Atmospheric Pollution Research 6 (2015) 476-483

Atmospheric Pollution Research

www.atmospolres.com

Spatial outlier detection in the PM_{10} monitoring network of Normandy (France)

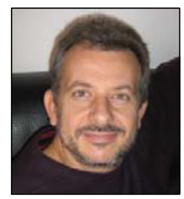
Michel Bobbia [1], Michel Misiti [2], Yves Misiti [2], Jean-Michel Poggi [2,3], Bruno Portier [4]

[1] *Air Normand, 3, place de la pomme d'or, 76 000 Rouen, France*
[2] *Université d'Orsay, Lab. de Mathématiques, 91405 Orsay, France*
[3] *Université Paris Descartes, Paris, France*
[4] *Normandie Université, INSA de Rouen, BP 08 – 76800 Saint–Etienne du Rouvray, France*

ABSTRACT

We consider hourly PM_{10} 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 focuses on the statistical detection of outliers of the hourly PM_{10} concentrations from a spatial point of view. The general strategy uses a jackknife type approach and is based on the comparison of the actual measurement with some robust spatial prediction. Two spatial predictions are considered: the first one is based on the median of the concentrations of the closest neighboring stations which directly consider weighted concentrations while the second one is based on kriging increments, instead of more traditional pseudo–innovations. The two methods are applied to the PM_{10} 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*



Corresponding Author:
Jean-Michel Poggi
☎ : *+33-0 1 69 15 57 44*
🖷 : *+33-0 1 69 15 72 34*
✉ : jean-michel.poggi@math.u-psud.fr

Article History:
Received: 16 July 2014
Revised: 01 December 2014
Accepted: 02 December 2014

doi: 10.5094/APR.2015.053

1. Introduction

In France, air quality is monitored in each region by an official association. In Normandy (consisting of two regions), Air Normand, based in Rouen and Air C.O.M. (Air COM for short), based in Caen, monitor air quality. In addition to these primary functions, their role is also to inform the population regarding air quality. Thus, to fulfill their missions, Air Normand and Air COM measure air quality with automatic analyzers scattered throughout the region, and make these measurements publicly available, mainly through the website to inform the public on exposure to air pollution. Indeed, Air Normand and Air COM work closely together to publish their measurements on a common website (www.airnormand.fr). In particular, measurements are spatially interpolated to produce maps of air quality, also available from the website. More precisely, Air Normand provides a map of air quality on the Normandy region updated every hour. The maps of air quality for two pollutants (O_3 and PM_{10}) are obtained combining hourly measurements of concentrations and the maps provided by the numerical model outputs. Each pollutant is mapped by correcting the numerical model outputs by the measurements provided by the monitoring stations, using assimilation methods (Grancher et al., 2005; de Fouquet et al., 2011). Thus undiagnosed measurement errors could seriously affect the quality of the spatial reconstruction of concentrations leading to erroneous maps.

The aim of this work is to provide tools for outlier detection in the spatial sense, which could help in the validation of measurement of each specific location of the monitoring network. More precisely, we consider in this paper the problem of spatial outlier detection in the context of particulate matter and especially PM_{10}, which is the more crucial pollutant in Normandy, but this is a general pattern and it can be applied to many other contexts.

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 (Ben–Gal, 2005; Planchon, 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. (2012), use analytic tools to explore spatial data and to deduce some outlier detection procedures as in Filzmoser et al. (2014). In the case of spatial data, a classical distinction is to be made 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 (detected using standard tools), local atypical only and the last one, local and global atypical. In this paper, we are interested in detecting local atypical observations. We can notice that a local atypical observation is often defined as an observation that differs from the closest observations, so it is implicitly assumed that the data exhibit a positive spatial autocorrelation. Of course it is important to check that this autocorrelation is realistic in each specific application. Some references particularly favor the detection of spatial anomalous observations. Cerioli and Riani (1999), define a procedure based on kriging schemes and dedicated to multiple outliers while Shekhar et al. (2003), Lu et al. (2003) and Kou et al. (2006), develop some simple, intuitive and robust ways to detect spatial outliers. Let us finally mention the very recent paper of Li et al. (2013) about outlier hypothesis testing studied in a universal setting.

The basic idea of the detection algorithm we propose consists in comparing the measured concentration to some spatial prediction, following a classical jackknife type approach (i.e. leaving out the observation at the considered location from the dataset used to calculate the estimate, see Efron, 1982). The decision rule is then based on thresholds coming from the distributions of prediction residuals along time. We consider two methods to perform the detection of spatial outliers depending on the prediction method. The first one, inspired by Lu et al. (2003) and by Kou et al. (2006), follows a non–stationary spatial way by directly comparing the concentration of a given site to the median of the weighted concentrations of its neighbors with respect to a pre–specified neighborhood system. The second one is based on kriging increments, namely the difference between the current observations and a reference set of past observations or numerical model outputs, and not directly on concentrations.

The paper is organized as follows. Section 2 presents the PM_{10} monitoring network, the Measurements Quality Control process and the PM_{10} data. Finally some basics about kriging methodology are recalled. Section 3 describes the general principle of the outlier detection procedure and the two methods for spatial outlier detection. Section 4 first presents the results of detection procedures of spatial outliers on some recent database and then proposes a discussion. Finally, some concluding remarks are collected in Section 5.

2. Materials and Kriging Methodology

Normandy (over 3.3 million people in 2008) is located in Northwestern France, along the south coast of the English Channel and at the Northwest of Paris. Normandy is composed of two regions: Basse–Normandie and Haute–Normandie. Haute–Normandie is heavily industrialized with two large urban areas, Rouen and Le Havre including more than 490 000 and 250 000 inhabitants respectively, while Basse–Normandie is more agricultural with only one urban area of significant size, Caen with more than 400 000 inhabitants.

2.1. PM_{10} monitoring network

We have a set of PM_{10} hourly mean concentrations coming from 22 monitoring stations located in Basse–Normandie and Haute–Normandie regions and also in the neighboring regions (see in Figure 1 the location of PM_{10} monitoring stations and main cities of Normandy). All considered monitoring stations are either urban background stations or rural ones.

The spatial outlier detection only concerns PM_{10} measurements coming from stations located in Normandy, namely AIL (near Dieppe), HRI, NEI (Le Havre), JUS, PQV, POS (Rouen), EVT (Evreux) for the Air Normand network (Haute–Normandie) and CHD (Cherbourg), SLO (Saint–Lo), CAE, IFS (Caen), LIS (Lisieux), ALE (Alençon) and MRA for the Air COM network (Basse–Normandie). The last eight stations ARR, BOV, DRE, FRE, LUC, PRU, MAZ and REN are stations of other monitoring agencies of the neighboring regions. Stations AIL, POS, MRA and ARR are rural stations while the others are urban background ones. Monitoring devices are mainly TEOM (Tapered Element Oscillating Microbalance) except for stations CHD, IFS, LIS and MRA where Beta gauges are used.

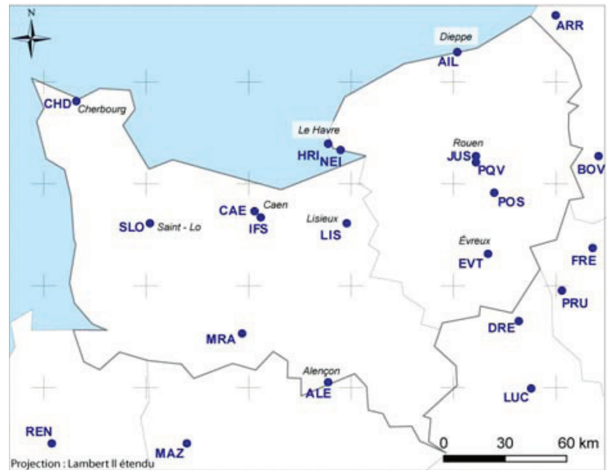


Figure 1. *Location of PM_{10} monitoring stations and main cities.*

2.2. Measurements quality control process

Air Normand and Air COM make measurements freely available and update it at a rate depending on the communication media that is going to be used (i.e., paper report, web). Of course, the most common one is the Internet: measurements are collected at various monitoring stations, stored in the database and automatically published every hour. To ensure a high quality of service, Air Normand and Air COM have developed several procedures for data validation.

First, the maintenance of the analyzers is the primary task of any air quality monitoring association. Technicians regularly calibrate the analyzers, according to the manufacturers' recommendations and the references prescribed by CEN standards (see www.cen.eu).

Then technicians check measurements twice a day (morning and evening). It is a first validation level performed on a strictly technical basis. At this stage, the physical state of measuring equipment is controlled, leading to three decisions: validation, invalidation or sometimes correction of the data in the database. The website is then updated accordingly. A second level of validation is performed daily by a single expert on a different temporal scale examining the sequence of measurements on each site. Finally, an environmental validation of the measurements is performed every month during a meeting of experts. At this stage, the measurement network is considered as a whole: rather than reviewing site by site, from a metrological point of view, it is examined spatially.

Of course, the notion of outlier is not necessary the same as invalidated data. For example, in case of a smooth drift, an expert will invalidate the data even if it is not an outlier. Conversely an outlier could be a validated data in case of local pollution episode. In summary, it is not possible to perform the complete validation of measurements in real–time. The scope of this work is to provide some automatic ways to identify possible spatial outliers using statistical methods. These possible outliers will then be invalidated or not.

2.3. Data

For each station, we have hourly or bi–hourly average PM_{10} concentrations. Each value can be a NaN (Not A Number) when the