle de la qualité des mesures. Quelques résultats numériques seront fournis sur des données récentes de janvier 2013 à janvier 2014 pour illustrer, comparer les deux méthodes et en montrer la complémentarité.

Bibliographie

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

[2] Ben-Gal I., 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*, 2005

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

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

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

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

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

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

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

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

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

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

[13] Lu C.-T., Chen D., Kou. Y. Algorithms for Spatial Outlier Detection. *Proceedings of the Third IEEE International Conference on Data Mining* (ICDM'03), 2003

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

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