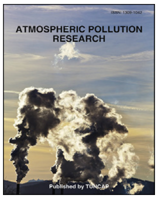
Atmospheric Pollution Research 9 (2018) 27–36

HOSTED BY

Contents lists available at ScienceDirect

Atmospheric Pollution Research

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



Semiparametric outlier detection in nonstationary times series: Case study for atmospheric pollution in Brno, Czech Republic

Jan Holešovský [a], [*], Martina Čampulová [b], [c], Jaroslav Michálek [b]

[a] *Brno University of Technology, Faculty of Civil Engineering, Institute of Mathematics and Descriptive Geometry, Veveří 95, 60200, Brno, Czech Republic*
[b] *University of Defence, Faculty of Military Leadership, Department of Econometrics, Kounicova 65, 66210, Brno, Czech Republic*
[c] *Mendel University in Brno, Faculty of Business and Economics, Department of Statistics and Operation Analysis, Zemědělská 1, 61300, Brno, Czech Republic*

ARTICLE INFO

Article history:
Received 13 February 2017
Received in revised form
5 June 2017
Accepted 13 June 2017
Available online 21 June 2017

Keywords:
Air pollution
Outlier
Kernel smoothing
Local bandwidth
Extreme value
Extremal index

ABSTRACT

Large environmental datasets usually include outliers which can have significant effects on further analysis and modelling. There exist various outlier detection methods that depend on the distribution of the analysed variable. However quite often the distribution of environmental variables can not be estimated. This paper presents an approach for identification of outliers in environmental time series which does not impose restrictions on the distribution of observed variables. The suggested algorithm combines kernel smoothing and extreme value estimation techniques for stochastic processes within considerations of nonstationary expected value of the process. The nonstationarity in variance is evaded by change point analysis which precedes the proposed algorithm. Possible outliers are identified as observations with rare occurrence and, in correspondence to extreme value methodology, the confidence limits for high values of observed variables are constructed. The proposed methodology can be especially convenient for cases where validation of the data has to be carried out manually, since it significantly reduces the number of implausible observations. For a case study, the technique is applied for outlier detection in time series of hourly PM_{10} concentrations in Brno, Czech Republic. The methodology is derived on solid theoretical results and seems to perform well for the series of PM_{10}. However its flexibility makes it generally applicable not only to series of atmospheric pollutants. On the other hand, the choice of return level turns out to be crucial in sensitivity to the outliers. This issue should be left to the practitioners to decide with respect to specific application conditions.

1. Introduction

Air pollution has negative impact on human health, ecosystem and the climate, and hence provides an important and complex problem. Air pollutants are emitted primarily directly from both natural and anthropogenic sources or formed secondary in the atmosphere from precursors. The local concentration of many air pollutants is problematic, especially in urban areas it may be also increased by long-range transport. Improving of air quality in Europe is therefore one of the priorities of present environmental policy. To move towards the air quality that does not have significant adverse effects on human health and the environment, both the Ambient Air Quality Directive of the Council and the European Parliament (EU, 2008) and Air Quality Guidelines (WHO, 2005) of WHO set limits for ambient concentrations of air pollutants.

One of the most significant pollutants in Europe with respect to negative impacts on human health is atmospheric aerosol (particulate matter, PM) with aerodynamic diameter of particles smaller than 10 μm, namely PM_{10}. Even relatively low concentrations of PM_{10} may noticeably affect human health and ecosystem. Numerous epidemiological studies have shown a positive association between PM_{10} exposure and negative health effects including increased mortality and morbidity, cardiovascular diseases and respiratory problems (see e.g. Pope et al., 1995; Pope and Dockery, 2006; Abrutzky et al., 2012; Restrepo et al., 2012). PM_{10} also causes damage to plants (Jimoda, 2012), reduces visibility and influences climate (Davison et al., 2005).

Although some improvements of the air quality have been achieved, PM_{10} concentration is still exceeding limits of EU as well

[*] Corresponding author.
E-mail address: holesovsky.j@fce.vutbr.cz (J. Holešovský).
Peer review under responsibility of Turkish National Committee for Air Pollution Research and Control.

http://dx.doi.org/10.1016/j.apr.2017.06.005
1309-1042/© 2017 Turkish National Committee for Air Pollution Research and Control. Production and hosting by Elsevier B.V. All rights reserved.

as stricter limits of WHO in large urban areas of Europe (Air Quality e-reporting database EEA, 2015). According to the short-term (24-hour) limit of European Union the daily average PM_{10} must not exceed the limit of $50 \cdot 10^{-6}$ g m^{-3} on more than 35 days in a calendar year. The long-term (annual) PM_{10} limit value is set at $40 \cdot 10^{-6}$ g m^{-3}.

Primary PM_{10} originates from a variety of natural and anthropogenic sources, while secondary particles are formed in the atmosphere by complex processes from gaseous precursors such as NO_2, SO_2 NH_3 and VOCs. Continuous monitoring of concentrations and composition of PM_{10} is essential for air pollution investigation as well as for the prediction and evaluation of periods with high-concentration of PM_{10}. However, not only the measurements are prerequisite of a good assessment of the air quality. It is known that large datasets often include outliers, which can significantly affect data analysis and modelling. The presence of outliers can also lead to misspecification in air quality evaluation with possible high expenses for its improvement. Measurements which are outlying from the other observed values may result from experimental errors as well as from abnormal behaviour of the observed variable. Detection and interpretation of outliers is, therefore, a critical and important part of data analysis.

It should be emphasized that from the perspective of practitioners it is only employed a visual inspection of the data supported eventually by the logs of device errors. This means that the outlier detection is in many cases provided purely by manual investigation of the given time series. Hence, in the context of atmospheric pollution, only evidently outlying observations are removed from the series while the less obvious values remain preserved. From a statistical point of view this seems to be inappropriate solution of the problem.

One of the first works for outlier detection in time series can be found in Fox (1972) and Burman and Otto (1988). Recently, various methods for outlier detection and data mining algorithms in both univariate and multivariate data have been proposed, for example in Gupta et al. (2014); Barnett (2004); Ben-Gal (2010); Chandola et al. (2009); Lee et al. (2000); Čampulová et al. (2017); Bobbia et al. (2015); Shaadan et al. (2015). Several methods enabling detection of outliers in multivariate time series have been discussed in Minguez et al. (2012). Weekley et al. (2010) focused on outlier detection procedures based on image processing and cluster analysis. In the context of atmospheric processes, the Grubb's test is often applied for the outlier detection (see Gerboles and Buzica, 2008; Gerboles et al., 2011). However the independence and normality of the observed data is required. Clearly, as we aim to, the test is not suited for observations in form of a time series, since the dependence can seriously harm the inference. Of course advanced parametric as well as non-parametric methods, which can be used to detect outlier observations in time-series, are still being proposed. The improvements comprise mostly the involvement of covariates. From the view of atmospheric observations this may lead to an extensive need of accompanying time series of all species as discussed in section 2 below. Another methods applied for air pollution time series which are based on clustering can be found in D'Urso et al. (2015, 2017).

However relatively little attention has been paid to extreme value (EV) models used for the purpose of outlier detection. These techniques are primarily based on own behaviour of the observed series. Some approaches have been the object of study, for example, in Roberts (1999); Dupuis and Field (2004); Burridge and Taylor (2006); Holešovský and Kůdela (2016); D'Urso et al. (2016), but the analysis is mostly done under very specific settings. Dupuis and Field (2004) proposed a robust procedure for fitting a distribution to high values, whereby each observation is assigned a weight. The weights are than compared against datasets generated artificially under the assumption of model validity. Similar to Burridge and Taylor (2006), the methodology is suitable solely for independent and identically distributed (i.i.d.) random variables, and thus inappropriate for long-run time series validation. The local EV estimation described in Roberts (1999) seems to be more adequate, but only the Gumbel distribution case is here considered. D'Urso et al. (2016) developed fuzzy clustering models with time-dependent EV-parameters. The estimates are obtained at the basis of annual maxima separated from the series (see further section 3.2). The parameters are estimated with large variability.

In this paper we present a novel semiparametric technique for outlier identification in time series without any need of accompanying covariates. The method is based on EV estimation of high threshold exceedances with no additional constraints on particular distributional form or EV domain of attraction. Generalization of EV theory to stationary processes is described in the literature (see e.g. Leadbetter et al., 1983; Beirlant et al., 2004). However EV estimation for a nonstationary series can be limited to specific instances only, assuming the form of dependence is known. In order to handle this issue and to develop a methodology applicable to a wide range of cases, we propose a two step procedure which uses results obtained by kernel smoothing performed prior to EV estimation for stationary series. The use of kernel smoothing for outlier detection has been already investigated by Čampulová et al. (2017) in combination with control charts and six sigma methodology. Both control charts and six sigma based algorithms, in contrast to the method proposed in this paper, can label only a segment of time series which could suffer from outliers. The principle of the methods suggested in (Čampulová et al., 2017) is to smooth the data and subsequently analyse the residuals using control charts and six sigma methodology. The aim is to find the segments where the residual process behaves unstable and incapable due to the presence of outliers. The method based on EV quantile estimation indicates exact points, leading to simplification or even to complete removal of manual inspection of the data.

Note that the true reason for the presence of outliers can not be specified using the presented method and the quality of the automatically detected outliers must be further evaluated manually. The value of the proposed methodology is that the number of observations for manual data control is reduced.

The paper is organized in the following manner. In the next section we give an overview of the data and conditions under which PM_{10} concentrations were observed. In section 3 we introduce the methodology. Particularly, we describe a local weighted kernel smoothing procedure, and give outline of EV estimation for stationary processes. The methodology for outlier detection is summarized to the end of the section. The discussed technique is applied to PM_{10} concentrations in section 4. Finally, in section 5, we give conclusions.

2. Data

The PM_{10} concentrations were hourly recorded at 5 monitoring stations in Brno, Czech Republic operated by Brno City Municipality (BCM). Brno is the second largest city of the Czech Republic with population of 430,000 inhabitants, and thus represents an area with significant air pollution mostly originating from industrial sources. The stations are equipped with diverse measurement systems, dependent on the level of modernization. For the purpose of our case study, we select two particular stations, namely Arboretum and Zvonarka whose observation period was from November 2007 and from November 2006, respectively, until November 2015. At the first one the PM_{10} concentrations are collected by radiometric dust-meter using the absorption of beta radiation, the latter one is equipped with optoelectronic device. The observation locations are placed at 250 m and 200 m above sea level, and cover areas with various demands on air quality.

The site Zvonarka is a heavily loaded traffic spot within an industrial area surrounded by parking lot and the bus station. The traffic intensity is around 43,000 vehicles per day from which there is about 10% haulage, average speed is 40 km/h. Zvonarka is by BCM classified as traffic-urban station. Its placement is 10 m distant from the roadway and 50 m from the road-intersection. In the vicinity to the monitoring site there is the railway station and other heavy-traffic roads.

On the other hand, the site Arboretum is installed in the botanical garden at the campus of Mendel University. There is an important traffic intersection close to the station (distant 105 m) with around 11,000 vehicles per day, 5% haulage, and average speed 70 km/h. Nevertheless, the surroundings consist mostly of residential housing and thus the character of the site is rather suburban. According to BCM classification the location is labelled as background-suburban.

Clearly, for the purpose of outlier detection both records may exhibit quite different behaviour. Moreover, the site Zvonarka is located in rather flat area, while the site Arboretum is placed in upper part of a mild hill. Thus one can expect heavier upper tails in the PM_{10} distribution for the first mentioned one.

As pointed out in Hübnerová and Michálek (2014); Mikuška et al. (2017); Křůmal et al. (2017), the dust aerosol is influenced by miscellaneous factors including presence of a weekday, heating season, cloud cover, or wind speed. Hence, the series in the monitored period is strongly nonstationary and a suitable model can suffer from excessive complexity. More details on the data can be also found in Hrdličková et al. (2008) where attention has been paid to identification of significant factors affecting the air pollution in Brno.

3. Methods

3.1. Kernel smoothing

The first step of the proposed method is based on kernel smoothing, which is a statistical technique widely used to estimate a regression function from noisy observations in case that no parametric model for the function is available.

Because a heteroscedastic model is considered for smoothing, the data are divided into several intervals such that the variability of the observed variable is similar (approximately constant) on each interval. The partitioning is performed by using change point analysis which enables to find time instants in which the distribution of the observed variable changes. We suggest to estimate the change points by using the Pruned Exact Linear Time (PELT) algorithm of Killick et al. (2012), which is computationally fast, but other algorithms, discussed e.g. in Scott and Knott (1974); Auger and Lawrence (1989), can be used as well.

Assuming that the unknown variable Y has been measured at n different times $t_1, \dots, t_n$ lying in an interval $[a, b]$ determined by change point analysis, the measurements $Y_1, \dots, Y_n$ should satisfy

$$Y_i = m(t_i) + \sigma(t_i)\varepsilon_i, \quad i = 1, \dots, n, \tag{1}$$

where m denotes an unknown regression function, ε_i are i.i.d. random variables with zero mean and unit variance, and $\sigma(t_i)$ is a standard deviation function expressing the variance of Y_i. The functions $\sigma(t)$ and $m(t)$ are supposed to meet standard regularity assumptions specified e.g. in Herrmann (1997).

The regression function at a point is estimated as a weighted mean of the neighbouring observations where the weights are defined by a suitable choice of kernel function. In principle there exist several estimators of the kernel regression function. Here we focus on the Gasser-Müller estimator which estimates the regression function on the interval $\langle a + h_t, b - h_t \rangle$ by

$$\widehat{m}(t, h_t) = \sum_{i=1}^{n} Y_i \int_{l_{i-1}}^{l_i} \frac{1}{h_t} K\left(\frac{t-u}{h_t}\right) du, \tag{2}$$

where $K(\bullet)$ denotes kernel function (shortly kernel) of order $(0, k)$ (Gasser et al., 1985), $h_t = h(t)$ is bandwidth at point t, and limits of integration are given by $l_0 = a, l_i = 0.5(t_i + t_{i+1})$ for $i = 1, \dots, n-1$, $l_n = b$.

From variable kernel functions the Epanechnikov kernel (see Gasser et al., 1985), which belongs to the most widely used, is preferred. However the bandwidth choice, which is still being discussed, influences the estimate of regression function much more than the kernel itself. The problem of bandwidth selection is, therefore, the critical and inevitable part of the kernel smoothing.

The classical methods for global bandwidth estimation are usually based on cross-validation techniques (Wand and Jones, 1995), Akaike information criterion, and its improved version (Hurvich et al., 1998; Harrold et al., 2001). A different approach for selection of optimal smoothing parameters is done in order to minimize the asymptotic integral mean square error (AMISE) of the fit. Hereby the plug-in principle is usually applied, i.e. the distribution characteristics are replaced by their empirical counterparts. The unknown functions in the expression of AMISE, such as derivative of regression function, standard deviation function, and design density, are estimated from sample points. Plug-in algorithms which can be used to select global bandwidth in kernel and local polynomial regression were proposed for example in Gasser et al. (1991) and Ruppert et al. (1995), respectively. Several plug-in methods focusing on construction of local bandwidth as minimiser of asymptotic mean squared error (AMSE) have been proposed in Fan and Gijbels (1995); Ruppert (1997); Herrmann (1997).

The estimate of the regression function based on local bandwidth, which can adapt to the data structure locally and capture complicated features in the data, leads to better practical results (Müller and Stadtmüller, 1987). Thus for the purpose of the outlier detection method presented in this paper local bandwidth estimated by using local plug-in algorithm is preferred (Herrmann, 1997). The algorithm, which generalizes methodology discussed by Brockmann et al. (1993) for heteroscedastic models with non-equidistant design, enables to find the local estimate of smoothing parameter h_t iteratively. The total number of iterations is fixed and equal to $(k+1)(2k+1)+1$. In the first $(k+1)(2k+1)$ iterations a sequence of global bandwidths is generated, whereby the local bandwidth is estimated based on global bandwidth from the preceding iteration.

3.2. Extremal models

In the second step the kernel smoothing residuals $X_1, \dots, X_n$, constant in variance, are analysed by using EV theory. EV theory provides adequate framework for frequency estimation of high values as documented by numerous papers (see e.g. Caeiro and Gomes, 2010; Fawcett and Walshaw, 2012; Holešovský et al., 2016; Madsen et al., 2002; Fawcett and Walshaw, 2016). The inference is usually based on the assumption that $X_1, \dots, X_n$ are i.i.d. random variables with an unknown cumulative distribution function (c.d.f.). It can be shown that, under some regularity conditions (see de Haan and Ferreira, 2006), the sample maxima $M_n = \max\{X_1, \dots, X_n\}$ follows a generalized extreme value (GEV) distribution, say, with c.d.f. $G(x)$. The distribution is three parametric with μ for location, σ for scale, and ξ for shape parameter. We write GEV(μ, σ, ξ) to emphasize the specific parameters of the distribution. Specifically, the parameter ξ has substantial effect on tail properties of the GEV distribution (with heavy tail for $\xi > 0$), and has to be properly estimated. Classical method on EV estimation, the block maxima approach, is based on the distribution of maxima taken from blocks large enough. See e.g. Beirlant et al. (2004) for more details.

Recently, the peaks-over-threshold (POT) approach is often preferred in the literature (Fawcett and Walshaw, 2012, 2016; Madsen et al., 2002; Silva et al., 2016; Alonso et al., 2014). For a threshold value u large enough and a random variable X, it can be shown that the distribution of variable $X - u$ conditioned by $X > u$ follows a generalized Pareto (GP) distribution approximately, i.e. $\Pr(X \le x \mid X > u) \approx H(x-u)$. The GP c.d.f. $H(x)$ is for $x > u$ defined as follows

$$H(x) = 1 - \left(1 + \xi \frac{x}{\sigma_u}\right)_+^{-1/\xi}, \tag{3}$$

where $a_+ := \max(a, 0)$. Parameters ξ and $\sigma_u > 0$ are shape and scale parameters respectively, and we denote the case GP(σ_u, ξ). Particularly, shape parameter of a GP distribution corresponds directly to the shape parameter of the related GEV distribution for sample maxima.

Given a threshold high enough, the exceedances are supposed to follow a GP distribution, and at this basis are estimated the parameters and other parametric functions of interest. However, a proper threshold selection still belongs to unsolved problems of the POT method. There is traditional trade-off between significant bias if threshold is set too low, and increasing variability for threshold too large. Usually, an optimal threshold is selected as low as possible in order to ensure a reasonable fit of the exceedances by a GP distribution. Review of techniques for proper threshold selection, including several obsolete ones, is given in Scarrott and MacDonald (2012). Nowadays, many authors focused on development of automated methods. Some of them can be found in Northrop and Coleman (2014); Draisma et al. (1999); Neves and Alves (2004), for example.

To the moment we focused on i.i.d. observations only. The EV theory can be extended to stationary series framework, being more suitable in real situations. The properties of random variables remain homogeneous in time, however some dependencies can arise in the series. In such cases either additional sampling has to be applied or one has to deal with dependence in the series. The first way consists in designing a separation scheme which allows to draw out approximately independent observations. In this fashion it is solely proceeded in practical applications. However as discussed in literature (see e.g. Ancona-Navarrete and Tawn, 2000), the separation scheme often suffers from choice of auxiliary parameters, suggested by various rules of thumb. Hence, estimation of the dependence structure seems to be more adequate in this case.

Suppose the underlying i.i.d. series $X_1, \ldots, X_n$ replaced with a stationary series satisfying a short-time dependence. A mixing type $D(u_n)$ condition of Leadbetter et al. (1983) is being considered in order to obtain time-distant variables being nearly independent. It can be shown that the limiting distribution of sample maxima M_n (drawn now from the stationary series) remains a GEV distribution, namely GEV$(\mu_\theta, \sigma_\theta, \xi_\theta)$ with c.d.f. $G_\theta(x)$. In correspondence to c.d.f. $G(x)$ for the i.i.d. maxima, the functions $G(x)$ and $G_\theta(x)$ are related by the equality

$$G_\theta(x) = [G(x)]^\theta, \tag{4}$$

where $0 \le \theta \le 1$, the so-called extremal index, is a measure of short-time dependence at extremal levels (Beirlant et al., 2004). Under the presence of dependence, the extreme values tend to cluster, i.e. a high value is more likely followed by another. According to possible interpretation of θ derived by Leadbetter et al. (1983), the value θ^{-1} specifies the expected cluster size. Obviously, $\theta = 1$ for an i.i.d. series, however the opposite implication does not hold (Ancona-Navarrete and Tawn, 2000).

Since the limiting distribution of M_n remains a GEV distribution, the inference can still be based on the POT model. The estimated parameters are corrected by the value of θ afterwards. Several advanced estimators for the extremal index have been proposed in recent years, see e.g. Northrop (2015); Süveges (2007); Ferro and Segers (2003); Ancona-Navarrete and Tawn (2000); Gomes (1993). Detail comparison of the latter two shows (Holešovský et al., 2014), that the estimator of Ancona-Navarrete and Tawn (2000) gains advantage in smaller variability, and can be more suitable for θ near its boundary. However, the estimator proposed by Gomes (1993) reveals overall better stability to the choice of auxiliary parameters, and will be further preferred.

For the purpose of outlier identification, we want to estimate a high quantile of the data. In EV theory, this is referred to r-observation return level z_r, i.e. value that is exceeded once every r observations on average. Hence, the return level z_r is the $(1 - r^{-1})$ quantile of a model distribution, specifically for an i.i.d. POT model and $\xi \ne 0$ can be z_r obtained in the form

$$z_r = u + \frac{\sigma_u}{\xi}\left[(\lambda_u r)^\xi - 1\right], \tag{5}$$

where $\lambda_u := \Pr(X > u)$. The return level estimate $\hat{z}_r$ can be obtained substituting all parameters by their (for example maximum likelihood) estimates; specifically, λ_u is estimated as relative frequency of the number of exceedances. If a stationary series is considered, the estimates need to be corrected by the value of θ, namely the return level z_r corresponds to $(1 - r^{-1})^{\theta^{-1}}$ quantile of i.i.d. series with the same marginal distribution (see Fawcett and Walshaw, 2012, 2016). Thus, the member $(\lambda_u r)^\xi$ in relation (5) needs to be replaced with

$$\left(\lambda_u^{-1}\left[1 - \left(1 - r^{-1}\right)^{\theta^{-1}}\right]\right)^{-\xi}. \tag{6}$$

Variability of parameter estimates is usually determined on the basis of asymptotic normality of the estimators. Hence, the variability of a more complex parametric function, such as return level z_r, can be estimated by approximation of the delta method (Beirlant et al., 2004).

3.3. Outlier detection methodology

The aim of our outlier identification method is to achieve a sort of confidence limits for extremely high values of a series. Given the data $Y_1, \ldots, Y_n$ measured at time instants $t_1, \ldots, t_n$ which belong to an interval where the variance of the observed variable is approximately constant, the algorithm can be formally written as follows:

1. For each $t_i, i = 1, , n$, calculate the estimate of regression function $\hat{m}(t_i, h_{t_i})$ with local bandwidth h_{t_i} estimated by plug-in algorithm (Herrmann, 1997).
2. Determine the smoothing residuals $X_i = Y_i - \hat{m}(t_i, h_{t_i})$, $i = 1, \ldots, n$, an approximately stationary series.
3. Choose a proper threshold value and compute GP(σ_u, ξ) parameter estimates; estimate extremal index θ.

4. For a given value r estimate return level $\widehat{z}_r$ using (5) and (6) by substitution of all parameters by their estimates.
5. Compose $\widehat{m}$ and $\widehat{z}_r$ to obtain confidence limits for high values; detect possible outliers where $Y_i > \widehat{m}(x_i, h_{x_i}) + \widehat{z}_r,\ i = 1, \ldots, n,$ holds.

For step 3, the threshold selection, various techniques can be applied. With intent to possible automation of the outlier detection procedure, we choose the value of threshold as a fixed high empirical quantile of the data. Other automated techniques, as mentioned above, can be used, however these are often followed by increased computational demands.

Note that the foregoing methodology lies no constraints to a specific distributional form of the observed data. The maximum likelihood (ML) method is being often used for the purpose of GP parameter estimation. The advantage of ML method is that it produces asymptotically normal distributed estimates without any additional conditions on the underlying distribution as it is required by non-parametric methods (see de Haan and Ferreira, 2006). However, as it was discussed by Smith (1985) and Zhou (2009), the ML method is suitable only for the case $\xi > -1/2$. Nevertheless, usually no attention has to be paid to this restriction while the value of ξ mostly fulfils this restriction in environmental issues.

On the other hand, a proper change point detection plays a crucial presumption of the above described methodology. Omitting changes in variance of the underlying series can significantly harm the suitability of extremal models. The violation of stationarity may misspecificate the estimation of the GP parameters and the extremal index, whereby the period with smaller variance is by the POT model rather neglected. This leads to substantially biased estimation.

In case that the interval determined by change point analysis contains a relatively small amount of observations the plug-in algorithm of Herrmann (1997) can lead to undersmoothing in the local extremes of regression function. This problem, which is caused by the fact that the global bandwidth from first iteration is based on the sample size, can be solved by smoothing with global bandwidth which can be estimated by using an algorithm proposed e.g. in Gasser et al. (1991).

It should be also mentioned that, although it is not the object of interest in our case, confidence limits for extremely low values can be obtained in similar way. It is then required to carry out the EV analysis for negated residuals in steps 3–5 owing to the relation $\min\{X_1, \ldots, X_n\} = -\max\{-X_1, \ldots, -X_n\}$.

4. Results

In this section we demonstrate the above discussed outlier identification procedure, and present the results for atmospheric pollution time series from the city of Brno. Recall from section 2, the PM_{10} concentrations were recorded hourly at stations with various geographic, weather, and pollution conditions. From those we concentrate only to two of them, to stations Zvonarka and Arboretum, which as we expect may exhibit different behaviour of outliers. Particularly, because of extensive range of the data, for detailed inference we reduced the time series to specific periods. Namely, for station Arboretum we take the period from January 13th to May 6th 2015, and for station Zvonarka we concentrate to the observations from May 14th to July 24th 2014.

This restriction is done for several reasons. First, these periods were indicated variance-stationary by the change point PELT algorithm. Hence necessary condition for our outlier identification procedure is satisfied. Second, by visual inspection of all possible periods determined by PELT these two seem to suffer from evidently outlying values and with high potential to presence of less obvious outliers as well. Finally, the reduction was also needed for a reasonable graphical visualization of the results.

Since the methodology for outlier detection described in foregoing section is based purely on own behaviour of the time series, we do not consider necessary the periods for two stations to overlap. On the other hand, the authors are aware that more detail analysis should be carried out if the outlier detection is conducted. Particularly relevant are the periods corresponding to the seasons of inversion, flowering etc. But the results shown below can be taken into account as demonstrative use of the methodology. Moreover, we still keep in mind that our technique would support the manual validation procedure rather than stand alone as an automatic method.

In Fig. 1 are shown monthly box-plots of PM_{10} concentrations during the years 2014 and 2015. Lower temperatures in winter and spring months are associated with household heating, which is significant source to PM_{10}. Moreover, the decrease in low temperature uncertainty during the summer leads to narrower boxes in this period. The differences between the two observation sites are well visible from the number of box-plot outliers, whereby these indicate overall heavier tails in PM_{10} concentrations observed at the site Zvonarka. The whiskers are set to show 1.5 inter-quantile range.

The estimates of the regression functions for the selected periods are visualized on Fig. 2 and Fig. 3 for station Arboretum and Zvonarka, respectively. The smoothing residuals are also plotted, forming an approximately stationary series meant for extreme quantile estimation. The estimate of regression function based on local bandwidth adapts to the data and the curvature of the regression line changes based on the variability of PM_{10} concentration. This phenomenon is visible especially in local extremes of the regression function where the smoothing line has sharp peaks.

As can be seen from Fig. 2 several steady increases and sharp declines occur in the concentrations of PM_{10} aerosols measured at site Arboretum during the studied measurement period especially from March 1st to April 1st. Although that the unusual behaviour seems to be a consequence of malfunction of the measurement equipment, there was no measurement or experimental error during the considered period as verified by inspecting the station logbook of interventions. Detailed examination revealed that all changes in the concentration took several hours, the only exception was a change in March 11th that was much faster. All these changes are associated with a change in meteorological conditions. Analysis of the basic meteorological parameters measured simultaneously with the concentrations of PM_{10} aerosols verified that except of changes in the wind direction no abrupt changes in the trend of other observed meteorological variables were present. A sharp decline in PM_{10} concentration during March 11th also results from different meteorological situation as was proven by backward trajectories calculated for sampling point in Arboretum that show sharp change in trajectories of air masses coming to Brno from March 15th to March 16th.

The steady increases and sharp declines similar to those shown in Fig. 2 were observed also for the concentrations of PM_{10} measured at remaining four stations in Brno. It can be concluded that the changes in PM_{10} concentrations at all Brno monitoring stations operated by BCM are caused by changes in meteorological conditions, especially wind direction.

The residuals are dispersed around the zero horizontal axis. Relatively high values occur mostly in time instants corresponding to the peaks of regression function. This willing property was expected and is the advantage of local kernel smoothing. Global bandwidth selection, in the opposite to a local choice we use in our case, would rather lead to over-smoothing of the outliers as well as to introduction of significant bias to the estimate of the regression

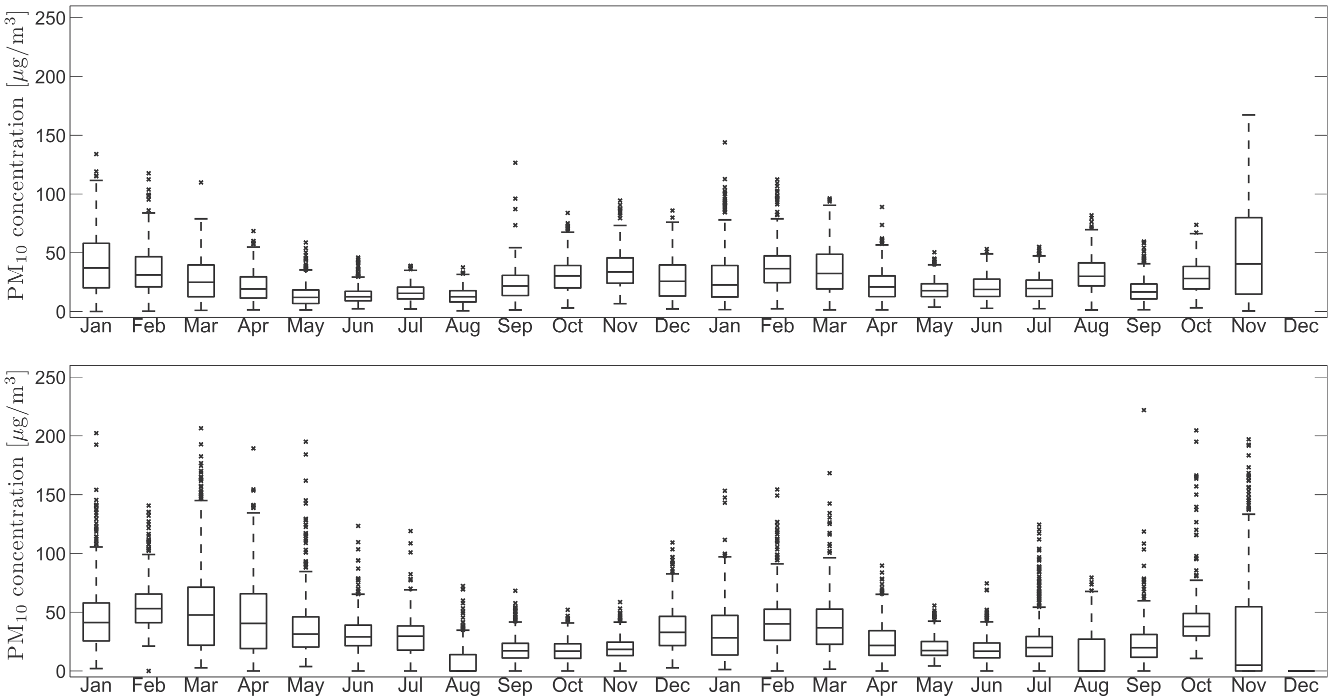


Fig. 1. Monthly box-plots of PM$_{10}$ concentrations during the years 2014 and 2015 for the site Arboretum (upper fig.) and the site Zvonarka (lower fig.). Whiskers show 1.5 interquartile range.

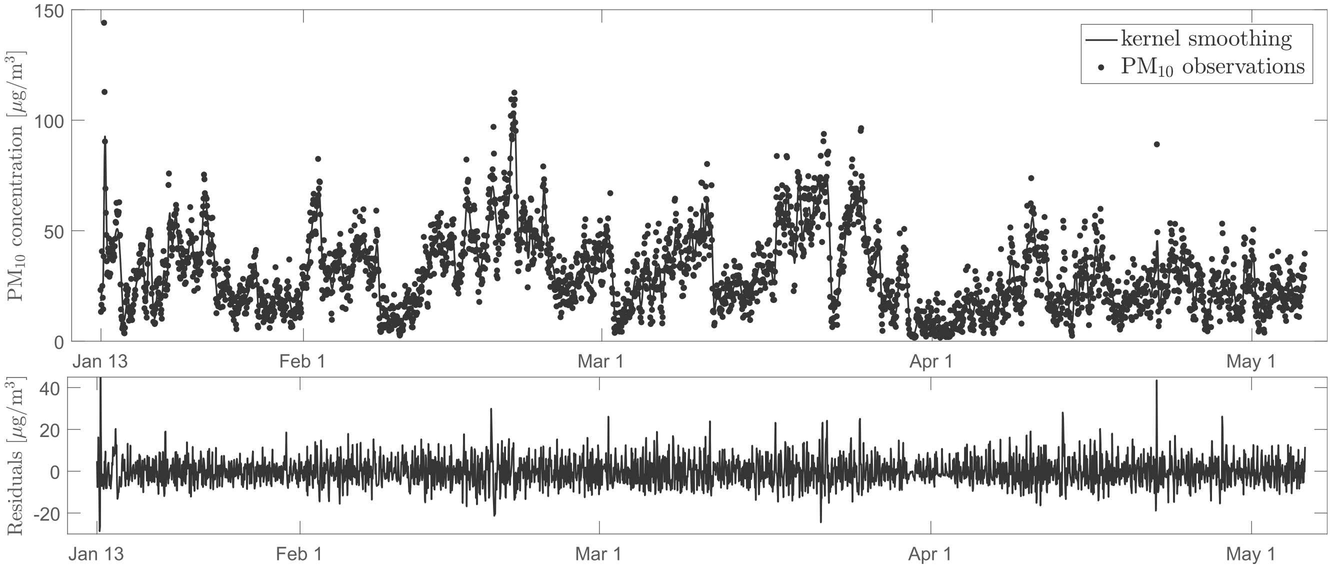


Fig. 2. PM$_{10}$ observations and kernel smoothing regression estimate (upper fig.), and smoothing residuals (lower fig.) for the period January 13th – May 6th 2015 at site Arboretum.

function. Specifically, the largest residuals are present at the beginning of January and at the end of April 2015 for Arboretum station and at the beginning of April 2014 for Zvonarka station. Furthermore, several other relatively large outliers are contained in the residual series for both stations.

On this basis can be estimated the GP parameters meant for determination of PM$_{10}$ return level. Prior to that, a proper threshold needs to be selected. As already mentioned, we choose the threshold value as high enough empirical quantile from the residuals, namely the 90% quantile. This approach gains advantage especially in its simplicity and suitability towards possible automation of the algorithm with no increase in computational demands. The 90% quantile is considered high enough to ensure a reasonable fit of the threshold exceedances. In correspondence to observation periods shown on Figs. 2 and 3, the threshold was determined as $u = 8.138 \cdot 10^{-6}$ g m^{-3} at the site Arboretum and $u = 3.363 \cdot 10^{-6}$ g m^{-3} at the site Zvonarka.

Considering the threshold exceedances only, we use the ML method to evaluate the estimates. Fig. 4 shows empirical quantiles plotted against quantiles of the estimated GP distribution. The fit is rather in good agreement with the limiting GP distribution. Several extraordinary deflections in the plots acknowledge the presence of outliers in both series. Obviously, since the outliers are located at high quantiles only, no other choice of any higher threshold can

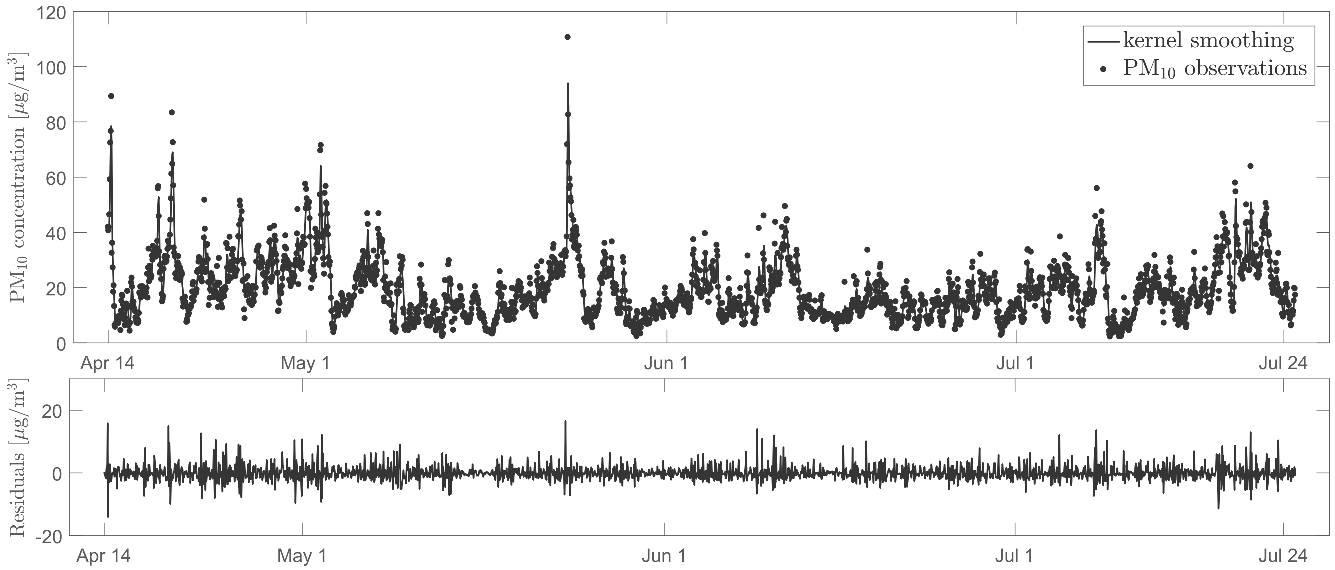


Fig. 3. PM$_{10}$ observations and kernel smoothing regression estimate (upper fig.), and smoothing residuals (lower fig.) for the period May 14th – July 24th 2014 at site Zvonarka.

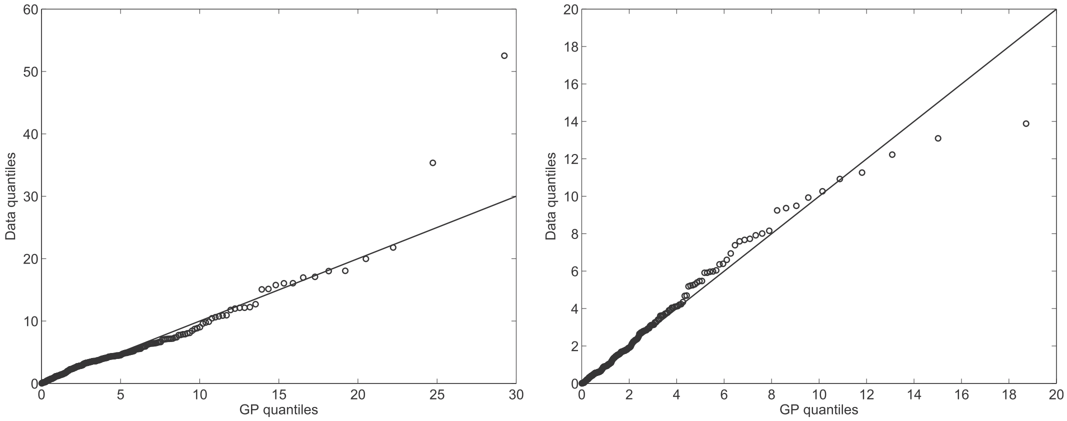


Fig. 4. Plots of threshold exceedances empirical quantile against fitted GP quantile for the series from station Arboretum (left) and Zvonarka (right). Threshold value was set to 90% empirical quantile of the residuals.

lead to their omission. For the assessment of the fit the Pearson χ^2 and Kolmogorov-Smirnov (KS) goodness-of-fit tests were performed (summarized in Table 1). All p-values indicate good agreement of the exceedances with the GP distribution, and the corresponding threshold values are thus considered reliable at both sites. Specific results, including standard deviation of the parameter estimates, are given in Table 2 for both particular cases. These are in correspondence to tail characteristics visible from Fig. 1, i.e. larger value of shape parameter ξ reveals heavier tail for the series from station Zvonarka. One could use the normal asymptotic properties of ML estimates for confidence interval estimation of ξ obtained by multiplication of the standard deviation by a suitable normal quantile. At site Arboretum this particularly means that the confidence interval for ξ, being the interval $[-0.010, 0.202]$, includes the point zero at significance level 0.95, and the hypothesis of heavy tail is rejected for the series observed at the site Arboretum (unlike to the station Zvonarka).

Table 1
The p-values obtained from goodness-of-fit tests performed to assess the agreement of threshold exceedances with a limiting GP distribution.

Site	p-values	
	χ^2 test	KS test
Arboretum	0.1210	0.1541
Zvonarka	0.4578	0.5290

Table 2
Maximum likelihood estimates of parameters of POT model for time periods of interest. Standard deviations of the estimates are given in parentheses.

Site	Estimated parameters			
	Shape ξ	Scale σ_u	Frequency λ_u	Extr. index θ
Arboretum	0.096 (0.054)	3.945 (0.321)	0.100 (0.006)	0.780 (0.028)
Zvonarka	0.211 (0.008)	1.800 (0.038)	0.100 ($4 \cdot 10^{-5}$)	0.957 (0.051)

On the basis of GP estimates were determined the return levels z_r to be rarely exceeded. Subsequently the confidence limits for high values of PM$_{10}$ could be set. The r-observation confidence limits were evaluated as combination of the regression function and an return level z_r, where we consider the values $r = 24, 48, \ldots, 240$. These correspond to the bounds that should be exceeded once in $1, 2, \ldots, 10$ days. Specific estimates are shown in

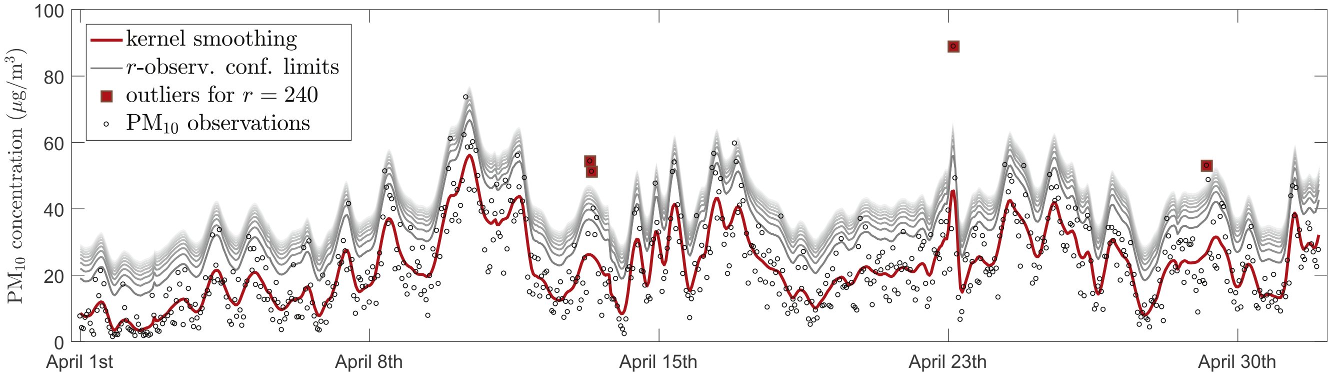


Fig. 5. Station Arboretum: *r*-observation confidence limits estimation obtained as combination of regression function and *r*-observation return level estimation for the values $r = 24, 48, \ldots, 240$ (from bottom to top). Threshold value selected as 90% empirical quantile. Square markers indicate outliers identified over conf. limit for $r = 240$.

Fig. 5 and in Fig. 6 for station Arboretum and Zvonarka, respectively. The square markers indicate outliers that have been identified as observations exceeding the 240-observation confidence limit, i.e. $r = 240$.

Since the return period *r* controls the width of the confidence interval, a proper choice of this parameter is the main issue to be determined. Especially, it should be chosen with respect to empirical experience including historical data and specific application requirements. A huge advantage of the application of extremal models is the ability to take into account the dependencies of a time series. This is a crucial benefit that it is gained in contrast to the models based on intermediate value theory described by the Central limit theorem. However, it should be expected that such dependency is considered only at extremal levels. In correspondence to this fact the return period *r* should be chosen large enough to ensure only a small number of observations out of the confidence bounds, whereby from theoretical point of view the models may turn out to be less suitable for *r* small.

All computations were performed in Matlab MathWorks Inc., 2016 environment and in the software R version 3.3.1 using packages ”lokern” (Herrmann, 2014) and ”changepoint” (Killick et al., 2015). For the case study we considered a ”fixed” threshold equal to 90% empirical quantile. More advanced techniques can be of course applied, see section 3. Particular interest may be paid to the adaptive methods. While these adaptive techniques gain support from some theoretical aspects, they are mostly accompanied by extensive growth in computational demands. Under the fixed-threshold settings are the computational times negligible; similarly in terms of kernel smoothing. Further details and scripts are available from the authors.

5. Conclusion

A method for outlier identification in environmental time-series has been introduced. The core idea of the method is to detect outliers by comparison of the original data with values exceeded on average once a specified period. Considering the methodology, the procedure consists of kernel smoothing with local bandwidth and extreme value estimation of high threshold exceedances. The result is a confidence limit for high values of the observed variable, which is constructed from composition of regression function estimates and estimates of EV return levels. Analogically, confidence limits for low values of observed time-series can be constructed.

In comparison to other outlier identification techniques for time series, as discussed in Section 1, the EV methodology enables to take into account the dependency between consecutive observations. This is a huge advantage to widely applied methods that are usually based on the Central limit theorem. The outliers are determined according to an EV-based criterion, and under a suitable setup enables full automation of the process. Of course, the parameters used in the proposed methodology must be selected based on the experience with historical data in collaboration with the specialists.

Since the proposed method is not able to distinguish the outliers caused by measurement and experimental errors from the outliers that result from unusual measurement conditions or from natural variability of the observed variable, the quality of the automatically detected outliers must be further evaluated by a specialised researcher. The value of the suggested procedure is that the number of observations for manual data inspection is reduced. This is helpful from the perspective of practitioners, who evaluate the

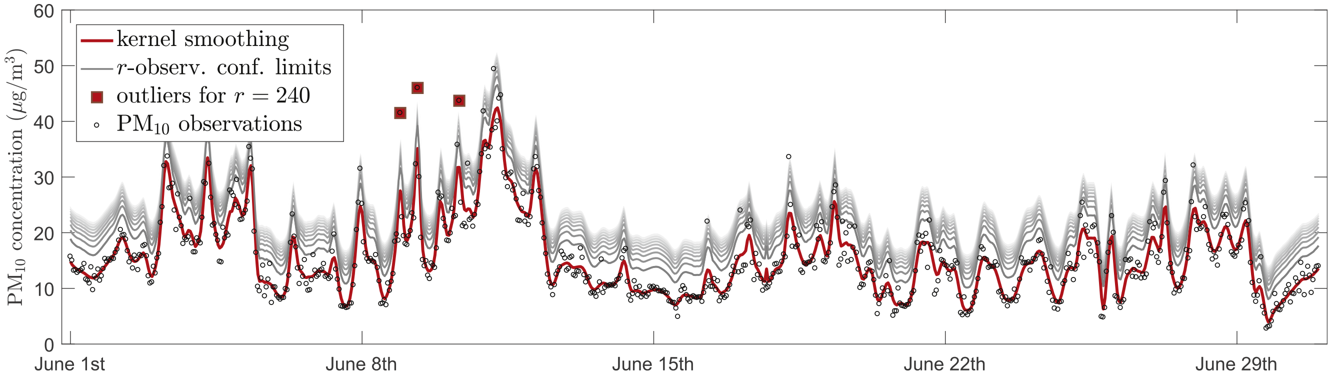


Fig. 6. Station Zvonarka: *r*-observation confidence limits estimation obtained as combination of regression function and *r*-observation return level estimation for the values $r = 24, 48, \ldots, 240$ (from bottom to top). Threshold value selected as 90% empirical quantile. Square markers indicate outliers identified over conf. limit for $r = 240$.

quality of the data that are measured continuously with high temporal resolution. It is obvious that the larger the dataset is the more time for manual data validation is needed. Using the automatic detection of outliers can save a lot of time, because a practitioner specialised on a studied data can concentrate only on the detected outliers and does not need to evaluate the quality of the entire data set.

The suggested method has been applied to solve the problem of high-value outlier detection of hourly PM_{10} concentrations measured at two stations in Brno, Czech Republic. Nevertheless the methodology is generally applicable, and it is suited for data from various environmental areas. We expect that the method can be effectively applied as a part of a data validation procedure, wherever data control is required.

Acknowledgement

The authors would like to thank to Pavel Mikuška for helpful discussions about the atmospheric pollution. This paper has been worked out under the project No. LO1408 “AdMaS UP - Advanced Materials, Structures and Technologies,” supported by Ministry of Education, Youth and Sports under the “National Sustainability Programme I” and under the project PASVŘ II (DZRO K-110) supported by Ministry of Defence.

References

Abrutzky, R., Dawidowski, L., Matus, P., Lankao, P., 2012. Health effects of climate and air pollution in Buenos Aires: a first time series analysis. J. Environ. Prot. 3, 262–271.
Alonso, A.M., de Zea Bermudez, P., Scotto, M.G., 2014. Comparing generalized Pareto models fitted to extreme observations: an application to the largest temperatures in Spain. Stoch. Environ. Res. Risk Assess. 28, 1221–1233.
Ancona-Navarrete, M., Tawn, J., 2000. A comparison of methods for estimating the extremal index. Extremes 3 (1), 5–38.
Auger, I.E., Lawrence, C.E., 1989. Algorithms for the optimal identification of segment neighborhoods. B. Math. Biol. 51 (1), 39–54.
Barnett, V., 2004. Environmental Statistics: Methods and Applications, first ed. Wiley Series in Probability and Statistics. Wiley.
Beirlant, J., Goegebeur, Y., Segers, J., Teugels, J., de Waal, D., Ferro, C., 2004. Statistics of Extremes: Theory and Applications, first ed. Wiley.
Ben-Gal, I., 2010. Outlier detection. In: Mainon, O., Rokach, L. (Eds.), Data Mining and Knowledge Discovery Handbook, second ed. Springer, pp. 117–130.
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 (3), 476–483.
Brockmann, M., Gasser, T., Herrmann, E., 1993. Locally adaptive bandwidth choice for kernel regression estimators. J. Am. Stat. Assoc. 88 (424), 1302–1309.
Burman, J., Otto, M., 1988. Outliers in Time Series. Statistical Research Division Report Series CENSUS/SRD/RR-88/14. Bureau of the Census.
Burridge, P., Taylor, M., 2006. Additive outlier detection via extreme-value theory. J. Time Ser. Anal. 27 (5), 685–701.
Caeiro, F., Gomes, M., 2010. Semi-parametric tail inference through probability-weighted-moments. J. Stat. Plan. Infer. 141, 937–950.
Čampulová, M., Veselík, P., Michálek, J., 2017. Control chart and six sigma based algorithms for identification of outliers in experimental data, with an application to particulate matter PM10. Atmos. Pollut. Res. 8 (4), 700–708.
Chandola, V., Banerjee, A., Kumar, V., 2009. Anomaly detection: a survey. ACM Comput. Surv. 41 (3) article 15.
Davison, C., Phalen, R., Solomon, P., 2005. Airborne particulate matter and human health: a review. Aerosol Sci. Tech. 39, 737–749.
de Haan, L., Ferreira, A., 2006. Extreme Value Theory: an Introduction, first ed. Springer.
Draisma, G., de Haan, L., Peng, L., Pereira, T., 1999. A bootstrap-based method to achieve optimality in estimating the extreme-value index. Extremes 2 (4), 367–404.
Dupuis, D., Field, C., 2004. Large wind speeds: modeling and outlier detection. J. Agric. Biol. Envir. St. 9 (1), 105–121.
D'Urso, P., De Giovanni, L., Massari, R., 2015. Time series clustering by a robust autoregressive metric with application to air pollution. Chemom. Intell. Lab. 141, 107–124.
D'Urso, P., Maharaj, E., Alonso, A., 2016. Fuzzy clustering of time series using extremes. Fuzzy Set. Syst. 318, 56–79.
D'Urso, P., Massari, R., Cappelli, C., De Giovanni, L., 2017. Autoregressive metric-based trimmed fuzzy clustering with an application to pm10 time series. Chemom. Intell. Lab. 161, 15–26.
EEA, 2015. Air Quality in Europe. Report 5. EEA iSBN 978-92-9213-702-1.
EU, 2008. Directive 2008/50/ec of the European Parliament and of the Council of 21 may 2008 on ambient air quality and cleaner air for Europe. Off. J. Eur. Commun. L 152, 1–44.
Fan, J., Gijbels, I., 1995. Data-driven bandwidth selection in local polynomial regression: variable bandwidth selection and spatial adaptation. J. Roy. Stat. Soc. B Met. 57 (2), 371–394.
Fawcett, L., Walshaw, D., 2012. Estimating return levels from serially dependent extremes. Environmetrics 23, 272–283.
Fawcett, L., Walshaw, D., 2016. Sea-surge and wind speed extremes: optimal estimation strategies for planners and engineers. Stoch. Environ. Res. Risk Assess. 30, 463–480.
Ferro, C., Segers, J., 2003. Inference for clusters of extreme values. J. Roy. Stat. Soc. B Met. 65 (2), 545–556.
Fox, A., 1972. Outliers in time series. J. Roy. Stat. Soc. B Met. 34 (3), 350–363.
Gasser, T., Kneip, A., Kohler, W., 1991. A flexible and fast method for automatic smoothing. J. Am. Stat. Assoc. 86, 643–652.
Gasser, T., Müller, H.-G., Mammitzsch, V., 1985. Kernels for nonparametric curve estimation. J. Roy. Stat. Soc. B Met. 47 (2), 238–252.
Gerboles, M., Buzica, D., 2008. Intercomparison Exercise for Heavy Metals in PM10. Tech. rep., EUR 23219 EN. Office for Official Publications of the European Communities, Luxembourg. http://dx.doi.org/10.2788/63349. http://publications.europa.eu/.
Gerboles, M., Buzica, D., Brown, R., et al., 2011. Interlaboratory comparison exercise for the determination of As, Cd, Ni and Pb in PM10 in Europe. Atmos. Environ. 45 (20), 3488–3499.
Gomes, M., 1993. On the estimation of parameter of rare events in environmental time series. In: Statistics for the Environment. Vol. 2 of Water Related Issues. Wiley, pp. 225–241.
Gupta, M., Gao, J., Aggarwal, C., 2014. Outlier detection for temporal data: a survey. IEEE T. Knowl. Data En. 26 (9), 2250–2267.
Harrold, T.I., Sharma, A., Sheather, S., 2001. Selection of a kernel bandwidth for measuring dependence in hydrologic time series using the mutual information criterion. Stoch. Environ. Res. Risk Assess. 15 (4), 310–324.
Herrmann, E., 1997. Local bandwidth choice in kernel regression estimation. J. Comput. Grap. Stat. 6 (1), 35–54.
Herrmann, E., 2014. Lokern: Kernel Regression Smoothing with Local or Global Plug-in Bandwidth. R Package Version 1.1–6. Packaged for R and enhanced by Martin Maechler. https://CRAN.R-project.org/package=lokern.
Holešovský, J., Fusek, M., Blachut, V., Michálek, J., 2016. Comparison of precipitation extremes estimation using parametric and nonparametric methods. Hydrol. Sci. J. 61 (13).
Holešovský, J., Fusek, M., Michálek, J., 2014. Extreme value estimation for correlated observations. In: Proc. of 20th International Conference on Soft Computing MENDEL 2014. Brno University of Technology, pp. 359–364.
Holešovský, J., Kůdela, J., 2016. Outlier identification based on local extreme quantile estimation. In: Proc. of 22th International Conference on Soft Computing MENDEL 2016. Brno University of Technology, pp. 255–260.
Hrdličková, Z., Michálek, J., Kolář, M., Veselý, V., 2008. Identification of factors affecting air pollution by dust aerosol PM10 in Brno city, Czech Republic. Atmos. Environ. 42 (37), 8661–8673.
Hübnerová, Z., Michálek, J., 2014. Analysis of daily average PM10 predictions by generalized linear models in Brno, Czech Republic. Atmos. Pollut. Res. 5 (3), 471–476.
Hurvich, C., Simono, J., Tsai, C.-T., 1998. Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. J. Roy. Stat. Soc. B 60 (2), 271–293.
Jimoda, L., 2012. Effects of particulate matter on human health, the ecosystem, climate and materials: a review. Facta Univ. Ser. Work. Living Environ. Prot. 9 (1), 27–44.
Killick, R., Fearnhead, P., Eckley, I.A., 2012. Optimal detection of changepoints with a linear computational cost. J. Am. Stat. Assoc. 107 (500), 15901598.
Killick, R., Haynes, K., Eckley, I.A., 2015. Changepoint: an R Package for Changepoint Analysis. R package version 2.2. http://CRAN.R-project.org/package=changepoint.
Křůmal, K., Mikuška, P., Večeřa, Z., 2017. Characterization of organic compounds in winter PM1 aerosols in a small industrial town. Atmos. Pollut. Res. 8 (5), 930–939.
Leadbetter, M., Lindgren, G., Rootzén, H., 1983. Extremes and Related Properties of Random Sequences and Series, first ed. Springer.
Lee, W., Stolfo, S., Mok, K., 2000. Adaptive intrusion detection: a data mining approach. Artif. Intell. Rev. 14 (6), 533–567.
Madsen, H., Mikkelsen, P., Rosbjerg, D., Harremöes, P., 2002. Regional estimation of rainfall intensity-duration-frequency curves using generalized least squares regression of partial duration series. Water Resour. Res. 38 (11), 21–1–21–11.
MathWorks Inc, 2016. MATLAB 9.0. Natick, Massachusetts.
Mikuška, P., Kubátková, N., Křůmal, K., Večeřa, Z., 2017. Seasonal variability of monosaccharide anhydrides, resin acids, methoxyphenols ans saccharides in PM2.5 in Brno, the Czech Republic. Atmos. Pollut. Res. 8 (3), 576–586.
Minguez, R., Reguero, B.G., Luceno, A., Mendez, F.J., 2012. Regression models for outlier identification (hurricanes and typhoons) in wave hindcast databases. J. Atmos. Ocean. Technol. 29, 267–285.
Müller, H.-G., Stadtmüller, U., 1987. Variable bandwidth kernel estimators of regression curves. Ann. Stat. 15 (1), 182–201.
Neves, C., Alves, M.F., 2004. Reiss and Thomas' automatic selection of the number of extremes. Comput. Stat. Data An. 47, 689–704.

Northrop, P., 2015. An efficient semiparametric maxima estimator of the extremal index. Extremes 18 (4), 585–603.

Northrop, P., Coleman, C., 2014. Improved threshold diagnostic plot for extreme value analyses. Extremes 17, 289–303.

Pope, C., Dockery, D., 2006. Health effects of fine particuate air pollution: lines that connect. J. Air. Waste Manage. 56, 709–742.

Pope, C., Dockery, D., Schwartz, J., 1995. Review of epidemiological evidence of health effects of particulate air pollution. Inhal. Toxicol. 7, 1–18.

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

Roberts, S., 1999. Novelty detection using extreme value statistics. In: IEE P-Vis. Image Sign, vol. 146, pp. 124–129.

Ruppert, D., 1997. Empirical-bias bandwidths for local polynomial nonparametric regression and density estimation. J. Am. Stat. Assoc. 92 (439), 1049–1062.

Ruppert, D., Sheather, S., Wand, M., 1995. An effective bandwidth selector for local least squares regression. J. Am. Stat. Assoc. 90, 1257–1270.

Scarrott, C., MacDonald, A., 2012. A review of extreme value threshold estimation and uncertainty quantification. REVSTAT 10 (1), 33–60.

Scott, A.J., Knott, M., 1974. A cluster analysis method for grouping means in the analysis of variance. Biometrics 30 (3), 507–512.

Shaadan, N., Jemain, A., Latif, M., Deni, S., 2015. Anomaly detection and assessment of PM10 functional data at seeral locations in the Klang Valley, Malaysia. Atmos. Pollut. Res. 6 (2), 365–375.

Silva, A.T., Naghettini, M., Portela, M.M., 2016. On some aspects of peaks-over-threshold modeling of floods under nonstatioanrity using climate covarites. Stoch. Environ. Res. Risk Assess. 30, 207–224.

Smith, R., 1985. Maximum likelihood estimation in a class of nonregular cases. Biometrika 72 (1), 67–90.

Süveges, M., 2007. Likelihood estimation of the extremal index. Extremes 10, 41–55.

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

Weekley, R.A., Goodrich, R.K., Cornman, L.B., 2010. An algorithm for classification and outlier detection of time-series data. J. Atmos. Ocean. Technol. 27, 94–107.

WHO, 2005. Air Quality Quidelines for Particulate Matter, Ozone, Nitrogen Dioxide and Sulfur Dioxide, Global Update 2005. World Health Institution [Available online at. http://www.euro.who.int/_gerbol#03A9data/assets/pdf_file/0005/78638/E90038.pdf.

Zhou, C., 2009. Existence and consistency of the maximum likelihood estimator for the extreme value index. J. Multivar. Anal. 100, 794–815.