Environmental Modelling & Software 101 (2018) 34–50

Contents lists available at ScienceDirect

Environmental Modelling & Software

journal homepage: www.elsevier.com/locate/envsoft

Automated data scanning for dense networks of low-cost air quality instruments: Detection and differentiation of instrumental error and local to regional scale environmental abnormalities

Maryam Alavi-Shoshtari [a, *], Jennifer Ann Salmond [b], Ciprian Doru Giurcăneanu [c], Georgia Miskell [a], Lena Weissert [a], David Edward Williams [a, **]

[a] School of Chemical Sciences, Private Bag 92019, Auckland Mail Centre, Auckland 1142, New Zealand
[b] School of Environment, Private Bag 92019, Auckland Mail Centre, Auckland 1142, New Zealand
[c] Department of Statistics, Private Bag 92019, Auckland Mail Centre, Auckland 1142, New Zealand

ARTICLE INFO

Article history:
Received 23 December 2016
Received in revised form
2 December 2017
Accepted 2 December 2017

Keywords:
Change-point detection
Linear multi-regression
Sensor networks
Data reliability

ABSTRACT

Recent improvements in low-cost air quality instrumentation make deployment of dense networks of sensors possible. However, the shear volume of data from these networks means that traditional methods for data quality control and data analysis are no longer viable. We propose a real-time data scanning routine that detects local and regional variability within the data sets. This can be used to differentiate errors resulting from instrument malfunction or calibration drifts from natural (environmentally driven) regional changes in ambient concentrations. Our case study considered hourly-averaged ozone data from Texas and from two networks in Vancouver. We used 7 and 28 days of data for the algorithm initialisation with simulated and real instrumental changes. The algorithm output can be used as part of a limited resource maintenance schedule for sensor networks, and to improve understanding of air quality processes and their relation to environmental and public health data.

1. Introduction

One of the primary objectives for deployment of air quality monitoring networks is to provide an accurate assessment of the risk high concentrations of pollutants pose to public health and the environment. Understanding air quality variation is essential to meet the objective, especially within an urban environment where the temporal and spatial heterogeneity of emission patterns and the complexity of the urban surface result in strong gradients in vertical and horizontal pollutant concentrations (Colvile et al., 2001). The primary focus of air quality managers is on conditions when pollutant concentrations exceed or are about to exceed standard thresholds and to identify locations where exceedances are more frequent (Adams et al., 2001; Gulliver & Briggs, 2005). However, raw data, even from regulatory monitors, cannot be trusted for analysis since all instruments are vulnerable to drifts and errors. This is why network managers plan regular on-site inspections, routine on-site calibration and the data are not released before all quality control procedures have been completed (Fiebrich et al., 2006; Stolarski & Frith, 2006).

The number of monitoring sites in air quality networks has therefore traditionally been limited by the cost of deploying and managing the instruments used. Sparse network data are useful for forecasting, exploring long-term variation in pollutants concentrations at regional scales and examining the effectiveness of practiced air quality standards in reducing the emissions (Jaffe & Widger, 2012; Patton et al., 2015; Vingarzan, 2004). However, there is increasing recognition that the variations within an urban environment, particularly within urban canyons, are typically so large that combinations of sparse network data and air quality models do not give reliable information (Fiebrich et al., 2006; Genikhovich et al., 2002; Rojas, 2014). Therefore, there has been an increasing demand to establish high-density monitoring networks which provide good quality data at dense spatial and temporal resolutions. The demand has been addressed by recent

* Corresponding author.
** Corresponding author.
E-mail addresses: m.alavi@auckland.ac.nz (M. Alavi-Shoshtari), j.salmond@auckland.ac.nz (J.A. Salmond), c.giurcaneanu@auckland.ac.nz (C.D. Giurcăneanu), georgia.miskell@gmail.com (G. Miskell), lwei999@aucklanduni.ac.nz (L. Weissert), david.williams@auckland.ac.nz (D.E. Williams).

https://doi.org/10.1016/j.envsoft.2017.12.002
1364-8152/

developments in technology for low-cost monitoring instruments (Bart et al., 2014; Jiao et al., 2016; Williams et al., 2013). As networks grow in size and density, traditional data quality control procedures, especially regular on-site inspections, rapidly become infeasible due to prohibitive costs and limited resources. In addition, semi-analytical techniques (such as visualisation or scanning the data for outliers) are not effective air quality management tools since they cannot detect subtle changes in such large datasets in real-time (or near real-time) especially when the data reliability cannot be assumed.

This paper addresses the need for an automated real-time or near real-time change detection routine for large air quality data sets, that identifies whether the change was due to *true* variation in pollutant concentrations at local or regional scales, or caused by calibration drift of the monitoring instrument. The proposed algorithm is a tool for network and air quality managers to make decisions on instrument maintenance and/or for issuing air quality advisories. Unlike semi-analytical routines of data which scan for outliers, our routine considers both positive and negative variations in concentration as part of the change-detection routine. Positive local or regional variations may trigger alarms for high pollutant concentrations whilst negative local or regional variations show the impact of temporary or long-term pollutant sinks as well as the effectiveness of practiced air quality standards in emissions reduction. Positive and negative instrumental variations are equally important in the evaluation of the instrument performance. Therefore, detecting and highlighting episodes of variation and identification of the cause have been the dominant consideration in the design of this algorithm.

The algorithm consists of four subroutines: i) a proxy model to define a measure or norm for data variation at a high temporal resolution from a pre-existing data set, with minimal requirements for the initialisation data; ii) change (or decision) criteria for data deviation (positive or negative) from the norm, iii) classification of the detected change as local/regional or instrumental change, and iv) an adaptive approach for updating and rolling forward the measure and proxy model. For the algorithm to be practical and generally applicable, the subroutines have been developed to make minimal assumptions about data availability, since periods of missing data impose a serious practical constraint.

For performance evaluation, we selected hourly-averaged ozone data from two regulatory networks with different topography: the TCEQ network in Houston, Texas (TCEQ, 2016) and the Metro Vancouver (MV) network in Vancouver, British Columbia (MetroVancouver, 2008). We selected 10 sites of the TCEQ network and 14 sites of the MV networks as case studies due to their data availability and diversity in land-use and geographical locations. The input data for the algorithm were ozone measurements only and, because we specified minimal assumptions about data availability, no other pollutant or meteorology data were included in the measure. We chose 7 and 28 days (about 672 data measurements for each monitoring site) as an acceptably short data set for the initial definition of the norm. We simulated calibration slope and offset errors for instrumental changes. The algorithm performance was evaluated using data for the five years from January 2010 to December 2014. These data sets are not from high-density networks simply because such data do not at present exist with the controls needed to check the performance of the algorithms. We have used data from regulatory networks and imposed artificial variations since this gives something clear that we can unambiguously detect as an instrumental change. We have also used those data from a low-cost network where the devices were co-located with regulatory instruments since again this provided a necessary check on the performance of the algorithm. We considered 6 sites of a low-cost gas-sensitive semiconducting oxide (GSS) network. This network was installed in the Lower Fraser Valley in 2012 (Bart et al., 2014) and we assessed the algorithm performance in comparison with the regulatory MV network. The algorithm was successful in detection of local and regional changes that had been advised over this period by air quality authorities. The simulated instrumental changes were also successfully detected with a clear distinction from local and regional changes. The results confirmed that the proposed algorithm was a simple but powerful tool to detect changes and identify causes and can be applied for network and air quality management purposes.

The rest of this paper is organised as follows; in Section 2 we review the previous literature related to the purpose of the algorithm. The algorithm formulation is given in Section 3, with the experimental results of the regulatory and low-cost networks in Section 4. Discussion about the algorithm's parameters are given in Section 5 and concluding remarks in Section 6.

2. Literature review

The tools and techniques of network management and air quality management have traditionally been different for sparse networks. Regular on-site inspection with data visualisation and scanning have been recommended traditionally for data quality control as the number of monitoring sites was assumed to be limited (Jiao et al., 2016; Le & Zidek, 2006; Lewis et al., 2016). In our previous works (Alavi-Shoshtari et al., 2013; Miskell et al., 2015) we addressed the need for automated routines for data quality control as the network grew in size and density. The data validation technique proposed by (Miskell et al., 2015) compared the large-sized data sets of air quality networks in high temporal resolution with some *proxies*. The proxies could be measurements from some reference instruments or derived by simple physical or statistical models. The proxies were not predictors (otherwise, the network would have unnecessarily extrapolated the data), however, the statistical characteristics of the difference were expected to remain stable under the no-change assumption. Statistically significant change in the difference between the network data and the proxy was indicative of a potential malfunction. The changes could also be the result of instability in the proxies or change in the instruments' local or regional conditions. This approach was successful in detection of subtle variation in the low-cost instruments' calibration settings in the field and was practical due to its minimal assumptions about data availability for initialisation and low computational cost (Miskell et al., 2015). The procedure did not classify the cause of changes and was sensitive to the choice of proxies. We also proposed a theoretical based decision approach to sensor network data quality control (Alavi-Shoshtari et al., 2013), where data received from every monitoring site were compared with an individual proxy using simple linear regression models. The approach aimed to control the network average maintenance cost by highlighting the instruments that were more likely to malfunction. The approach could detect small simulated calibration drifts provided that the nodes of the network were sufficiently correlated to one another but could not differentiate between instrument malfunction and natural variability in the data and required large pre-existing data sets for initialisation.

Statistical methods have been extensively used for network management purposes, however, they assume that the reliability of the initial data is confirmed prior to the analysis. Assuming there was no instability or drift in the instruments, calibration of heterogeneous networks has been implemented by the Geostatistical Dynamical Calibration Model (GDC) (Fasso et al., 2007; Sahu & Nicolis, 2009) or blind calibration (Balzano et al., 2008). Both approaches are more suitable for instrument measurement correction in the long term rather than quick detection of subtle malfunctions.