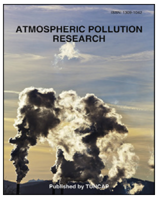
HOSTED BY

ELSEVIER

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 xxx

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/

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

Please cite this article in press as: Holešovský, J., et al., Semiparametric outlier detection in nonstationary times series: Case study for atmospheric pollution in Brno, Czech Republic, Atmospheric Pollution Research (2017), http://dx.doi.org/10.1016/j.apr.2017.06.005