Atmospheric Pollution Research 8 (2017) 700–708

HOSTED BY

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 22 January 2017

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

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

found based on the classic estimates from individual segments as follows:

Denoting $\bar{x}_1, \ldots, \bar{x}_k$ sample means and $s_1, \ldots, s_k$ sample standard deviations from preliminary subgroups, the parameter μ is estimated by sample mean $\bar{\bar{x}} = \sum_{i=1}^{k} \bar{x}_i$, and the estimate of parameter σ is obtained from mean $\bar{s} = k^{-1}\sum_{i=1}^{k} s_i$, which is corrected to unbiasedness by $\bar{s} = \mathrm{k}^{-1}\sum_{\mathrm{i}=1}^{\mathrm{k}} \mathrm{s_i}\bar{s} = \mathrm{k}^{-1}\sum_{\mathrm{i}=1}^{\mathrm{k}} \mathrm{s_i}$ normalising by correction factor $C_4(n) = \Gamma(n/2)\sqrt{2/(n-1)}/\Gamma(0.5(n-1))$ (SAS/QC, 1999), where $\Gamma(.)$ denotes gamma function. The resulting estimate of parameter σ is thus of the form

$$\widehat{\sigma}_s = \frac{\bar{s}}{C_4(n)}. \tag{2}$$

Another estimate of parameter σ can be obtained from sample ranges $R_1, \ldots, R_k$ in preliminary subgroups. Such estimate is based on mean $\bar{R} = k^{-1}\sum_{i=1}^{k} R_i$, and is given by

$$\widehat{\sigma}_R = \frac{\bar{R}}{d_2(n)}, \tag{3}$$

where $d_2(n) = \int_{-\infty}^{\infty} [1 - (1 - \Phi(x))^n - (\Phi(x))^n]dx$ (SAS/QC, 1999) and $\Phi(x)$ denotes cumulative distribution function of normal distribution.

For further analysis of the process X the control charts and Six sigma methodology described in next paragraphs are used. Both methods are frequently used in technical applications.

3.2.1. Control charts

Control charts represent a classical statistical method which permits statistical stability of a residual process X to be controlled. The principle is to graphically illustrate the changes of the residual mean and variability over time, which allows to identify time segments when the process X gets out of statistical control. Thus control charts visualise characteristics from disjoint subgroups, and two horizontal lines – upper and lower control limit UCL and LCL. These visualisation permits inessential instability in the behaviour of the residuals to be eliminated and the stability of the process X to be described. Note that the residual process X gets out of control due to changes in mean and variance of the studied variable. Typically, outlier observations are caused by unstable process. If the process remains statistically stable, all characteristics from disjoint subgroups fall within control limits, and it is presumable that the variability of the process results only from natural variation of the studied variable. Characteristic from disjoint subgroups falling outside the control limits indicates that the process is out of control and the cause of the instability needs to be investigated.

To detect segments of observations, where the outliers responsible for the change in mean and variance of the residuals occur, $\bar{x}$ chart and R chart constructed from characteristics from disjoint subgroups and graphically representing sample means $\bar{x}_1, \ldots, \bar{x}_k$ and sample ranges $R_1, \ldots, R_k$ is used, respectively. Analogous to R chart, s chart, which graphically represents sample standard deviations $s_1, \ldots, s_k$ from disjoint subgroups, can be used to describe the variability of the residual process.

The control limits are based on Chebyshev's inequality, which states that irrespective of the distribution type of the residuals maximally $(1/L^2)\%$ of sample means $\bar{x}_1, \ldots, \bar{x}_k$ falls outside the limits $\mu_{\bar{X}} \pm L\sigma_{\bar{X}}$, where $\mu_{\bar{X}} = \mu$ and $\sigma_{\bar{X}}$ denotes mean (expectation) and standard deviation of $\bar{x}_1, \ldots, \bar{x}_k$, respectively. Thus for usual choice $L = 3$ the Chebyshev's inequality states that at least $0,\overline{8}\%$ of sample mean from disjoint subgroups lies within three standard deviations of the mean μ, i.e. between the limits $\mu_{\bar{X}} \pm 3\sigma_{\bar{X}}$ (3-sigma rule). Similarly, for $L = 4$, $L = 5$, $L = 6$ the probability that sample mean from disjoint subgroup falls within the limits $\mu_{\bar{X}} \pm 4\sigma_{\bar{X}}$, $\mu_{\bar{X}} \pm 5\sigma_{\bar{X}}$, $\mu_{\bar{X}} \pm 6\sigma_{\bar{X}}$ is 93.75%, 96%, $97,\overline{2}\%$, respectively. However for a normal distribution the probability that the sample mean falls within 3-sigma limits is 99.7%, and the limits $\mu \pm 4\sigma$, $\mu \pm 5\sigma$, $\mu \pm 6\sigma$ contain 99.994%, 99.99994%, 99.999999999% of observations.

Therefore the control limits of $\bar{x}$ chart are defined as $\mu_{\bar{X}} \pm L\sigma_{\bar{X}}$. Because the parameter $\sigma_{\bar{X}}$ is related to parameter σ through $\sigma_{\bar{X}} = \sigma/\sqrt{n}$, the $\bar{x}$ chart control limits can be derived in the form (Wild and Seber, 2000)

$$UCL = \bar{\bar{x}} + L\frac{\widehat{\sigma}_s}{\sqrt{n}}, \quad LCL = \bar{\bar{x}} - L\frac{\widehat{\sigma}_s}{\sqrt{n}}. \tag{4}$$

Analogous, the control limits of R chart can be obtained in the form

$$UCL = \bar{R} + Ld_3(n)\widehat{\sigma}_R, \quad LCL = \max(0, \bar{R} - Ld_3(n)\widehat{\sigma}_R), \tag{5}$$

where $d_3(n)$ is correction factor (SAS/QC, 1999), and the control limits of s chart can be derived as

$$UCL = \bar{s} + L\sqrt{\widehat{\sigma}_s^2 - \bar{s}^2}, \quad LCL = \max\left(0, \bar{s} - L\sqrt{\widehat{\sigma}_s^2 - \bar{s}^2}\right). \tag{6}$$

We will further suppose that L is equal to at least 3.

3.2.2. Six sigma

Six sigma methodology is used to control statistical capability of the residual process X. The behaviour of the residuals is again evaluated first on disjoint segments of size n, and subsequently the capability index C_p is for each day constructed based on the estimates of characteristic σ in preliminary segments. The capability index, which is used to assess whether the process shows a significant deviation from the specified mean or contains outliers caused by the variability of individual observations, is estimated by using the relation

$$C_p = \frac{\mathrm{USL} - \mathrm{LSL}}{6\sigma}, \tag{7}$$

where USL (respectively LSL) is upper (respectively lower) specification limit usually defined by a specialised operator.

The classification is based on the assumption that the standard deviation σ of the residual process X satisfies $\sigma \leq (\mathrm{USL} - \mathrm{LSL})/8$, which corresponds to $C_p = 1.33$. If $C_p > 1.33$ the process is considered highly capable, in case that $1 < C_p \leq 1.33$ the process is said to be medium capable and $C_p \leq 1$ signifies incapable process. As is apparent from paragraph 3.2.1., the value of $C_p = 1.33$ means that the residual process X with normal distribution keeps between the LSL and USL with probability 99.994%, and for residual process with other not exactly specified distribution at least with probability 93.75%.

For further analysis the estimate of capability index C_p given by

$$\widehat{C}_p = \frac{\mathrm{USL} - \mathrm{LSL}}{\widehat{\sigma}_s}, \tag{8}$$

where $\widehat{\sigma}_s$ given by (2) will be used.

The $(1-\alpha)100\%$ confidence interval of parameter C_p, $\alpha\varepsilon(0,1)$, is under the assumption of normality given by Montgomery (2009)

$$(\widehat{C}_{p,L};\widehat{C}_{p,U}) = \left(\widehat{C}_p + \frac{\widehat{C}_p u_{\alpha/2}\sqrt{1-C_4^2(n)}}{C_4(n)\sqrt{k}};\widehat{C}_p + \frac{\widehat{C}_p u_{1-\alpha/2}\sqrt{1-C_4^2(n)}}{C_4(n)\sqrt{k}}\right). \quad (9)$$

One-sided $(1-\alpha)100\%$ confidence intervals of the parameter C_p with left-sided interval limit $\widehat{C}_{p,L^*}$ and right-sided interval limit $\widehat{C}_{p,U^*}$ are obtained analogous to (9) by substituting $u_{\alpha/2}$ and $u_{1-\alpha/2}$ by the value u_α and $u_{1-\alpha}$, respectively.

Since statistical incapability is indicative of outliers present in the residual process the point and interval estimates of capability index can be used to direct detection of segments, when the outlier residuals occur. Considering the significance level α, for testing the null hypothesis $C_p = 1.33$ against the alternative hypothesis $C_p < 1.33$ the right-sided confidence interval $(-\infty;\widehat{C}_{p,U^*})$ is used. Rejecting the null hypothesis indicates the presence of outliers in the corresponding time intervals.

3.3. Outlier detection methodology

As previously noted, the core idea of both methods is to smooth the original data by using kernel regression and subsequently analyse the residuals. Because $\bar{x}$ chart and R chart is used to evaluate the statistical capability of the residual process on disjoint segments of size n, the method based on control charts permits these segments, where the outliers occur, to be detected. The method based on Six sigma methodology will be used to assess the residuals on 1-day intervals and thus detect the days, where the residual process is incapable due to the presence of outliers.

To construct the control charts both the size of subgroups n and constant L must be specified.

Larger subgroups containing more than 10 values are under the assumption of normality reasonably effective to detect segments containing outlier residuals deviated less than 2σ from the mean value of the process. To detect segments containing residuals deviated more than 2σ from the mean value of the process smaller values of n ($n = 3, 4, 5$) are suitable.

From (4), (5), and (6) is evident that higher values of parameter L, which is supposed to be at least 3, lead to wider limits of control charts.

In case that the upper and lower specification limits USL and LSL can not be specified by a specialised operator we suggest to determine them based on $\bar{x}$ chart control limits UCL and LCL. If the residuals are normally distributed USL and LSL set equal to UCL and LCL can be used. As Chebyshev's inequality suggests, if the residuals show deviations from the normality, USL and LSL set as appropriately chosen multiple of UCL and LCL, which ensures wider specification limits, are recommended. In case that the specification limits LSL and USL are based on $\bar{x}$ chart control limits (as we suggest) both smaller value of n and larger value of L lead to milder criterion for evaluation of statistical capability.

4. Results

In this section we present examples of the analysis of the data measured at two monitoring stations described in section 2. Particularly, for station Lany we focus on concentrations from October 2016, and for station Turany we analyse the measurements from July 2016.

For data smoothing kernel regression with Epanechnikov kernel and local bandwidth estimated by using plug-in algorithm (Herrmann, 1997) was used. The PM_{10} mass concentrations together with the estimates of regression functions are visualised in Fig. 1a) and Fig. 2a) for station Lany and station Turany, respectively. The smoothing residuals are shown in Fig. 1b) and Fig. 2b).

We can see that the estimate of regression function adapts to the data and the curvature of the regression line changes based on the

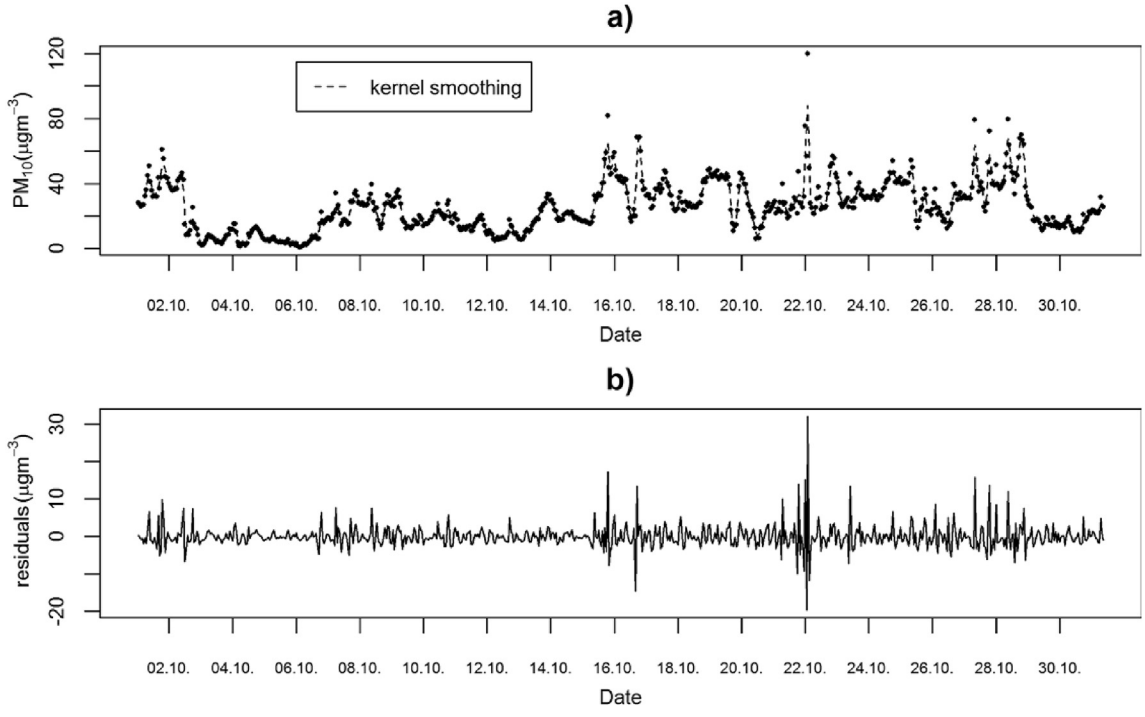

Fig. 1. Lany: a) PM_{10} concentrations and kernel smoothing regression estimate, b) regression residuals.

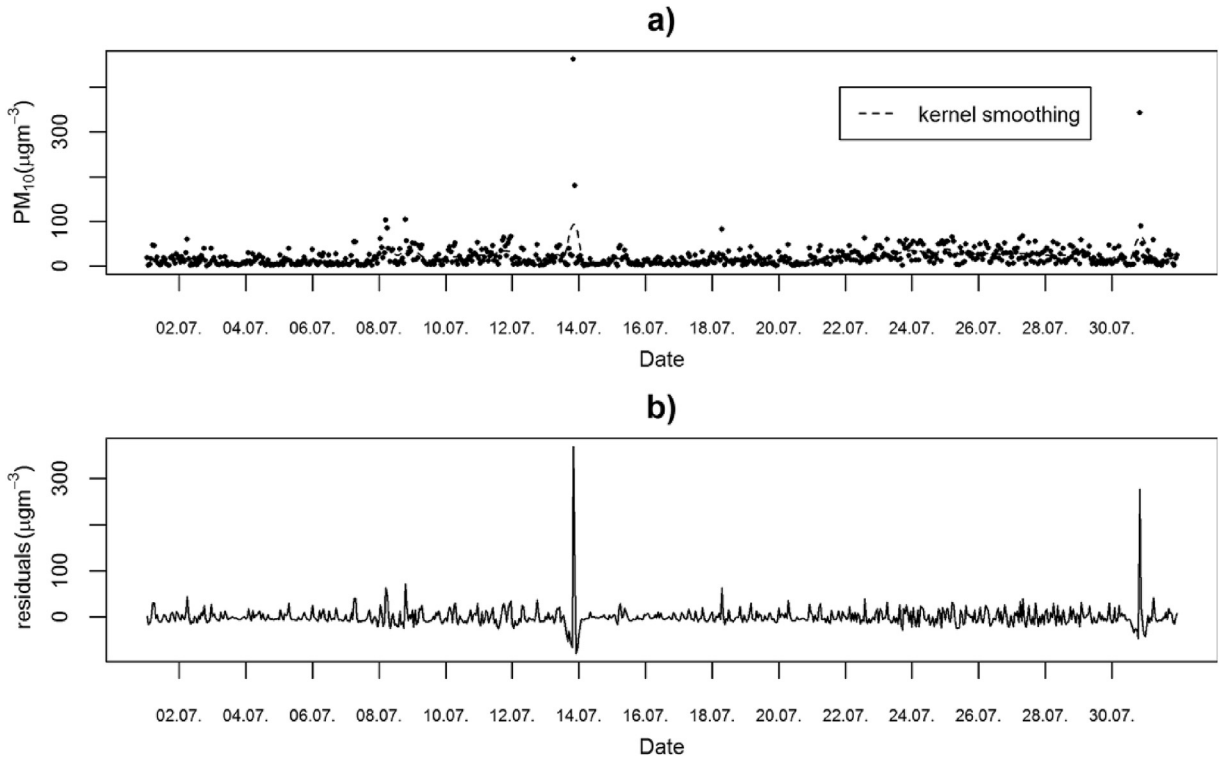

Fig. 2. Turany: a) PM_{10} concentrations and kernel smoothing regression estimate, b) regression residuals.

variability of the data. This phenomenon, typical for smoothing based on local bandwidth, is visible especially in local extremes of the regression function. The residuals are dispersed around the zero horizontal axis, and for station Lany relatively high values of the residuals are present mostly in the second half of the presented period. Considering the station Turany, residuals significantly deviated from the other residuals occur on 13th and 30th of July.

To evaluate the fit of the residuals to normality robust medcouple test (MC-LR) [Brys et al., 2008], which does not reject the normality when the outliers are present, was used. The null hypothesis that the distribution of the residuals is normal was rejected at 5% significance level for station Lany as well as for station Turany.

On this basis the control charts were constructed, and the capability of the residuals was evaluated on 1-day intervals. For this the point and right-sided interval estimate of capability index C_p was computed for each day of the considered period. The size of segments was set as $n = 3$, and the parameter L was set as $L = 5$ for both stations. The specification limits USL and LSL were set equal to 1.8 multiple of $\bar{x}$ chart control limits.

Fig. 3 shows control charts constructed from residuals corresponding to measurements at station Lany. Point and right-sided interval estimate (for $\alpha = 0.05$) of capability index, which is constant for each day, is displayed in a logarithmic scale as well. While all sample means fall within $\bar{x}$ chart control limits, both R chart and s chart indicate that the residual process gets out of statistical control several times. We can see that the results obtained by using R chart are comparable to those obtained by using s chart. Therefore for further analysis just the R chart is used.

From the right-sided confidence interval $(-\infty; \widehat{C}_{p,U^*})$ visualised in Fig. 3d) is clear, that the null hypothesis $C_p = 1.33$, which indicates the highly capable process, is rejected on 15th, 16th, 21st, 22nd and 27th of October, since the value $1.33 > \widehat{C}_{p,U^*}$ and thus the C_p value 1.33 is not included in $(-\infty; \widehat{C}_{p,U^*})$.

PM_{10} mass concentrations corresponding to sample means and sample ranges exceeding the limits of control charts are visualised in Fig. 4a) for station Lany and in Fig. 5a) for station Turany. Observations in days, when the null hypothesis $C_p = 1.33$ was rejected in favour of the alternative hypothesis $C_p < 1.33$, and which were thus detected by using the method based on Six sigma, are displayed in Fig. 4b) for station Lany and in Fig. 5b) for station Turany. The detailed graph of the residuals and PM_{10} mass concentrations from 8th of July and station Turany is shown in Fig. 6. Remind that both the segments and days detected by using control charts and Six sigma based method need to be further manually in detail investigated by a specialised operator for the presence of outlier and invalid measurements.

The identified segments and days for each station were compared with the results obtained by manual data control, which was performed by specialised operators from Council of the City of Brno and Czech Hydrometeorological Institute.

From Figs. 4 and 5 we can see that by using both proposed methods the segments containing outliers were identified on 15th, 16th, 21st, 22nd, 23rd and 27th of October for station Lany and on 8th, 13th and 30th of July for station Turany.

However, by using manual data control only four invalid measurements at night from 21st to 22nd of October were detected for station Lany. These measurements were detected as invalid due to their large, during night hours very unlikely, deviations from the rest of the values. However, the reason for the presence of such values in the data set could not be clarified, since no information about the unusual measurement conditions or measurement device failure in the night hours was recorded.

Considering the station Turany, two invalid values on 13th of July and two invalid observations on 30th of July were identified manually. The latter two correspond to data calibration, the cause for the presence of the first two invalid measurements is unknown.

All observations, which were manually identified as invalid for Lany station, were included in the segments and days detected by

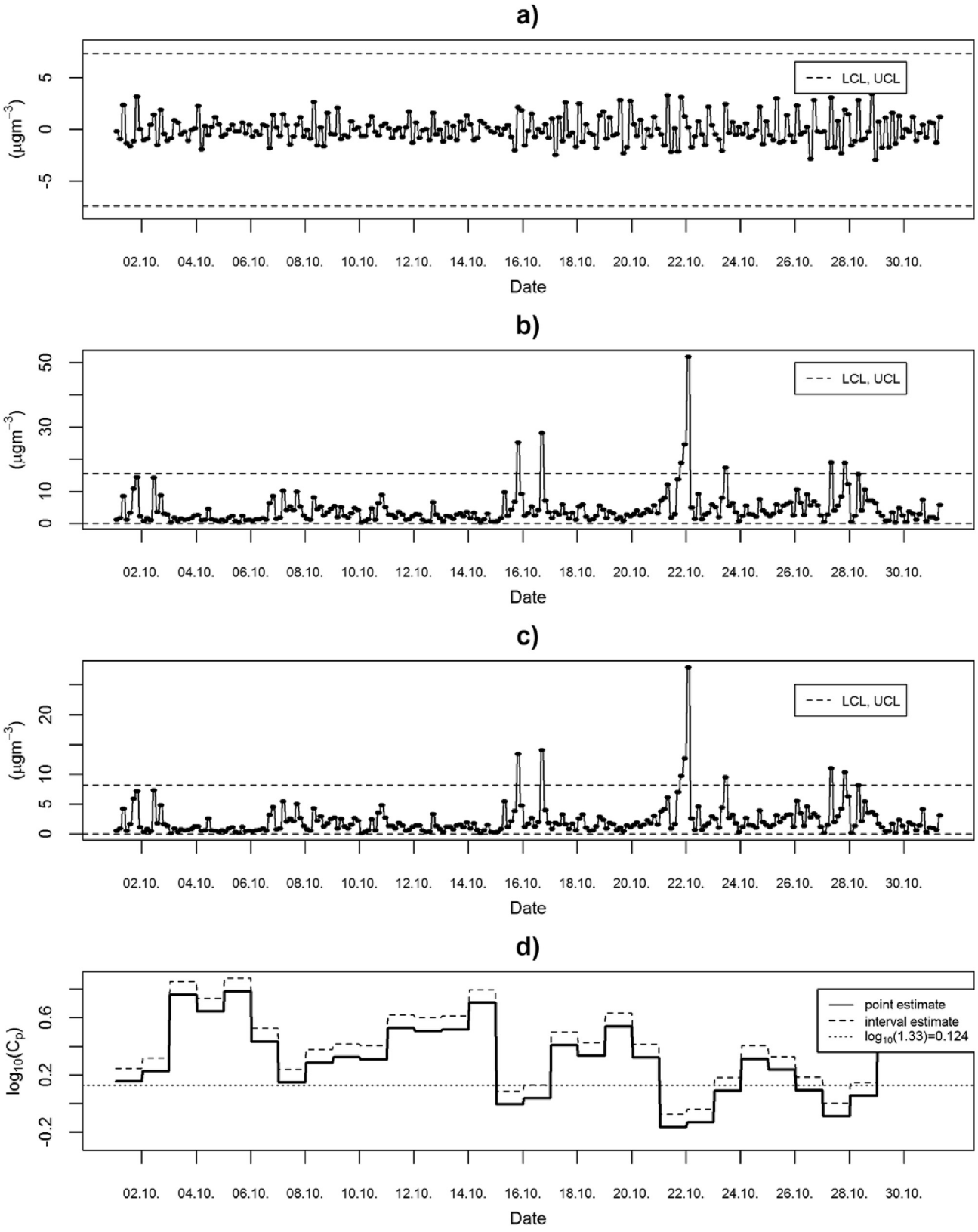

Fig. 3. Lany: a) $\bar{x}$ chart of the residuals, b) R chart of the residuals, c) s chart of the residuals, d) Point estimate $\hat{C}_p$ and right-sided interval estimate $\hat{C}_{p,U^*}$ for capability index, horizontal line at the level.$\log_{10}(1.33) \doteq 0.124$

using the proposed methods.

For station Turany the control charts based method did not find one segment containing measurement on 30th of July, which was manually detected as invalid. However the days detected by using Six sigma method included all observations, that were detected as invalid based on manual data control.

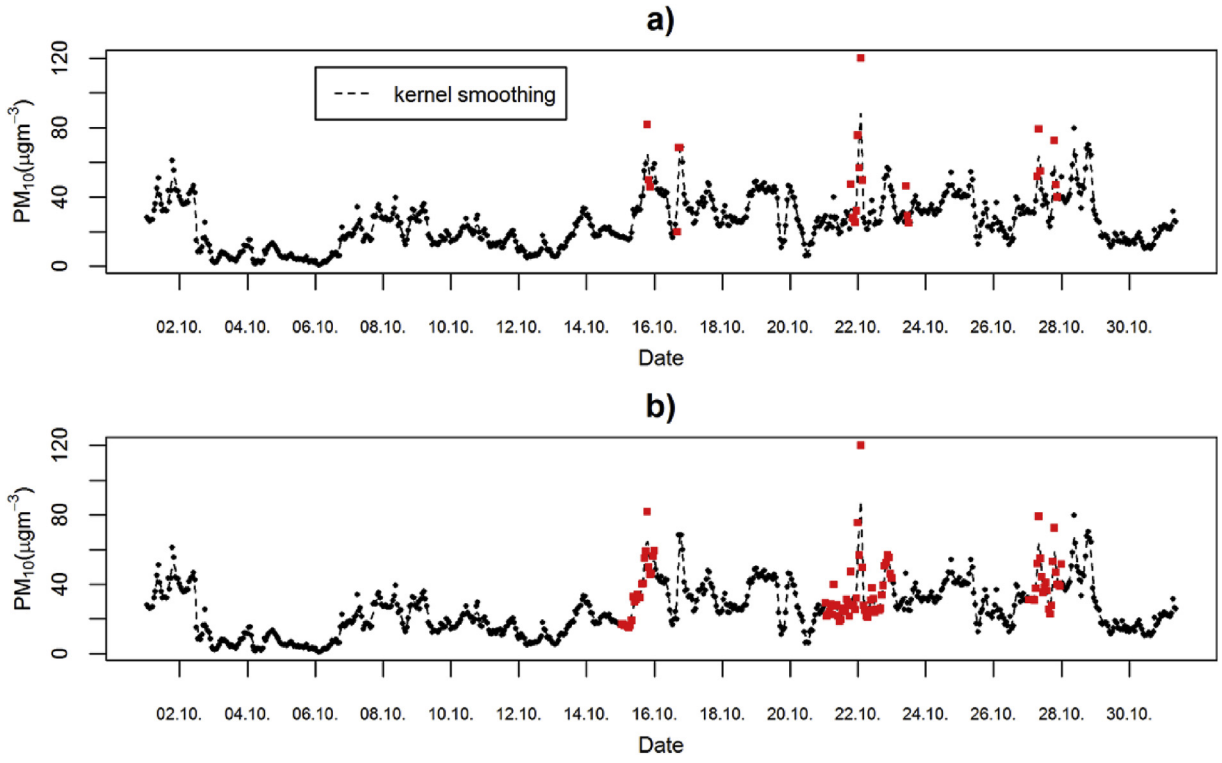

Fig. 4. Lany: a) Segments detected based on control charts, b) Days detected based on Six sigma methodology.

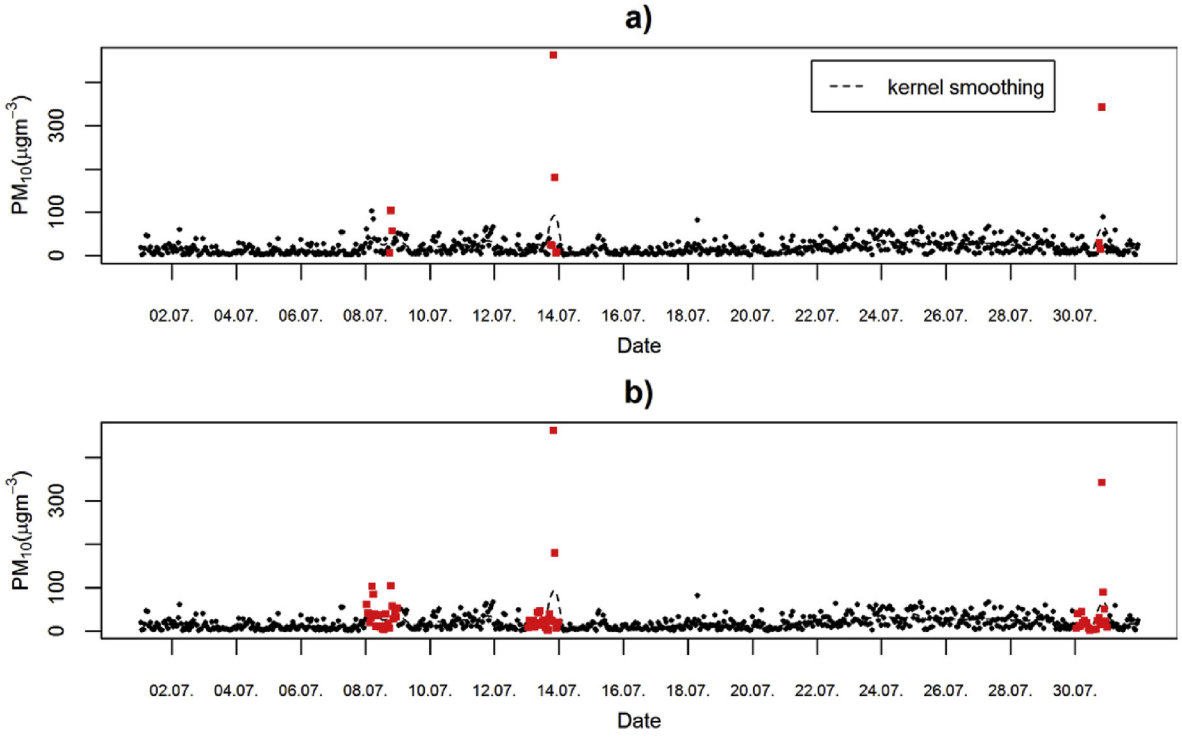

Fig. 5. Turany: a) Segments detected based on control charts, b) Days detected based on Six sigma methodology.

5. Discussion

By inspecting Fig. 4 we can see that for station Lany large number of observations that need to be further investigated was detected compared to the number of observations identified as invalid based on the manual data control. As can be seen from Fig. 1 the changes of variability of the residuals occur several times from 1st to 31st of July. Since the normality of the residuals was rejected the R chart control limits, which were constructed based on $L = 5$, can be regarded as 0.04 probability limits for sample ranges in

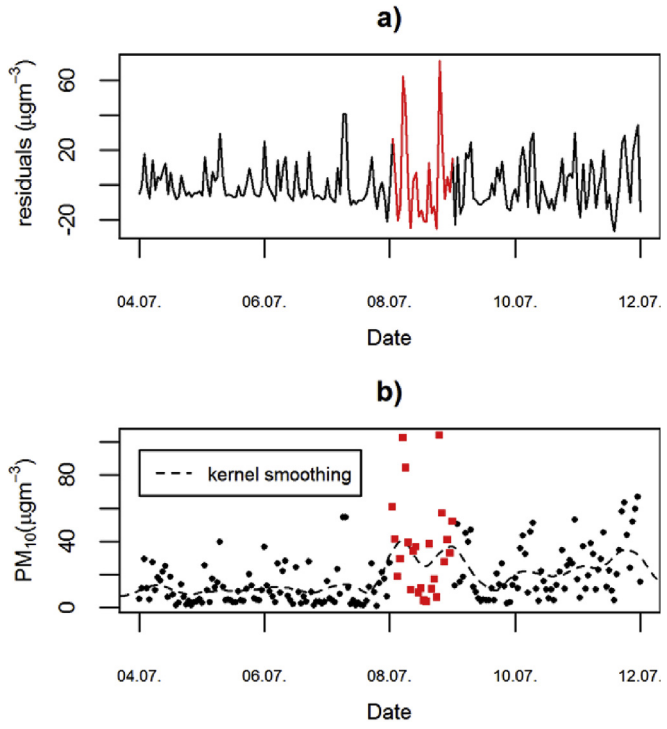

Fig. 6. Turany: a) Residuals corresponding to day - 8th of July, which was identified by using Six sigma based method, b) Measurements corresponding to day - 8th of July, which was identified by using Six sigma based method.

individual subgroups. Then, the probability that a point representing sample range falls above the upper control limit is four out of a hundred.

Considering the size of dataset presented in Fig. 1, the probability 0.04 corresponds to 29 observations, a value which exceeds the number of observations for further analysis. As described in section 3.2.1., the more the distribution of the residuals approaches to the normality the more the probability approaches to the value 0.0000006. By using the method based on Six sigma, which quantifies the variability of the residual process and thus assess the uniformity of the residuals, all days when the changes of the variability occur, were identified regardless to the cause of the variability change.

Although that except two significant changes of variability on 13th and 30th of July the residual process from station Turany appears to be stable, segments and days containing outliers were detected also on 8th of July by using both proposed methods. The reason is that a variability of the residuals on the considered period is relatively large (see Fig. 6) and standard deviation of the residual process thus exceeds the specification limits.

As pointed in section 1, the true reason for the presence of outliers in the data set can not be specified automatically and the value of the presented methodology is that the number of observations for manual treatment is reduced. Of course a stricter or milder criterion for the detection of outliers can be achieved by different choice of parameters n and L. As described in section 3.2.3., both of these parameters affect the amount of detected outliers. Because by using control charts based method all observations corresponding to sample mean or sample range exceeding the control limits are detected as outliers, and because Six sigma based method automatically detects 1-day intervals, the parameters n and L must be chosen also with regards to the number of observations that need to be further investigated.

By choosing the smaller value of n the number of observations that need to be further investigated is reduced. For this reason and also because we focus on outliers exceeding 3σ, smaller values of n ($n = 2, 3, 4$) are recommended.

The parameter L must be chosen with regards to the character of the data, because each monitoring station is site specific. Therefore we suggest to determine the value of L based on the analysis of historical data, that have already been manually validated.

We previously noted that for station Turany the proposed method based on control charts did not detect segment containing one invalid measurement from 30th of July, which has second highest value relative to the neighbouring observations. Inadequate detection of outliers occurring in a cluster of outliers is a general property and weakness of the method. The reason is that the estimate of the regression function based on the local bandwidth is locally influenced by the outliers occurring in a cluster and a value of the residuals can be small relative to the other residuals although that the corresponding observations of the analysed variable appear deviated from the neighbouring values. As can be seen from the first graph of Fig. 2 the estimate of regression function on 30th of July is influenced by the highest daily value and the residuals, that correspond to PM_{10} mass concentrations appearing deviated from the neighbouring measurements, are thus smaller than the residuals, that correspond to PM_{10} mass concentrations appearing consistent with the neighbouring values. Therefore in a case that the control charts based method identifies outlier in a cluster of observations significantly deviated from the other measurements, detailed inspection of all observations from the cluster is recommended. Also, the observations exceeding allowed physical range can be detected and omitted from the further analysis prior to kernel smoothing.

Another disadvantage of Six sigma based method is that setting relatively low value of parameter L (such that the residuals deviated less than 3 sigma from the other residuals are detected) results in the identification of too many observations for subsequent manual control. For this reason the Six sigma based method can not be used to detect outliers which are deviated less than three standard deviations from the mean value of the neighbouring observations. Therefore by using the method based on Six sigma the workload of the specialised researcher is cut down at the expense of missing some genuine outliers that are less extreme and that can not be correctly detected by using the suggested procedure.

6. Conclusions

Two methods for the automatic detection of segments in environmental data, where the outliers occur, have been presented. The first step of the procedures is based on kernel regression with variable bandwidth which is used to smooth the original data and this way to remove an influence of unknown covariates for further analysis. In the second step of the proposed methods, the segments, where the variability of the residuals is relatively high, are found by using control charts and Six sigma methodology.

Since none of the methods is able to exactly identify the outliers and invalid measurements, quality of the observations corresponding to detected segments must be further evaluated by a specialised researcher. The value of the proposed method is that the number of observations for manual data control is reduced. This is especially useful in controlling the quality of the data that are measured continuously with high temporal resolution. In such a case the data set contains large amount of measurements and manual control of data quality is very time demanding. Using the automatic detection of segments containing outliers can save a lot of time, because the specialised operator can deal only with the observations corresponding to detected segments (where the observations are in some way inconsistent with the rest of the data),

and does not need to control the quality of the entire data set.

The proposed methods were applied to problem of detection outliers in PM_{10} mass concentrations measured at two monitoring stations in Brno (Czech Republic). However, the procedures can be adapted to data from various environmental areas. We expect that suggested methods can be effectively implemented in data validation procedure in numerous types of data.

The presented analysis was performed using software R version 3.3.1. (R Core Team, 2013). The R script can be obtained by the first author [martina.campulova@unob.cz].

Acknowledgments

This work was supported by Ministry of Defence - project PASVŘII - DZRO K110. The authors would like to thank to Council of the City of Brno and to Czech Hydrometeorological Institute for providing the data.

References

Abrutzky, R., Dawidowski, L., Matus, P., Lankao, P.R., 2012. Health effects of climate and air pollution in Buenos Aires: a first time series analysis. J. Environ. Prot. 3, 262–271.

Barnett, V., Lewis, T., 1978. Outliers in Statistical Data. John Wiley, Chichester. New York.

Bobbia, M., Misiti, M., Misiti, Y., Poggi, J.-M., Portier, B., 2015. Spatial outlier detection in the PM10 monitoring network of Normandy (France). Atmos. Pollut. Res. 6, 476–483.

Brockmann, M., Gasser, T., Herrmann, E., 1993. Locally adaptive bandwidth choice for kernel regression estimators. J. Am. Stat. Assoc. 88, 1302–1309.

Brys, G., Hubert, M., Struyf, A., 2008. Goodness-of-fit tests based on a robust measure of skewness. Comput. Stat. 23, 429–442.

Burman, J., Otto, M., 1988. Outliers in Time Series. Statistical Research Division Report Series CENSUS/SRD/RR-88/14, vol. 44. Bureau of the Census.

Cao-Abad, R., González-Manteiga, W., 1993. Bootstrap methods in regression smoothing. J. Nonparametr. Stat. 2, 379–388.

Dupuis, D.J., Field, C.A., 2004. Large wind speeds: modeling and outlier detection. J. Agric. Biol. Environ. Stat. 9, 105–121.

Fan, J., Gijbels, I., 1995. Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation. J. Roy. Stat. Soc. B 57, 371–394.

Fan, J., Hall, P., Martin, M., Patil, P., 1996. On local smoothing of nonparametric curve estimators. J. Am. Stat. Assoc. 91, 258–266.

Fox, A., 1972. Outliers in time series. J. Roy. Stat. Soc. B Met. 34, 350–363.

Franchini, M., Mannuci, P.M., 2007. Short-term effects of air pollution on cardiovascular diseases: outcomes and mechanisms. J. Thromb. Haemost. 5, 2169–2174.

Gasser, T., Müller, H.G., 1979. Kernel estimation of regression functions. In: Gasser, T., Rosenblatt, M. (Eds.), Smoothing Techniques for Curve Estimation, Lect. Notes Math., 757. Springer, pp. 23–67.

Gupta, M., Gao, J., Aggarwal, C., 2014. Outlier detection for temporal data: a survey. IEEE Trans. Knowl. Data En. 26, 2250–2267.

Henry, R., Norris, G.A., Vedantham, R., Turner, J.R., 2009. Source region identification using kernel smoothing. Environ. Sci. Technol. 43, 4090–4097.

Herrmann, E., 1997. Local bandwidth choice in kernel regression estimation. J. Comput. Grap. Stat. 6, 35–54.

Keuken, M.P., Moerman, M., Voogt, M., Blom, M., Weijers, E.P., Rockmann, T., Dusek, U., 2013. Source contributions to PM2.5 and PM10 at an urban background and a street location. Atmos. Environ. 71, 26–35.

Kim, K.H., Kabir, E., Kabir, S., 2015. A review on the human health impact of airborne particulate matter. Environ. Int. 74, 136–143.

Kokalj, M., Rihtarič, M., Kreft, S., 2011. Commonly applied smoothing of IR spectra showed unappropriate for the identification of plant leaf samples. Chemom. Intell. Lab. Syst. 108, 154–161.

Michálek, J., 2009. Capability and Performance Indices of Manufacturing Process. ÚTIA Prague (in Czech).

Montgomery, D.C., 2009. Introduction to Statistical Quality Control, sixth ed. John Wiley & Sons, New York.

Pope, C.A., Dockery, D.W., 2006. Health effects of fine particulate air pollution: lines that connect. J. Air Waste Manag. Assoc. 56, 709–742.

R Core Team, 2013. R: a Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

Restrepo, C.E., Simonoff, J.S., Thurston, G.D., Zimmerman, R., 2012. Asthma hospital admissions and ambient air pollutant concentrations in New York City. J. Environ. Prot. 3, 1102–1116.

Russell, A.G., Brunekreef, B., 2009. A focus on particulate matter and health. Environ. Sci. Technol. 43, 4620–4625.

Samek, L., 2016. Overall human mortality and morbidity due to exposure to air pollution. Int. J. Occup. Med. Environ. Health 29, 417–426.

SAS/QC User's Guide, Version 8, 1999. SAS Institute, Cary, N.C.

Shaadan, N., Jemain, A.A., Latif, M.T., Deni, S.M., 2015. Anomaly detection and assessment of PM10 functional data at several locations in the Klang Valley, Malaysia. Atmos. Pollut. Res. 6, 365–375.

Shewhart, W.A., 1931. Quality control chart. Bell Syst. Tech. J. 5, 593–603.

Šmejdířová, J., 2016. Statistical Analysis of Selected Time Series Monitoring the Air Pollution. University of Defence, Faculty of Military Leadership, Brno, p. 96. Diploma thesis (in Czech).

Wand, M.P., Jones, M.C., 1995. Kernel Smoothing. Chapman and Hall, London.

Wild, C.J., Seber, G.A.F., 2000. Chance Encounters: a First Course in Data Analysis and Inference. John Wiley, New York.