Environmental Research 160 (2018) 183–194

Contents lists available at ScienceDirect

Environmental Research

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

Enhanced data validation strategy of air quality monitoring network

Mohamed-Faouzi Harkat[a], Majdi Mansouri[b,*], Mohamed Nounou[a], Hazem Nounou[b]

[a] Chemical Engineering Program, Texas A & M University at Qatar, Doha, Qatar
[b] Electrical and Computer Engineering Program, Texas A & M University at QATAR, Doha, Qatar

ARTICLE INFO

Keywords:
Data validation
Air quality monitoring network
Exponentially weighted moving average
Generalized likelihood ratio test
Midpoint-radii
Principal component analysis

ABSTRACT

Quick validation and detection of faults in measured air quality data is a crucial step towards achieving the objectives of air quality networks. Therefore, the objectives of this paper are threefold: (i) to develop a modeling technique that can be used to predict the normal behavior of air quality variables and help provide accurate reference for monitoring purposes; (ii) to develop fault detection method that can effectively and quickly detect any anomalies in measured air quality data. For this purpose, a new fault detection method that is based on the combination of generalized likelihood ratio test (GLRT) and exponentially weighted moving average (EWMA) will be developed. GLRT is a well-known statistical fault detection method that relies on maximizing the detection probability for a given false alarm rate. In this paper, we propose to develop GLRT-based EWMA fault detection method that will be able to detect the changes in the values of certain air quality variables; (iii) to develop fault isolation and identification method that allows defining the fault source(s) in order to properly apply appropriate corrective actions. In this paper, reconstruction approach that is based on Midpoint-Radii Principal Component Analysis (MRPCA) model will be developed to handle the types of data and models associated with air quality monitoring networks. All air quality modeling, fault detection, fault isolation and reconstruction methods developed in this paper will be validated using real air quality data (such as particulate matter, ozone, nitrogen and carbon oxides measurement).

1. Introduction

Maintaining high air quality is a major environmental concern that has a profound impact on human health and the ecosystem. Various industrial effluents, human activities, and meteorological factors contribute to the pollution of air by pollutants, such as carbon oxides, nitrogen oxides, ozone, and particulate matter. Air quality monitoring networks are usually used to monitor the quality of air, not only to make sure that air quality standards are maintained, but also to allow taking any necessary preventive or corrective measures to minimize the effect of possible undesirable changes in some of these pollutants. Proper data validation of air quality networks is crucial to achieve their intended purpose. Therefore, the objective of this paper is to develop a general framework technique that aims at enhancing the data validation of air quality networks by developing:

- modeling technique that can accurately predict the behavior of air quality monitoring networks and any changes in pollution and/or meteorological conditions using different types of air quality data,
- monitoring technique that can quickly detect sensor faults or serious anomalies in air quality data,
- fault isolation method that can identify the root cause(s) of the detected fault(s), and
- fault estimation and data correction methods that allow providing meaningful information about the detected fault(s) that can ultimately be shared with the public.

Modeling and monitoring of air quality networks are crucial to ensure safety and protection of humans and the environment. In general, monitoring approaches (Venkatasubramanian et al., 2003a, 2003b) can be classified as: model-based or data-driven approaches. Model-based monitoring approaches utilize predictions of process models to make decisions regarding the existence or absence of faults (Kinnaert, 2003; Nyberg and Nyberg, 1999). Hence, the effectiveness of such approaches is greatly influenced by the quality of the process models. In the case where the difference between the model prediction and process measurement is relatively small, this indicates that the process is operating normally and no fault exists. However, when such a difference is relatively large, this is an indication that a fault has occurred (Kinnaert, 2003; Nyberg and Nyberg, 1999). Several model-based monitoring approaches have been developed, such as the parity space approaches (Staroswiecki, 2001; Ding and Frank, 1990; Patton

* Corresponding author.
E-mail address: majdi.mansouri@qatar.tamu.edu (M. Mansouri).

http://dx.doi.org/10.1016/j.envres.2017.09.023
Received 15 March 2017; Received in revised form 19 September 2017; Accepted 20 September 2017

and Chen, 1991; Chow and Willsky, 1984), observer-based approaches (Clark et al., 1975; Patton et al., 1989; Xu, 2002), and interval approaches (Adrot, 2000; Adrot et al., 2002; Benothman et al., 2007). In many situations, however, many key biological variables are not measured online and thus need to be estimated for monitoring purposes. Several researchers have worked on developing estimation algorithms to estimate some of such key variables (Mansouri et al., 2014b, 2012, 2014a).

Data-based approaches, on the other hand, assume that process measurements, when the process operates in normal (fault-free) conditions, are available (Venkatasubramanian et al., 2003b). Such fault-free process measurements are often used to construct empirical models that can be used for process monitoring. Several data-based monitoring approaches are available in the literature, such as latent variable regression (LVR) approaches such as partial least square (PLS) regression, principal component analysis (PCA), support vector machine (SVM) approaches (Dehestani et al., 2011), canonical variate analysis (CVA) approaches (Chaing et al., 2001; Venkatasubramanian et al., 2003b), pattern recognition approaches (Mohammadi and Asgary, 2005), as well as approaches that are based on fuzzy systems (Dexter and Benouarets, 1996) and neural networks (Subbaraj and Kannapiran, 2010). Monitoring approaches that are based on LVR models have been extensively used in practice to monitor various applications, such as air quality monitoring (Harkat et al., 2006), agriculture (Magyar and Oros, 2012; Mansouri et al., 2016a), water treatment (George et al., 2009; Tharrault, 2008), pharmacology (Nascimento and Martins, 2012), and health (Belasco et al., 2012). In a previous works (Mansouri et al., 2016b, 2016c; Botre et al., 2016; Sheriff et al., 2017), we have developed PCA, PLS, kernel PCA (kPCA) and kernel PLS (kPLS)-based generalized likelihood ratio test (GLRT) fault detection techniques, in which PCA, PLS, kPCA and kPLS have been used as a modeling framework for fault detection and the faults are detected using the GLRT chart.

Data-based approaches for modeling and monitoring of air quality networks, and various real applications, depend greatly on the quality of data used. Such data is usually imprecise due to uncertainties induced by measurement errors or specific experimental conditions. One way to deal with such uncertainties, which are common in modeling and monitoring of air quality data, is to represent the data as intervals rather than single values. Modeling and analysis using interval data have been considered by several researchers. Various interval PCA (IPCA) methods have been developed and used in process monitoring. The centers PCA (CPCA) (Cazes et al., 1997) and the vertices PCA (VPCA) (Douzal-Chouakria, 1998) were among the first IPCA methods to be developed. The CPCA method uses the centers matrix of the input interval data set to compute the principal components. Thus, it ignores the variations within each interval. This can be useful in cases where the interval is narrow, but can be a limitation if there exist large variations in the data. On the other hand, the VPCA method calculates the principal components using the vertices matrix, which is the matrix of all possible distinct classical observations extrapolated from the interval data set. This has three limitations. First, it may negate the inherent benefit of using IPCA, which is to aggregate data sets and minimize computational complexity. Second, it assumes that classical observations extracted from the interval data are independent, which is not necessarily the case. Finally, it neglects some of the variances covered by the interval structure of the data set.

To deal with the shortcomings of the VPCA method, the authors in (Lauro and Palumbo, 2000) presented the symbolic object and range-transformation IPCA methods. The symbolic object method introduces an additional Boolean transformation matrix, the purpose of which is to remove any interdependency between the vertices. However, this approach still suffers from the limitation of utilizing only the variances between the vertices, thereby ignoring some of the data set's internal variance. On the other hand, the range-transformation method transforms the matrix of ranges then calculates the principal components. As expected, this may also lead to misleading results since it ignores the centers of the intervals. In addition, selecting a method to calculate a suitable transformation matrix is vague and computationally expensive. In (Le-Rademacher, 2008), three techniques that take into consideration the internal structure of symbolic variables have been presented: i) a technique that extends the conventional PCA to an analysis of interval-valued measurements based on symbolic variance-covariance structure, ii) a technique that extends the PCA-based interval-valued data method to a PCA-based histogram-valued data, which uses histogram-valued measurements as a generalization of interval-valued measurements, and iii) a technique that constructs the likelihood functions for the symbolic data. The author of (Jie, 2008) has applied symbolic data analysis methods to extract the dynamic features of Copper futures market of Shanghai Futures Exchange (SHFE) between 2005 and 2006, where the author used three-way principal component analysis of interval data for this purpose. Later, another IPCA method called midpoint-radii PCA (MRPCA) was developed (Lauro and Palumbo, 2005; Palumbo and Lauro, 2003; Lauro et al., 2008), which treats the centers and radii of the input interval data set as two separate variables, then combines their respective PCA models to compute the interval principal components. MRPCA is the most popular interval multivariate statistical methods, able to tackle the issue of uncertainties on the models and one way to improve the data validation abilities. Thus, to achieve the modeling objective, MRPCA modeling technique will be developed to allow predicting the concentrations of various pollutants (such as carbon oxides, nitrogen oxides, ozone, and particulate matter concentrations) and to help understand the behavior of air quality networks.

To achieve the monitoring objective, on the other hand, statistical fault detection method that can effectively and quickly detect sensor faults or abnormal changes in key measured air quality variables will be developed. In this task, a new fault detection method will be developed, which combines the generalized likelihood ratio test (GLRT) and exponentially weighted moving average (EWMA). EWMA-GLRT is a filtered statistical hypothesis testing fault detection method that relies on maximizing the detection probability of faults for a particular false alarm rate. EWMA-GLRT statistic will aim at detecting changes in the mean of the residuals obtained from the air quality model. Changes in the mean of air quality models can be due to malfunctioning sensors (that get stuck to give biased measurements) or due to abnormal physical changes in the air quality conditions.

Important elements in the data validation of air quality motioning networks include the isolation and correction of the detected faults. Fault isolation refers to the identification of the root cause(s) of the detected fault(s), where data correction refers to the estimation of the actual value of the variable when a faulty measurement is obtained using a malfunctioning sensor. For fault isolation, EWMA-GLRT based technique would provide simultaneous detection and isolation. Fault isolation and correction can be achieved using a technique called ”Reconstruction method,” which is an iterative procedure that can be used to estimate the actual values of faulty measurements using an MRPCA model. For fault isolation, the measured variables are sequentially excluded and the fault detection index is computed without each excluded variable to identify the source of the fault. When the faulty variable is included in the data, the value of the fault detection statistic violates its threshold, but when the faulty variable is excluded, the fault detection index falls below its threshold, indicating that the excluded variable is the faulty one. The reconstruction method can be directly applied using single-valued data. Therefore, in this paper, an extension of the reconstruction method, that can be used in fault isolation and correction of interval-valued air quality data, will be developed. In summary, this paper will develop technique that aims at enhancing the data validation of air quality monitoring networks. The developed method will integrates modeling, fault detection, fault isolation, and data correction methods with improved performances. These techniques will be able to deal with interval-valued air quality data. The