Atmospheric Pollution Research xxx (2017) 1–9

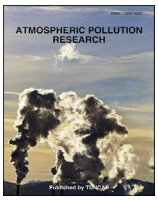
HOSTED BY

ELSEVIER

Contents lists available at ScienceDirect

Atmospheric Pollution Research

journal homepage: http://www.journals.elsevier.com/locate/apr



Control chart and Six sigma based algorithms for identification of outliers in experimental data, with an application to particulate matter PM_{10}

Martina Čampulová [a, b, *], Petr Veselík [a], Jaroslav Michálek [a]

[a] University of Defence, Faculty of Military Leadership, Department of Econometrics, Kounicova 65, 662 10 Brno, Czechia
[b] Mendel University in Brno, Faculty of Business and Economics, Department of Statistics and Operation Analysis, Zemědělská 1, 61300 Brno, Czechia

ARTICLE INFO

Article history:
Received 9 August 2016
Received in revised form
9 January 2017
Accepted 10 January 2017
Available online xxx

Keywords:
Outlier detection
Particulate matter
Kernel smoothing
Six sigma
Control charts

ABSTRACT

Outliers, which can have significant effects on further analysis and modelling, occur between continuously measured environmental data. Most methods for outlier detection depends on model or distribution of observed variable. However the distribution of environmental variables cannot be estimated quite often. This paper presents two procedures, which do not impose restrictions on the distribution of analysed variable, and which permit the intervals of the environmental observations, where the outliers occur, to be detected. The proposed procedures are based on smoothing original data and subsequent analysis of the residuals. The output of both methods is an interval of observations, where the residual process behaves substandard, and whose quality must be further manually assessed. Thus the value of the proposed methodology is that the number of observations for manual data control is reduced. Both methods are applied to problem of detection outliers in hourly PM_{10} measurements. However, the methodology is general and can be applied to different type of data whose quality control is required.

Contents

* Corresponding author. University of Defence, Faculty of Military Leadership, Department of Econometrics, Kounicova 65, 662 10 Brno, Czechia.
E-mail addresses: martina.campulova@unob.cz, martina.campulova@mendelu.cz (M. Čampulová), petr.veselik@unob.cz (P. Veselík), jaroslav.michalek@unob.cz (J. Michálek).
Peer review under responsibility of Turkish National Committee for Air Pollution Research and Control.

http://dx.doi.org/10.1016/j.apr.2017.01.004

Please cite this article in press as: Čampulová, M., et al., Control chart and Six sigma based algorithms for identification of outliers in experimental data, with an application to particulate matter PM_{10}, Atmospheric Pollution Research (2017), http://dx.doi.org/10.1016/j.apr.2017.01.004

1. Introduction

Outliers, the observations that appear inconsistent with the rest of the data set (Barnett and Lewis, 1978), occur sometimes in environmental measurements. The outliers might result from natural variability of analysed pollutant in the air, from erroneous measurements, unusual measurement conditions, or they may be caused by the presence of a new factor affecting the observed variable. Because outliers may significantly affect the results of other analyses, their detection and interpretation plays important role in the research of air pollution.

An overview of the methods for outlier detection on temporal data is given in Gupta et al. (2014), outlier detection techniques for time series are described for example in Burman and Otto (1988) or in Fox (1972). Recently, numerous methods, that permit the outliers in the environmental data to be detected and the quality of the data to be checked have been proposed (Kokalj et al., 2011; Bobbia et al., 2015; Shaadan et al., 2015; Dupuis and Field, 2004).

As most classical outlier detection techniques requires a priori knowledge of the distribution function or model of studied variable, it is problematic to apply them on data whose distribution is unknown. Observations from environmental areas, which may depend on accompanying variables, are example of data whose distribution can't be estimated due to the unknown dependence on accompanying variables.

The aim of this article is to propose two procedures, which can be used to automatic identification of segments in environmental data, where the outliers occur. The principle of both methods is to smooth the original data by using nonparametric regression with variable (local) bandwidth and subsequently detect the intervals, where the residual process behaves nonstandard due to the presence of outliers. The observations of original data corresponding to these intervals, which are found by using Six sigma methodology (Michálek, 2009; Montgomery, 2009) and control charts proposed by Shewhart (1931), need to be further manually investigated for the presence of outlier and invalid measurements.

The suggested methods are applied to identify outliers in hourly measurement of particulate matter PM_{10}. Particulate matter is released into the environment from natural sources (e. g. forest fires, volcanoes, dust storms, sea spray) as well as from anthropogenic sources (automotive transportation, industrial and agricultural activities, coal combustion, burning of waste and biomass, road dust etc.) (Keuken et al., 2013; Kim et al., 2015). Large number of epidemiological studies (Abrutzky et al., 2012; Franchini and Mannuci, 2007; Pope and Dockery, 2006; Restrepo et al., 2012; Russell and Brunekreef, 2009; Samek, 2016) confirm the existence of a linkage between changes in the concentration of particulate matter in the air and negative impacts on the human health, especially of people suffering from cardiovascular and respiratory diseases. Continuous monitoring of concentrations and composition of PM_{10} particles is essential for the prediction and evaluation of periods with high-concentration of PM_{10}. The presence of outliers in the data set can lead to misspecification in identification of emission sources of aerosols with possible high expenses for its amendment.

The article is structured as follows: In the following section 2 we describe the data and measuring stations. In section 3 the methodology is introduced. Particularly, the description of kernel regression used for data smoothing and approaches for detection of residual outliers is given. The proposed methods are summarised at the end of section 3. The application of presented procedures on problem of detection outliers in PM_{10} concentrations is given in section 4 and discussed in section 5. Finally the findings are concluded in section 6.

2. Data

The proposed methods are applied to detect outliers in hourly measurements of concentrations of atmospheric aerosol PM_{10}. PM_{10} mass concentrations were measured at two monitoring stations (namely Lany and Turany) in the city of Brno, the second largest city of the Czech Republic with population of about 400 000. The data were provided by Council of the City of Brno and by Czech Hydrometeorological Institute. The station Lany, which is situated on the southern edge of the Bohunice housing estate, is protected against effects of the traffic by two rows of houses and grown up vegetation, but motorway D1 leads approximately 400 m south. Station Turany is situated in the area of Turany airport and the territory in immediate surroundings of the station can be defined as an area without buildings and without residents. More detailed description of the data and measuring stations can be found in diploma thesis by Šmejdířová (2016).

3. Methodology

3.1. Data smoothing

Because the observed environmental variable depends on many different factors, which are quite often unknown, the original data are not stationary. To compensate the influence of unknown covariates the original data are smoothed by using kernel regression (Wand and Jones, 1995) and smoothing residuals are obtained.

Kernel regression is a nonparametric smoothing technique, which estimates the mean value of dependent variable at a given point as a weighted average of surrounding noisy observations. The weights are defined by the choice of kernel function and the amount of observations used for averaging is determined by a parameter called bandwidth. The choice of bandwidth, which can be determined globally or locally, is crucial part of the analysis. The chemical applications are usually based on global bandwidth (Henry et al., 2009). However several algorithms for local bandwidth, which produces better practical results, have been suggested in the literature (Fan and Gijbels, 1995; Fan et al., 1996; Cao-Abad and González-Manteiga, 1993; Brockmann et al., 1993). For smoothing the environmental data, the best results were obtained by using local plug-in algorithm (Herrmann, 1997).

Denoting Y_i the observed concentration in time instant t_i, $i = 1, \ldots, N$, where N denotes the number of observations, the residuals in time instants t_i are given by

$$X_i = Y_i - \widehat{m}(t_i), \tag{1}$$

where $\widehat{m}(t_i)$ is the kernel estimate of the unknown regression function $m(t_i)$ in time t_i . The estimate $\widehat{m}(t_i)$ is obtained by using Gasser-Müller estimator (Gasser and Müller, 1979) with local bandwidth (Herrmann, 1997). The residuals given by (1) are not influenced by unknown accompanying covariates.

3.2. Analysis of the residual process

Suppose that the obtained residuals $X_1, \ldots, X_N$ represent observations of a process X with mean μ and standard deviation σ. For the further analysis the unbiased estimates of μ and σ are needed.

To obtain these estimates the residuals are partitioned into k disjoint segments (subgroups) of size n ($N = kn$) and the behaviour of the process X is evaluated first on these segments. Thus the classic estimates of the considered characteristics (sample mean for μ and sample standard deviation for σ) on individual segments are found and the estimates of μ and σ for the whole process X are

Please cite this article in press as: Čampulová, M., et al., Control chart and Six sigma based algorithms for identification of outliers in experimental data, with an application to particulate matter PM_{10}, Atmospheric Pollution Research (2017), http://dx.doi.org/10.1016/j.apr.2017.01.004