SCREENING TOOLS FOR DATA QUALITY AND OUTLIER DETECTION APPLIED TO THE AIRBASE AMBIENT AIR POLLUTION DATABASE

Oliver Kracht[1], Hannes Isaak Reuter[2] and Michel Gerboles[1]

[1]European Commission - Joint Research Centre, Institute for Environment and Sustainability, Climate Change and Air Quality Unit, Via E. Fermi, 21027 Ispra (VA), Italy
[2]gisxperts gbr, Eichenweg 42, 06849 Dessau, Germany

Abstract: Systematic collection of long term meso- to large-scale datasets of ambient air quality provides an indispensible means for air pollution monitoring. However, the quality of these monitoring data depends on the chosen method of measurements and the QA/QC procedures applied. We present a consolidated screening tool for the automatic detection of outliers in large data volume air quality monitoring records. The method is based on an adaption of the existing "Smooth Spatial Attribute Method", which considers both attribute values and spatio-temporal relationships. The method is demonstrated with application examples of warnings on abnormal values in time series of PM_{10} datasets reported in the European air quality database AirBase.

Key words: *air pollution, air quality, monitoring, spatial statistics, spatio-temporal outlier, quality control, harmonization*

INTRODUCTION

In order to provide scientifically sound information for regulatory purposes and environmental impact assessment, long term meso- to large-scale datasets of ambient air quality provide an important means for monitoring population exposure to harmful pollutants, climate change trends, and for model calibration, evaluation and validation as well. However, the collection of high quality datasets with suitable spatial coverage for air pollution management and decision support poses many challenges. It is thus critical to establish expedient tools for the efficient assessment and data quality control of air pollution measurements in large scale national and international monitoring networks.

The European Environmental Agency systematically receives measurements of ambient air pollution at more than 6000 monitoring stations from over 30 countries (European air quality database AirBase). The quality of these data depends on the chosen method of measurements and quality assurance and quality control (QA/QC) procedures applied by each country. In the context of our ongoing research at the Joint Research Centre (JRC), the objective is to develop geostatistical tools able to automatically extract quality indicators within these large datasets, e.g. trends of data, identification of outliers, estimation of analytical uncertainty, and the extent of the area of representativeness of individual monitoring stations. In particular, we are currently developing novel methodologies to automatically screen the AirBase records for internal consistency and to detect spatio-temporal outliers nested in the data.

METHODOLOGY

We implemented a spatio-temporal toolset for screening abnormal datapoints which considers both attribute values and spatial relationships. This procedure for outlier detection was designed based on already existing literature. Specifically, we adapted the "Smooth Spatial Attribute Method" that was first developed for the identification of outlier values in networks of traffic sensors. Such methodological background and applications have for example been presented by Lu, CH.-T., D. Chen and Y. Kou (2003) and by Shekhar, S., CH.-T. Lu and P. Zhang (2003).

Concept for spatio-temporal outlier screening

Our algorithms are based on the definition of a neighbourhood for each air quality measurement, corresponding to a spatio-temporal domain limited by time (e.g., +/- 2 days) and distance (e.g., +/- 1 spherical degrees) around the individual locations of ambient air monitoring stations. The objective of the method is that within such a given spatio-temporal domain, in which the attribute values of neighbours have a relationship due to the emission, transport and reaction of air pollutants, abnormal values can be detected by extreme values of their attributes compared to the attribute values of their neighbours. This comparison basically requires a spatio-temporal smoothing, i.e. a specific rule by which data points are averaged within a neighbourhood. The calculation of such reference basis has the effect of a low pass

filter, meaning that high frequencies of the signal are removed from the data while preserving low frequencies. In this context, the choice of an appropriate kernel smoother function (e.g., nearest neighbour smoother, weighted kernel average smoother, etc.) is of particular importance.

Data analysis scheme

For brevity, only a simplified conceptualisation of the data reduction scheme can be described here. More details and a step by step description of the computational methods can be obtained from Kracht, O., H. I. Reuter and M. Gerboles (2013).

In the following, we use x to denote a spatial object which attributes are (i) its location, and (ii) the corresponding pollutant measurement value. For each monitoring station value x, the set of spatio-temporal neighbours is identified. After a log-transformation of non-Gaussian data, we compute (i) the weighted average of measurement values within a neighbourhood of x according to equation (1), and (ii) the differences Sx between the pollutant concentration (non-spatial attribute value f(x)) at x and the corresponding average value of its neighbours according to equation (2).

$$\overline{x}_n = \frac{\sum_{i=1}^{n} w_i \; x_{n,i}}{\sum_{i=1}^{n} w_i} \tag{1}$$

$$Sx = f\left(x\right) - \overline{f\left(x_n\right)} \tag{2}$$

Suitable methods for the calculation of the weighting factors w_i are the squared normalized Euclidean distance or the inverse squared Mahalanobis distance.

Within each neighbourhood, the Sx values of the central station are normalised to center data at 0 with a standard deviation of 1 using equation (3). In this equation $\overline{Sx}$ and s_{Sx} are the weighted average and the weighted standard deviation of all Sx_i attribute values within the neighbourhood of x. As a result, a time series z_i of z-values is obtained for each monitoring station.

$$z = \frac{Sx - \overline{Sx_n}}{s_{Sx_n}} \tag{3}$$

Finally, the test statistics for detecting a spatio-temporal outlier, given in equation (4), searches for z_i values exceeding a limit value of θ plus / minus a predefined threshold of 1.96.

$$\theta - 1.96 < z_i < \theta + 1.96 \tag{4}$$

For the exercise runs presented here, we have chosen to obtain the reference basis θ by applying a Kolmogorov-Zurbenko (KZ) filter to the individual z_i time series of each station. In this way, we allow for long-term trends and natural spatial variations in the air pollutions properties not to be detected as a measure of outlierness. However, in order to more strictly consider the full spectrum of possible spatial and temporal outlierness, one would prefer θ to be set to zero in this screening step.

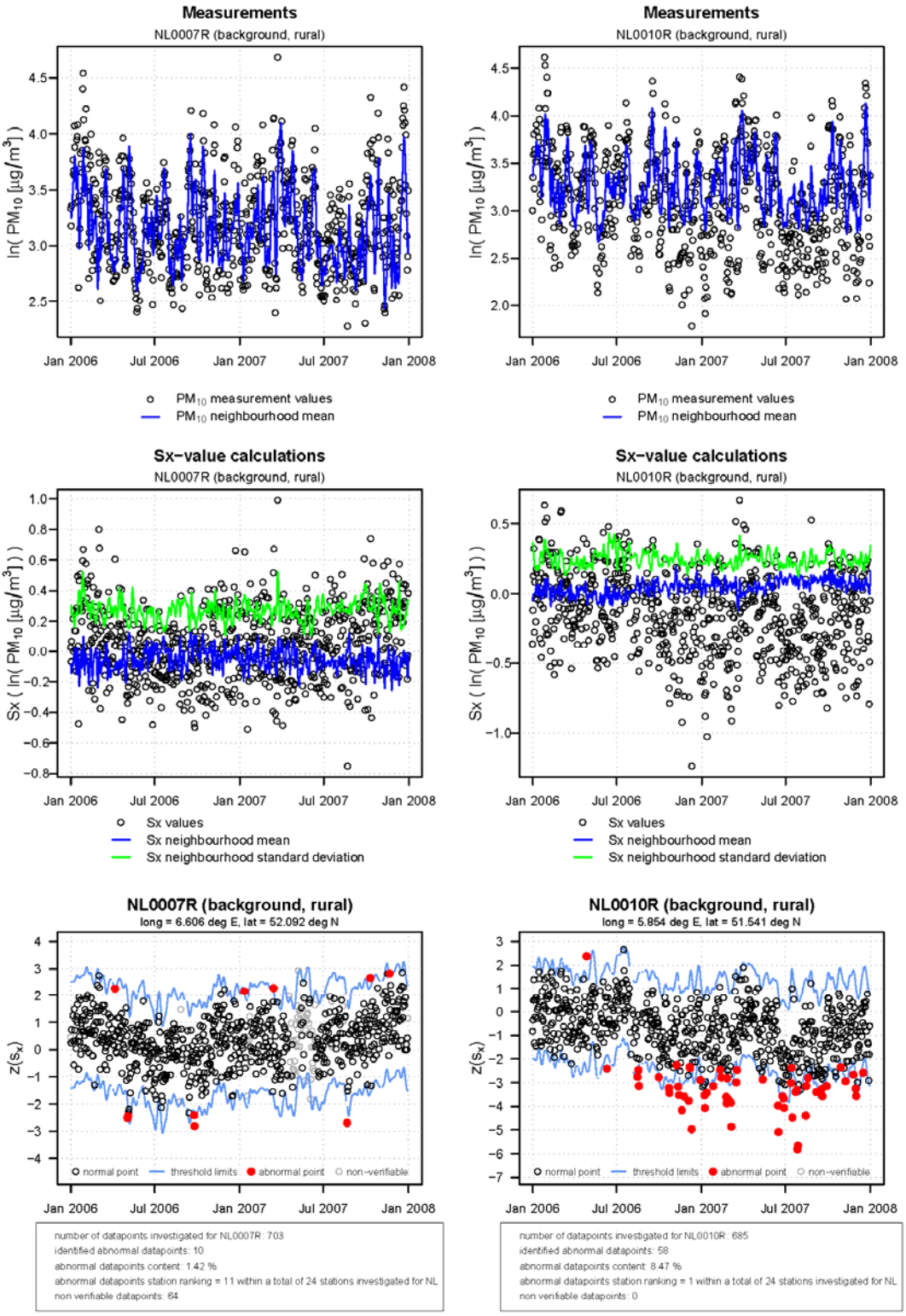

Figure 1. Data analysis scheme and summary of final outcomes for the AirBase stations NL0007R (panels in left column) and NL0010R (panels in right column). Note that measurement values have been log-transformed.

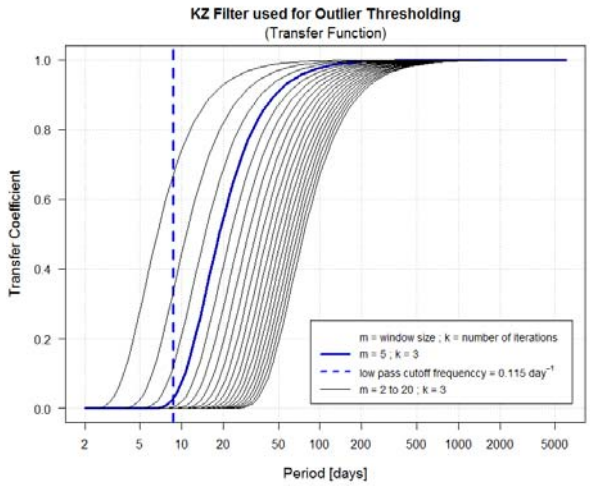

Figure 2. Properties of the Kolmogorov-Zurbenko (KZ) filter used in the computation of the continuous $\theta(z_i)$ timeseries (reference for the Sx-outlier tresholding). In our examples a filter with m = 5 days window size and k = 3 iterations has been used (solid blue line). Such filter effectively removes signal components with a periodicity of approximately less than 8.7 days (dashed blue line: low pass cutoff frequency of 0.115 day^{-1}). Filters with differing window extent (m = 2 to 20 days) are shown for comparison (black lines).

For safety of the conclusions, some quality limitations are applied to this last outlier thresholding step: First, in the case of $|z_i|$ exceeding a value of 1.96, the z_i is not taken into account for the calculation KZ-filtered series. Second, a minimum number of datapoints is required within a window to calculate a KZ-filtered value. Third, outliers are only flagged when the reference point neighbourhood contains a minimum number of data points (threshold set to 20). In consequence, the basis for outlier thresholding is unavailable for such cases, and datapoints are labelled as "non verifiable".

Figure 1 illustrates the sequence of different data processing steps applied to two examples selected from AirBase (monitoring stations NL0007R and NL0010R). In these examples, a KZ filter with a window size of m = 5 days and k = 3 iterations has been applied to the z_i time series before the final outlier thresholding step (figure 2).

APPLICATION STUDIES

The application of a first prototyped version of this screening method was tested by a comprehensive simulation and data analysis study based on the 2006 and 2007 AirBase records of daily PM_{10} values for a selection of 8 countries (AT, CZ, DE, ED, FR, GB, IT and NL) (Kracht, O., H. I. Reuter and M. Gerboles, 2013). Taking into account the high spatial variability of PM_{10} concentrations around industrial and traffic stations, it was decided to apply the screening method to the sole stations of background type, but for all area types (urban, suburban and rural).

In this first numerical exercise, the datasets covered a range of different country sizes which comprised between 16'135 and 165'543 records each. From these, the content of abnormal datapoints identified ranged between 2% and 4.1% of the records within the individual country datasets. However, not all records did fulfil the predefined selection criteria for being included into the computations. Furthermore, the design of the outlier test scheme also led to some mathematical dead ends restricting the verifiability of certain individual records. In consequence between 9% and 40% of the records per individual country had to be flagged as non-verifiable. In consequence, those non-verifiable datapoints had to be excluded from the screening for irregularities for safety of the conclusions.

Closer analyses and revision of this first implementation revealed that an important computational bottleneck was connected to the use of a simple moving average with window size 5 days for the calculation of the outlier thresholding reference $\theta(z_i)$. This was especially the case when calculations

stopped because of several z_i values exceeding the stipulated quality limitations, and in consequence prevented the continuous computation of the 5 days moving average of θ.

In an updated version of the method - which is presented here - we are now using a Kolmogorov-Zurbenko filter for the computation the $\theta(z_i)$ time series. In this way, it was possible to significantly reduce the number of non-verifiable datapoints. The number of non-verifiable data now ranges between 1% and 26% of the records per individual country. However, as the outcome of this improved method, the content of identified abnormal datapoints is now ranging between 1.9% and 14.2% of the records within the individual country datasets. This indicates that a large proportion of abnormal records was hidden within the non-verifiables.

For the remaining records, the possible non verifiability is no more a primarily consequence of mathematical dead ends, but likely more often due to the design of monitoring networks and the lack of stations within neighbourhoods. Indeed it can be observed that in regions with low spatial station density the number of neighbouring stations is often insufficient for this type of analysis.

We would like to emphasize that the reported figures about abnormal datapoints contents are unquestionably dependent on the parameter values chosen in the screening method, and that those are still going to be fine-tuned further. The choice of different functional parameters can affect the outcome of the abnormal value screening, and tuning this parameter in direction of “strict” values (e. g., low thresholds) naturally causes a larger number of abnormal values (and unverified records) in the evaluation. In this sense, an absolute definition for outlying stations is not feasible, but does indeed depend on the intended objectives of the use of the method.

OUTLOOK

The implemented method could be of interest as a data quality screening system when countries report their measurements to the European Environment Agency. Beyond this, it could also provide a simple solution to investigate the accuracy of station classification in AirBase. Long term research and development objectives can be linked with investigations like (i) application of the screening tool for checking of data quality in the framework of near to real time data reporting, and (ii) evaluating the perspectives and feasibilities to develop the screening tool into an online-application for operational use and accessibility by individual station managers

REFERENCES

Lu, CH.-T., D. Chen and Y. Kou, 2003: Detecting Spatial Outliers with Multiple Attributes. ICTAI, 15th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'03), IEEE 2003.

Shekhar, S., CH.-T. Lu and P. Zhang, 2003: A Unified Approach to Detecting Spatial Outliers. *GeoInformatica*, **7(2)**, 139-166.

Kracht, O., H. I. Reuter and M. Gerboles, 2013: A Tool for the Spatio-Temporal Screening of AirBase Datasets for Abnormal Values. *JRC Technical Reports* **78437**, EUR 25787 EN, ISSN 1831-9424 (online), ISBN 978-92-79-28286-7 (PDF), DOI 10.2788/81552, 209 pp.