ACTA UNIVERSITATIS AGRICULTURAE ET SILVICULTURAE MENDELIANAE BRUNENSIS

Volume 66 48 Number 2, 2018

https://doi.org/10.11118/actaun201866020453

COMPARISON OF METHODS FOR SMOOTHING ENVIRONMENTAL DATA WITH AN APPLICATION TO PARTICULATE MATTER PM_{10}

Martina Čampulová[1]

[1]Department of Statistics and Operation Analysis, Faculty of Business and Economics, Mendel University in Brno, Zemědělská 1, 613 00, Brno, Czech Republic

Abstract

ČAMPULOVÁ MARTINA. 2018. Comparison of Methods for Smoothing Environmental Data with an Application to Particulate Matter PM_{10}. *Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis,* 66(2): 453–463.

Data smoothing is often required within the environmental data analysis. A number of methods and algorithms that can be applied for data smoothing have been proposed. This paper gives an overview and compares the performance of different smoothing procedures that estimate the trend in the data, based on the surrounding noisy observations that can be applied on environmental data.
The considered methods include kernel regression with both global and local bandwidth, moving average, exponential smoothing, robust repeated median regression, trend filtering and approach based on discrete Fourier and discrete wavelet transform. The methods are applied to real data obtained by measurement of PM_{10} concentrations and compared in a simulation study.

Keywords: data smoothing, trend filtering, environmental data, particulate matter PM_{10}

INTRODUCTION

Univariate data smoothing techniques, which permit the trend in the data to be estimated from noisy observations, are frequently applied in many environmental applications (Holešovský *et al.*, 2017; Čampulová *et al.*, 2017; Lee and Kang, 2015; Hutchinson, 1995; Cook and Peters, 1981).

Smoothing methods commonly applied for environmental data include moving average (Kaneko and Funatsu, 2015), Savitzky-Golay smoothing (Sankaran *et al.* 2010; Kaneko and Funatsu, 2015), smoothing splines (He *et al.*, 2016), kernel regression techniques with global bandwidth (Henry *et al.*, 2009) or nonlinear smoothers (Kafadar and Morris, 2002).

Time series data is often smoothed using exponential smoothing based algorithms (Holt, 1957), regression-based approaches (Bowerman *et al.* 2005), Kalman Filters (Tsay, 2005) or decomposition methods (Hyndman and Koehler, 2008; Tsay, 2005), which are useful also for the prediction.

A number of methods are parametrised by a smoothing parameter which gives a rate of exchange between residual error and local variation. The choice of the smoothing parameter, which influences the final estimate of the smoothing line, is a crucial part of the analysis. Depending on the value of the parameter the smoothing line can vary from noisy function exactly explaining all data points to smooth function. If the value of the smoothing parameter is overestimated the data is over-smoothed and the detailed local information is lossed. On the other hand, underestimating the smoothing parameter results in disturbation of the general trend contained in the data by local features of noise.

The selection of optimal bandwidth is even more complicated in a case where the data contains intervals with both relatively low and relatively high variability. In such a case using a greater value of the smoothing parameter results in a better fit of flat parts and under-fitting of segments with increased variability of the data, while smoothing line based on a smaller value of the smoothing parameter over-fits the flat parts and fits well the intervals with relatively high variability of the data. This is a general problem associated with smoothing environmental data that is influenced by many different factors.

To overcome the problem of unsatisfactory local predictive accuracy of the smoothing line, the various methods based on variable smoothing parameter approach have been proposed. Methods derived from the classical smoothing methods and improved by substituting global smoothing parameter by local smoothing parameter include e.g. kernel regression based on local bandwith (Herrmann, 1997; Fann *et al.*, 2011) or adaptive spline smoothing (Wang *et al.*, 2013; Storlie *et al.*, 2010). A different approach adopted by wavelet methods (Alsberg *et al.*, 1997; Walczak and Massart, 1997) is based on the idea that a function representing the trend in the data in a given time instant can be expressed as a sum of basis functions of different scale.

Considering the problem of smoothing environmental data relatively little attention has been paid to the application of signal filtering methods which are often referred to as „data filtering“ or „noise denoising“ procedures.

Data filtering based on splines can be found in (Unser, 1999). Another method suggested for signal data filtering and often used for time series smoothing is median filtering (Tukey, 1977). In (Davies *et al.* 2004) robust repeated median regression, which estimates the unknown regression function using repeated medians (Siegel, 1982), is suggested. Such estimate improves median filtering by approximating underlaying data trend in a moving time window by a linear trend. In (Fried, 2004) robust regression suggested in (Davies *et al.*, 2004) is further modified by including rules for outlier detection and treatment to increase the robustness of the method. Another procedure, l1 trend filtering that is commonly used in signal data processing, has been proposed in (Kim *et al.*, 2009).

Specific class of methods for signal data processing are change-point detection algorithms (Chen and Gupta 2012; Bleakley and Vert, 2011; Neubauer and Vesely 2011; Zhang *et al.*, 2015) that are used to detect abrupt shifts and sharp changes in the mean values of the analysed data. However, considering environmental data, jump changes are not very common and ususally correspond to manipulation with the measuring instrument which means that the time of this change is known.

Of course, advanced methods for processing signal data that can be used for data smoothing are still being proposed. In (Ridel *et al.*, 2015) a fast least-square fitting method of parabolic cylinders for smoothing noisy data has been introduced. An effective iterative method for approximation of underlying data trend based on error measure has been presented in (Levin, 2015). A method that is based on convex optimization and that partitions the data into segments where the polynomail structure of the data is assumed was introduced recently in (Rajmic *et al.*, 2017).

In (Tibshirani, 2014) the trend filtering estimates were compared with smoothing splines and their alternative with better local adaptivity, so called locally adaptive regression splines (Mammen and Geer, 1997). It was shown that trend filtering estimates significantly outperform smoothing splines in local adaptivity. Considering locally adaptive regression splines the adaptivity properties were shown to be comparable. However, using trend filtering a lower computational complexity is achieved.

The aim of this paper is to present and give overview and comparison of the performance of different approaches for smoothing environmental data and to provide some insight to specialised operators. The considered methods include moving average, simple exponential smoothing (Holt, 1957), kernel regression with local (Herrmann, 1997) and global (Gasser *et al.*, 1991) bandwidth, robust repeated median regression (Fried, 2004), l1 trend filtering (Kim *et al.*, 2009) and smoothing based on discrete Fourier (Proakis, 1995) and discrete wavelet transform (Donoho and Johnston, 1995).

The present paper deals with comparison of kernel regression because it was implemented as a data smoothing method within outlier detection algorithms described in (Holešovský *et al.*, 2017; Čampulová *et al.*, 2017). Robust repeated median regression was selected because it represents signal filtering method which permits (analogous to kernel regression with local bandwidth) the smoothing parameter to be estimated locally. DWT was chosen because its principle is different from kernel regression approach and because the coefficients from DWT are estimated locally. Moving average method and exponential smoothing represent classical methods used for smoothing time series and l1 trend filtering is frequently used for signal data processing. Fast Fourier transform is a commonly used tool for frequency filtering.

The methods are applied to smooth hourly measurements of mass concentration of particulate matter $PM_{10.}$ Note that the methods are based on different models, which means they are hardly comparable. For this reason their performance is compared using simulated data.

MATERIALS AND METHODS

In this section the selected methods for smoothing measumenents of variable *Y* observed at *n* discrete time instants *t* are shortly described.

Moving average method

Moving average method estimates the data trend in a given time instant as the average of the noisy observations on a local segment. Supposing the measurements *Y(1), ..., Y(n)* and a parameter *h* called window size, the smoothed data *m(t)* in time *t* are given by

$$m(t)=\frac{1}{2h+1}\sum_{j=-h}^{j=h}Y(t+j).$$

It is obvious that the amount of observations for local averaging is determined by the choice of smoothing parameter *h*. Here the parameter is chosen using the approach described in (Barsanti and Gilmore, 2011).

Exponential smoothing

Whereas moving average method smooths the data in a given time instant using a fixed number of the most recent and equally weighted observations, using exponential smoothing all past observations are considered and exponentially weighted. The exponential weights to individual observations are assigned in a decreasing order such that the influence of the distant observations to the estimate in a given time instant diminishs exponentially over time.

Given the measurements *Y(t)* in time *t*, where *t = 2, ..., n*, the smoothed data *m(t)* in time *t* are given by (Paul, 2011)

$$m(t)=\alpha Y(t-1)+(1-\alpha)\hat{m}(t-1), \quad \alpha\in(0,1\rangle$$

where α is a smoothing parameter determining the weight assigned to the most recent measurement and $\hat{m}(t-1)$ is the estimate of smoothed data in time *t - 1*. As can be seen, the estimate of a smoothed data in a given time *t* is based on the observation corresponding to time where the estimate is computed and all past observations whose weights decrease exponentially over time. As for every smoothing parameter the choice of α determines the quality of the estimate. Here the parameter is chosen using a trial and error approach (Paul, 2011) such that mean squared error (MSE) is minimised.

Kernel regression

Kernel regression is a nonparametric smoothing technique which estimates the trend in the data (regression function) at a given time instant as a weighted mean of the surrounding noisy observations. The weights are determined by the choice of kernel function and the amount of noisy observations used for averaging is defined by a parameter called bandwidth.

Given the measurements $Y(t_i)$ observed at *n* discrete time instants t_i, $i = 1, ..., n$, lying in the interval [*a, b*] , the heteroscedastic regression model can be written in the form (Herrmann, 1997)

$$Y(t_i)=m(t_i)+\sigma(t_i)\varepsilon(t_i), \quad i=1,\ldots,n,$$

where $\varepsilon(t_i)$ are independent and identically distributed (i.i.d.) random errors with zero mean and unit variance and $\sigma(t_i)$ is standard deviation function expressing the variance of $Y(t_i)$. The functions $m(t_i)$ and $\sigma(t_i)$ are supposed to meet standard regularity assumptions which are given e.g. in (Herrmann, 1997).

Among various available estimators of the regression function $m(t_i)$ (Wand and Jones, 1995) Gasser-Muller convolution estimator (Gasser and Müller, 1984) is preferred. This estimator is given by

$$m(t,h_t)=\sum_{l=1}^{n}Y(t_i)\int_{x_{l-1}}^{x_l}\frac{1}{h_t}K\left(\frac{t-u}{h_t}\right)du$$

where h_t is smoothing parameter called bandwidth in point *t* and limits of integration are given by $x_0 = a$, $x_l = 0.5(t_{l+1} + t_l)$ for $l = 1, ..., n - 1$, $x_n = b$. *K* denotes kernel function of order k, $k \geq 2$ (Gasser *et al.*, 1985).

For regression function estimation the Epanechnikov kernel (Gasser *et al.*, 1985) which has the property of optimal kernels and which is commonly used in practise is preferred.

The smoothing parameter can be estimated both locally and globally. Considering global bandwidth the plug-in estimate (Gasser *et al.*, 1991) of the parameter is preferred. Such an estimate has lower variability than cross-validation estimators and is constructed such that Mean Integral Squared Error (MISE) is minimised. Note that since MISE is defined as the integral over non-negative function (Herrmann, 1997), the order of integration and mean can be reversed and denoted as Integral Mean Squared Error (IMSE).

The optimal local bandwidth, which is defined as the minimizer of the Mean Squared Error (MSE) (Herrmann, 1997) of the estimate of the regression function is estimated using local plug-in algorithm (Herrmann, 1997). The algorithm is iterative. During the first $(k + 1)(2k + 1)$ iterations a sequence of global bandwidths minimizing the Integral Mean Squared Error (MISE) (Herrmann, 1997) is generated and local bandwidth minimizing MSE is estimated in the last iteration.

Robust repeated median regression

Robust repeated median regression estimates the smoothed data in a given time instant by a linear trend which is estimated using repeated medians based on the surrounding noisy observations.

Given the measurements *Y*(1), ..., *Y*(*n*), robust repeated median regression is based on model (Gather, Fried, 2004)

$$Y(t)=m(t)+\sigma(t)\varepsilon(t)+\eta(t), \quad i=1,\ldots,n$$

where σ(*t*) denotes standard deviation expressing the variance of *Y*(*t*) and ε(*t*) are random errors with zero mean and unit variance. Although that σ and ε have the same interpretation as σ and ε for kernel regression, the algorithm for their estimation is different from the algorithm used in kernel regression estimation. Function $\eta(t)$ represents the outlier process which is responsible for sudden changes in mean and variability of the analysed variable *Y*. Note that the outlier process is zero most of the time and occasionaly exhibits large absolute values. To estimate the smoothed data, σ(*t*) and $\eta(t)$ are supposed to vary smoothly in time (Fried, 2004).

Supposing a parameter $h_t > 0$ the robust repeated median regression estimate of the smoothed data within a smoothing window {t – h_t, ..., $t + h_t$}, $0 < h_t \leq t$ is given by (Fried, 2004)

$$m(t+i)=\beta_1^t+i\beta_2^t,\ i=-h_t,\ldots h_t$$

where β_1^t and β_2^t is intercept and slope of the regression line within a smoothing window {t – h_t, ..., $t + h_t$}. Given a set of observation $Y(t - h_{t)}, ..., Y(t + h_{t)}$ corresponding to window {t – h_t, ..., $t + h_t$}, the estimate of parameters β_1^t and β_2^t is defined by (Siegel, 1982)

$$\hat{\beta}_1^t=\text{med}_{i\in\{-h_t,\ldots h_t\}}\left\{Y(t+i)-i\hat{\beta}_2^t\right\},$$

$$\hat{\beta}_2^t=\text{med}_{i\in\{-h_t,\ldots h_t\}}\left\{\text{med}_{i\neq j}\frac{Y(t+i)-Y(t+j)}{i-j}\right\}$$

To improve the robustness of the method the fit in the current time window is extrapolated to the next time and used for the evaluation of the outlyiness of the next measurement. The outlyiness of the next measurement is evaluated by comparing the size of exrapolating residual with the appropriately chosen multiple of the estimate of standard deviation corresponding to observations in the current window (Fried, 2004). In case that the next observation is evaluated as outlier it is trimmed or shrinked to prevail its impact on the local fit. This way the outliers are detected and treated prior to moving smoothing window.

The residuals are also used for the detection of level shifts. As described in (Fried, 2004), in case that level shift occurs in the smoothing window, the algorithm for computing the fit is restarted.

The procedure for robust repeated regression together with the rules for outlier and shift detection is described in detail in (Fried, 2004).

The window width h_t can be chosen globally or locally based on the adaptation described in (Gather and Fried, 2004). Here the recommendations given in (Gather and Fried, 2004) and adaptation algorithm (Borowski and Fried, 2011) are preferred.

Data smoothing using discrete Fourier transform

The principle of data smoothing using discrete Fourier transform (DFT) is the conversion of observed values from its original (usually time) domain to a representation in the frequency domain, where significant frequencies occuring in the observations are filtered. Subsequently the data is transferred back to time domain.

The transform is based on expressing the observations *y*(0), ..., *y*(*n* – 1), where *n* is supposed to be a power of two, as a weighted sum of orthonormal basis functions. These basis functions, which are formed by sine and cosine functions of different frequencies, are given by

$$c_k(i)=\cos(2\pi ki/n),\quad s_k(i)=\sin(2\pi ki/n)$$

where *i* = 1, ..., *n* – 1 and parameter *k* = 0, ..., *n*/2 determines the frequency. Using the Euler's formula the discrete Fourier transform of the observations *y*(*t*), *t* = 0, ..., *n* – 1, is given by (Proakis, 1995)

$$X(k)=\sum_{i=0}^{n-1}y(i)e^{-j2\pi ki/n}$$

and the conversion back to time domain is performed using the relation

$$y(i)=\frac{1}{n}\sum_{k=0}^{n-1}X(k)e^{j2\pi ki/n}$$

where *j* is complex unit. The principle of data smoothing is based on the assumption that high values of transform coefficients correspond to important changes in the sequence of the observations while low values of wavelet coefficients correspond to random errors. To suppress the random errors the transform coefficients are thresholded.

The idea of thresholding is to remove transform coefficients that are smaller than appropriately chosen threshold δ. In practise the thresholding is performed using soft thresholding function or hard thresholding function. Hard thresholding function sets all coefficients smaller than threshold value to zero and the remaining coefficients are left unchanged. Using soft thresholding function the coefficients below the threshold value are set to zero and the values of remaining coefficients are reduced.

Of course, the choice of the threshold value is important. Here the universal threshold, which belongs to widely used ones and which is easy to implement, is preferred. The universal threshold is defined as (Donoho, Johnstone 1995)

$$T=\sigma\sqrt{2\log(n)}$$

where σ represents the noise standard deviation which can be estimated as the absolute median deviation of the transform coefficients.

Data smoothing using discrete wavelet transform

The principle of data smoothing using discrete wavelet transform (DWT) is analogous to data smoothing using DFT. It means that the observed values are converted from its original domain to a time-frequency domain, subsequently significant frequencies are thresholded and finally the data is transferred using inverse transform back to time domain.

Similarly to DFT the DWT is based on expressing the observations $y(t)$, $i = 1, ..., n$, as a weighted sum of orthonormal basis functions of different frequencies. A contrary to DFT, where the base of functions is formed by sine and cosine functions, several wavelet families, forming the base of functions exist. These wavelet families – basis functions of the space $L^2(\mathbb{R})$ (the space of all measurable functions $f(t)$ satisfying $\int_{-\infty}^{\infty} |f(t)|^2 dt < \infty$, where $\mathbb{R}$ denotes the set of real numbers), which are concentrated in time, are derived from parental wavelets, namely from mother wavelet $\psi(t) \in L^2(\mathbb{R})$ and father wavelet $\phi(t) \in L^2(\mathbb{R})$ (Meyer, 1992). The relations, based on which the child wavelets (forming the orthonormal basis of space $L^2(\mathbb{R})$) are derived from parent wavelets, can be written as

$$\psi_{j,s}(t) = \sqrt{2^j}\psi\left(2^j t - s\right), \quad \phi_{j,s}(t) = \sqrt{2^j}\phi\left(2^j t - s\right)$$

where $s \in \mathbb{Z}$ ($\mathbb{Z}$ denotes the set of integers) is parameter of time shift and j is parameter corresponding to scale.

Given the parameter $J \in \mathbb{Z}$ determining the maximum resolution, the sequence of measurements $y(t) \in L^2(\mathbb{R})$ can be expressed as

$$y(t) = \sum_{s \in \boldsymbol{Z}} c_{0,s}\phi_{0,s}(t) + \sum_{0<j<J, j \in \boldsymbol{Z}} \sum_{s \in \boldsymbol{Z}} d_{j,s}\psi_{j,s}(t), \tag{1}$$

where discrete wavelet coefficients representing the weights of basis functions are given by

$$c_{j,s} = \int \quad y(t)\phi_{j,s}(t)dt, \quad d_{j,s} = \int \quad y(t)\psi_{j,s}(t)dt.$$

Using the relation (1) the sequence of observations $y(t)$, $t = 1, ..., n$, can be expressed as a "smooth" (approximative) part generated by child wavelets derived from father wavelet and detailed part generated by child wavelets derived from mother wavelet. Therefore the coefficients $c_{j,s}$ are called aproximative while the coefficients are called $d_{j,s}$ detailed coefficients.

For the analysis presented in this paper orthonormal Daubechies 8 (db8) wavelets (Daubechies, 1992) were used.

The DWT as well as backward reconstruction can be realized in $0(n)$ steps using algorithm (Mallat, 1998), which is based on lowpasss and highpass filters. The thresholding of DWT coefficients, which is performed analogous to thresholding of coefficients obtained using DFT, is applied only for detailed coefficients.

Trend filtering

Trend filtering estimate belonging to nonparametric regression estimates is constructed based on penalised least squares criterion, where the penalty term penalizes the changes in the discrete derivative of the estimate. The penalty term is based on l1 norm, which encourages sparsity of the discrete derivatives.

Given the measurements $Y(1), ..., Y(n)$ observed at n discrete equally spaced time instants (the extension for arbitrarily spaced time instants is described in (Tibshirani, 2014), the regression model can be written as (Kim *et al.*, 2009)

$$Y(t) = m(t) + \varepsilon(t), \quad i = 1, \ldots, n$$

where $\varepsilon(t)$ are independent errors. Denoting $\boldsymbol{Y} = (Y(1), ..., Y)(n))$ and supposing an integer $k \geq 0$, the k th order trend filtering estimate of $\boldsymbol{m} = (m(1), ..., m)(n))^{\mathrm{T}}$ can be written in the form (Kim *et al.*, 2009)

$$\hat{\boldsymbol{m}} = \arg\min_{\boldsymbol{m} \in \mathbb{R}^n} \frac{1}{2} \parallel \boldsymbol{Y} - \boldsymbol{m} \parallel_2^2 + \lambda \parallel \mathrm{D}^{(k+1)}\boldsymbol{m} \parallel_1,$$

where $\lambda \geq 0$ is a tuning parameter and $D^{(k+1)}$ represents discrete difference operator. As can be seen, the aim is to estimate the parameters such that the penalised sum of least squares is minimised.

For $k = 0$, $D^{(1)} \in \mathbb{R}^{(n-1)xn}$ is the first-order difference matrix

$$D^{(1)} = \begin{bmatrix} -1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 1 & & 0 & 0 \\ & \vdots & & \ddots & & \vdots \\ 0 & 0 & 0 & \cdots & -1 & 1 \\ & & & & & \end{bmatrix}$$

and the 0th order trend filtering reduces to the 1-dimensional fused lasso estimator (Tibshirani *et al.*, 2005), which is also called 1-dimensional total variation denoising (Rudin *et al.*, 1992).

For $k > 0$ the operator $D^{(k+1)}$ is defined recursively by

$$D^{(k+1)} = D^{(1)} D^{(k)}$$

Since the penalty term penalizes the discrete $(k + 1)$st derivative of the vector $\boldsymbol{m}$, the resulting trend filtering estimate $\hat{\boldsymbol{m}}$ has the structure of a piecewise polynomial of order k with data-based adaptive choice of knots.

The trend filtering estimate $\hat{\boldsymbol{m}}$ can be found using a Primal-Dual Interior-Point Method (Kim *et al.*, 2009) or using a path algorithm (Tibshirani, Taylor, 2011) which is preferred here.

Software

The computation was performed in the software R, v. 3.3.1, using packages „lokern" (Herrmann, 2014), „robfilter" (Fried *et al.*, 2014), „smooth"

(Svetunov, 2017), „genlasso“ (Taylor and Tibshirani, 2014) and „waveslim“ (Whitcher, 2015).

Data

The smoothing methods are applied to hourly mass concentrations of atmospheric aerosol (particulate matter, PM) with aerodynamic diameter of particles smaller than 10 µm, namely PM_{10}. The concentrations of PM_{10} were recorded with time resolution 1 hour at 5 monitoring stations situated in Brno, Czech Republic and operated by Brno City Municipality (BCM). The analysed datasets were provided by Brno City Municipality.

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 residential heating, traffic and industrial sources. For the purpose of the illustration of the performance of the methods, observations measured at station Lany were selected. The monitoring station Lany is located on the southern edge of the Bohunice housing estate. There is a motorway leading 450 m south from the location point and the surroundings of the station consists mostly of apartment buildings and full-grown vegetation.

As shown in (Hrdlickova *et al.*, 2008; Hübnerová and Michálek, 2014), the concentrations of the particulate matter is influenced by numerous factors including specific days of the week, heating season, cloud cover and meteorological conditions (temperature, relative humidity). Continuous monitoring of chemical composition and concentrations of PM_{10} is important for air pollution investigation.

RESULTS AND DISCUSSION

Real data

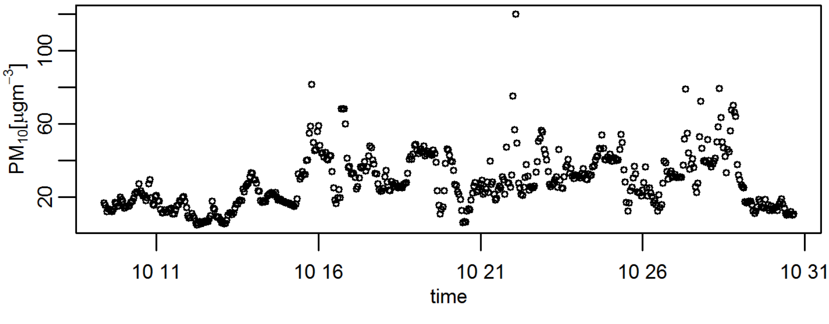
As described in section Data, the PM_{10} concentrations were recorded at 5 different monitoring locations. For a reasonable graphical visualization of the results the reduction of the extensive range of the data is needed. For this reason in this paper the concentration is paid only on data from Lany station and the period from October 9th to October 30th 2016. The corresponding sequence of measurements contains evidently outlying values and can be thus considered as a representative segment of the observations.

The analysed concentrations of PM_{10} aerosols are illustrated in Fig. 1, which shows that several steady increases and sharp declines in PM_{10} concentration occur over the studied period. As already mentioned, a few observations obviously deviated from the other measurements are present in the data during the studied measurement period.

Manual data control performed by specialised operators from Council of the City of Brno stated that all changes in the variability of PM_{10} concentrations are associated with a change in meteorological conditions. The only exception are four measurements in the night from the 21st to 22nd of October, which were evaluated as invalid, because they show large and during night hours very unlikely deviations from the rest of the values. However, inspection of the station logbook did not clarify the reason for the presence of these outliers.

The data presented in Fig. 1 were smoothed using the methods presented in section Materials and methods. As already described, individual methods are parametrised by a parameter specific to the corresponding methodology. Remember that the approaches used for the selection of individual parameters were described in the section Materials and methods. For computing moving average estimate several different window sizes were considered. The results presented in Fig. 2 are based on window width $h = 25$.

The PM_{10} concentrations together with the smoothing lines obtained using individual methods are visualised in Fig. 2. All data was analysed together, however, for graphical visualisation of the results the considered time period was partitioned into several segments and corresponding estimates were visualised in individual graph with different



1: *Concentrations of PM_{10} aerosols* range of y-axis. Since each graph corresponds to relatively short time interval the observations are plotted against the index of measurements.

As the figure shows the results obtained using trend filtering and kernel regreression based on both local and global bandwidth are quite comparable. It can also be seen that the variability of smoothing lines (considered for each smoothing line separately) obtained using exponential smoothing, kernel regression, trend filtering and DWT adapts to the data better than using the remaining methods. This phenomenon is obvious especially in local extremes of PM_{10} concentrations where the smoothing lines have sharp peaks.

The figure illustrates that the smoothing lines obtained using trend filtering and kernel regression vary from noisy curve detaily explaining variability in the data to smooth curve in time instants where the variability of the data is relatively small. While the estimate of the smoothed data based on DFT is quite variable, the smoothing lines obtained using moving averages and robust repeated median regression appear to be realively smooth.

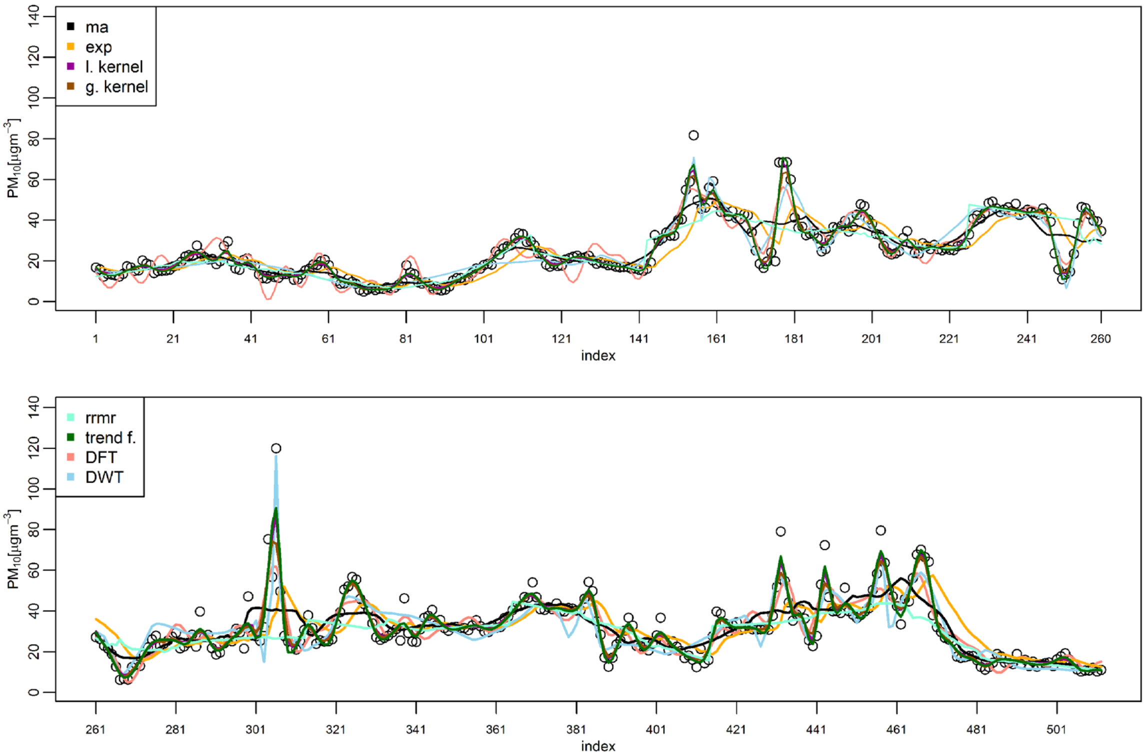
It can also be seen that the DWT based estimate is most influenced by outlier observation corresponding to index = 308. Inspecting the segment corresponding to index 430 – 470, where several outlying observations are present, it can be concluded that moving average method and robust repeated median regression does not at all appear to be influenced by outliers.

Simulated data

To compare the performance of individual smoothing methods two simulation studies were performed. For both studies the data was simulated based on the function defined on the interval [0 ; 0.85] and visualised in Fig. 3 together with the smoothing residuals generated from normal distribution with zero mean and standard deviation σ, where σ = 0.2, 0,3, ..., 0.8. For each value of σ 500 replicates was used.

The first simulation study was performed on datasets free of ouliers. However, for the purpose of the second simulation study six intentional outliers were assigned at random positions in the data set. These outliers were generated from the normal distribution $N(ky(x_i), \sigma^2)$, where k was generated from uniform distribution *unif* (3,6).

The simulated data was smoothed using the methods described in section Materials and methods and the smoothing parameters were estimated based on approaches introduced in the description of individual methods. Besides that different values of the smoothing parameters were tested. However, with regards to the amount of methods being compared only the results based on smoothing parameters chosen using approaches given in section Materials and methods



2: *Concentrations of PM10 aerosols together with the smoothing lines obtained using: moving average (ma), exponential smoothing (exp), local kernel regression (l. kernel), global kernel regression (g. kernel), robust repeated median regression (rrmr), trend filtering (trend f.), discrete Fourier transform (DFT) and discrete wavelet transform (DWT)* (which resulted in the best results) were presented. Specifically, the moving average method was performed using window size of length 25.

On the base of (Tibshirani, 2014), the performance of the smoothing procedures was evaluated by average squared error, which is defined as

$$\frac{1}{n}\sum_{i=1}^{n}\left(\hat{m}(x_i)-y(x_i)\right)^2$$

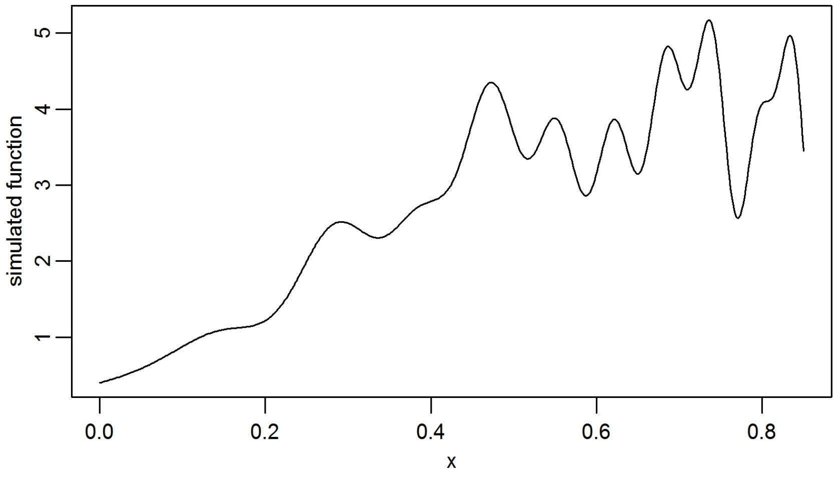
where (x_i) represent the smoothed data estimate and $y(x_i)$ is the original function.

The mean of average squared error obtained using different smoothing methods is plotted against the constant σ in Fig. 4 for the first simulation study and in the Fig. 5 for the second simulation study including outliers.

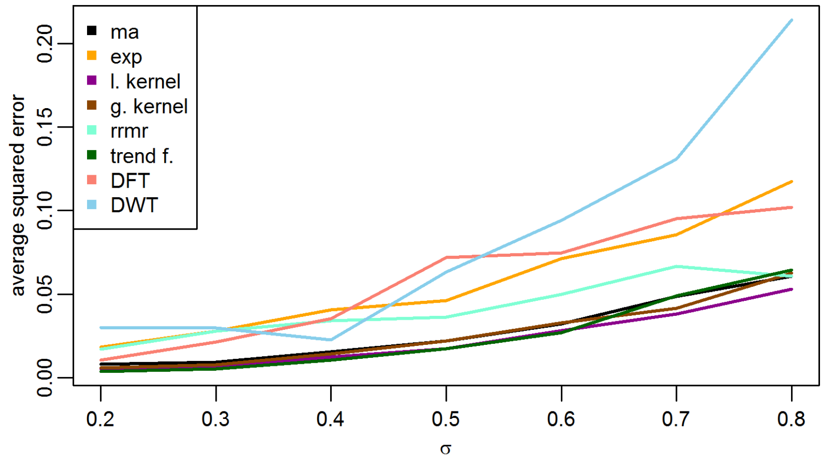
As the Fig. 4 and Fig. 5 show, the best results were obtained using kernel regression based on both local and global bandwidth, trend filtering and the moving average method. As expected, higher precision of the smoothing methods (smaller average mean squarred error) was achieved by applying the procedures on the datasets free of outliers. It can also be seen that the average squared error increases with increasing variance of the residuals used for the generation of the datasets for all presented procedures.

While the precision obtained using kernel regression, trend filtering and moving average is quite comparable, the results obtained using the remaining methods differ. Considering datasets free of outliers and σ > 0.5, the least accuracy was obtained using exponential smoothing, DWT and DFT based approach. As the Fig. 5 shows, the same conclusion can be stated for datasets containing outliers.

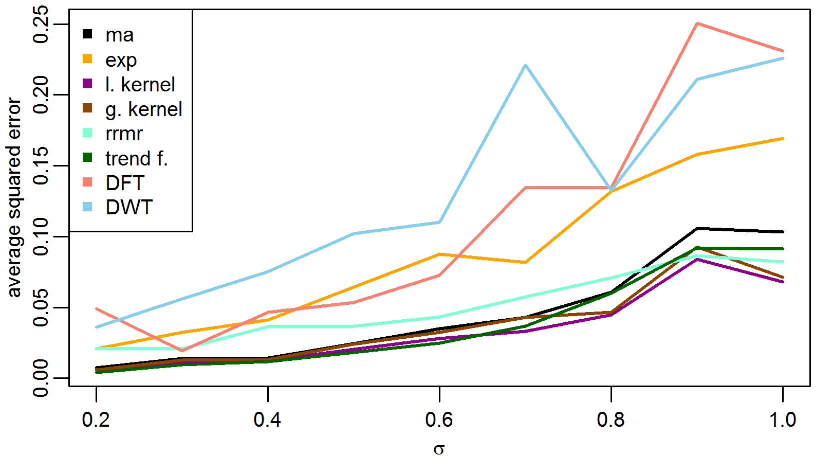
The results obtained using DFT was expected, since Fourier transform assumes periodic signals,



3: *Function used for the simulation studies*



4: *The results of the first simulation study: The mean of average squared error obtained using different smoothing methods plotted against the constant . The smoothing methods are: moving average (ma), exponential smoothing (exp), local kernel regression (l. kernel), global kernel regression (g. kernel), robust repeated median regression (rrmr), trend filtering (trend f.), discrete Fourier transform (DFT), discrete wavelet transform (DWT)*



5: *The results of the second simulation study: The the mean of average squared error obtained using different smoothing methods plotted against the constant . The smoothing methods are: moving average (ma), exponential smoothing (exp), local kernel regression (l. kernel), global kernel regression (g. kernel), robust repeated median regression (rrmr), trend filtering (trend f.), discrete Fourier transform (DFT), discrete wavelet transform (DWT)*

which is not true for the data presented in this paper. The violation of the periodicity assumption results in a bad fit of the smoothed curves.

Note that the analysis of a data file containing outliers is quite often problematic. Considering e.g. DWT, the outliers create significant detail coefficients at finest scale, which means that they cannot be simply thresholded. The situation with DFT is similar, since the outliers destroy the spectrum.

Based on the results obtained by analysing real and simulated data it can be concluded that the best results were obtained using kernel regression and trend filtering. The reason of this fact is that the corresponding smoothing lines adapt to the data better than the smoothing lines estimated using remaining procedures. Also the average squared errors computed based on the smoothing lines obtained using kernel regression and trend filtering are relatively small.

Of course the quality of the estimates is significantly influenced by the choice of smoothing parameters and different selection of the parameter values may result in more or less smooth estimates of the data trend. As already described, the presented results are based on smoothing parameters chosen using the approaches given in section Materials and methods. However, it was verified that by using smoothing parameters values close to the proposed ones the comparable results are obtained.

CONCLUSION

This paper presents the overview and comparison of different smoothing methods that can be used to smooth environmental data. The considered methods, which are applied to smooth PM_{10} concentrations and compared in a simulation study, include kernel regression with both local and global bandwidth, robust repeated median regression, trend filtering, moving average, discrete Fourier transform approach, discrete wavelet transform approach and exponential smoothing.

It was shown that the considered methods differ in their sensitivity to outliers and adaptivity to the data. It is known that outliers, the observations significantly deviated from the other measurements, may have a signifficant effect on data evaluation and modelling. The outliers occur in large environmental datasets quite often and their presence might result from numerous experimental errors, natural variability of the analysed variable, unusual experimental conditions or from abnormal behaviour of the observed variable. For this reason, the choice of the data smoothing technique depends on the application and requirements of the analyst.

Considering the adaptivity to the data and precision evaluated based on mean squared error the best results were obtained using kernel regression and trend filtering.

Of course the shape of all smoothing lines depends on the choice of parameters determining the amount of observations for local smoothing. As already described the results presented here were obtained using concrete values of smoothing parameters. However, by using smoothing parameter values close to the proposed ones comparable results are obtained.

Acknowledgement

The author would like to thank to Geno Reynolds for the proofreading of English.

REFERENCES

ALSBERG, B., WOODWARDS, A. and KELL, D. 1997. An introduction to wavelet transforms for chemometricians: A time-frequency approach. *Chemometrics and Intelligent Laboratory Systems,* 37(2): 215–239.

BARSANTI, R. J. and GILMORE, J. 2011. Comparing noise removal in the wavelet and Fourier domains. In: *IEEE 43rd Southern Symposium on System Theory*. 14–16 March, Auburn, AL, USA, pp. 163–167.

BLEAKLEY, K. and VERT, J. P. 2011. *The group fused Lasso for multiple change-point detection.* [Online]. Available at: https://hal.archives-ouvertes.fr/hal-00602121 [Accessed: 2017, November 13].

BOROWSKI, M. and FRIED, R. 2011. *Robust repeated median regression in moving windows with data-adaptive width selection*. Discussion paper 28/2011, SFB 823. Dortmund: Technische Universität.

BOWERMAN, B. L., O'CONNELL, R. T. and KOEHLER, A. B. 2005. *Forecasting, time series, and regression: an applied approach*. 4th Edition. Belmont, CA: Thomson Brooks/Cole.

CHEN, J. and GUPTA, A. 2012. *Parametric statistical change point analysis: with applications to genetics, medicine, and finance.* 2nd Edition. New York: Birkhäuser.

COOK, E. and PETERS, K. 1981. The smoothing spline: a new approach to standardizing forest interior tree-ring width series for dendroclimatic studies. *Tree-Ring Bulletin*, 41: 45–55.

ČAMPULOVÁ, M., VESELÍK, P. and MICHÁLEK, J. 2017. Control chart and six sigma based algorithms for identification of outliers in experimental data, with an application to particulate matter PM10. *Atmospheric Pollution Research*, 8(4): 700–708.

DAUBECHIES, I. 1992. Ten Lectures on Wavelets. In: *CBMS-NSF Regional Conference Series in Applied Mathematics*. Philadelphia: SIAM.

DAVIES, P. L., FRIED, R. and GATHER, U. 2004. Robust signal extraction from on-line monitoring data, *J. Statist. Plann. And Inference*, 122: 65–78.

DONOHO, D. and JOHNSTONE, I. 1995. Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association*, 90(432): 1200–1244.

FANN, N., LAMSON, A., ANENBERG, S., WESSON, K., RISLEY, D. and HUBBELL, B. 2011. Estimating the national public health burden associated with exposure to ambient PM2 and ozone. *Risk Analysis*, 32(1): 81–95.

FRIED, R. 2004. Robust Filtering of Time Series with Trends. *Journal of Nonparametric Statistics,* 16(3-4): 313–328.

FRIED, R. SCHETTLINGER, K. and BOROWSKI, M. 2014. *robfilter: Robust Time Series Filters*. R package version 4.1. Available at: https://CRAN.R-project.org/package=robfilter [Accessed: 2018, February 15].

GASSER, T. and MÜLLER, H.-G. 1984. Estimating regression functions and their derivatives by the kernel method. *Scandinavian Journal of Statistics*, 11(3): 171–185.

GASSER, T., MÜLLER, H.-G., and MAMMITZSCH, V. 1985. Kernels for nonparametric curve estimation. *Journal of the Royal Statistical Society: Series B*, 47(2): 238–252.

GASSER, T., KNEIP, A. and KOEHLER, W. 1991. A flexible and fast method for automatic smoothing. *Journal of the American Statistical Association,* 86(415): 643–652.

GATHER, U. and FRIED, R. 2004. Methods and Algorithms for Robust Filtering. In: ANTOCH, J. (Ed.). *COMPSTAT 2004: Proceedings in Computational Statistics*. Berlin - Heidelberg: Physika-Verlag, pp. 159–170.

HE, S., FANG, S., LIU, X., ZHANG, W., XIE, W., ZHANG, H., WEI, D., Fu, W. and PEI, D. 2016. Investigation of a genetic algorithm based cubic spline smoothing for baseline correction of raman spectra. *Chemometrics and Intelligent Laboratory Systems*, 152: 1–9.

HENRY. R., NORRIS. G., VEDANTHAM, R. and TURNER, J. 2009. Source region identification using kernel smoothing. *Environmental Science Technology*, 43(11): 4090–4097.

HERRMANN, E. 1997. Local bandwidth choice in kernel regression estimation. *Journal of Computational and Graphical Statistics*, 6(1): 35–54.

HERRMANN, E. 2014. Packaged for R and enhanced by Martin Maechler. *lokern: Kernel Regression Smoothing with Local or Global Plug-in Bandwidth.* R package version 1.1-6. [Online]. Available at: https://CRAN.R-project.org/package=lokern [Accessed: 2018, February 15].

HOLEŠOVSKÝ, H., ČAMPULOVÁ, M. and MICHÁLEK, J. 2017. Semiparametric outlier detection in nonstationary times series: Case study for atmospheric pollution in Brno, Czech Republic. *Atmospheric Pollution Research*, 9(1): 27–36.

HOLT, C.C. 1957. Forecasting trends and seasonals by exponentially weighted moving averages, *International Journal of Forecasting*, 20(1): 5–10.

HRDLIČKOVÁ, Z., MICHÁLEK, J., KOLÁŘ and VESELÝ, V. 2008. Identifcation of factors affecting air pollution by dust aerosol PM10 in Brno city, Czech Republic. *Atmospheric Environment*, 42(37): 8661–8673.

HÜBNEROVÁ, Z. and MICHÁLEK, J. 2014. Analysis of daily average PM10 predictions by generalized linear models in Brno, Czech Republic. *Atmospheric Pollution Research,* 5(3): 471–476.

HUTCHINSON, M. F. 1995. Interpolating mean rainfall using thin plate smoothing splines. *International Journal of Geographic Information Systems*, 9(4): 385–403.

HYNDMAN, R. J., KOEHLER, A. B., ORD, J. K. and SNYDER, R. D. 2008. *Forecasting with Exponential Smoothing: The State Space Approach (Springer Series in Statistics).* Berlin Heidelberg: Springer Verlag.

KAFADAR, K. and MORRIS, M. 2002. Nonlinear smoothers in two dimensions for environmental data. *Chemometrics and Intelligent Laboratory Systems*, 60(1-2): 113–125.

KANEKO, H. and FUNATSU, K. 2015. Smoothing-combined soft sensors for noise reduction and improvement of predictive ability. *Industrial Engineering Chemistry Research*, 54(50): 12630–12638.

KIM, S. J., KOH, K., BOYD, S. and GORINEVSKY, D. 2009. L1 trend Filtering. *SIAM Review*, 51(2): 339–360.

LEE, H., and KANG, K. 2015. Interpolation of missing precipitation data using kernel estimations for hydrologic modeling. *Advances in Meteorology*, 2015: 935868.

LEVIN, D. 2015. Between Moving Least-Squares And Moving Least- l1. *Bit Numerical Mathematics*, 55(3): 781–79.

MALLAT, S. 1998. *A wavelet tour of signal processing*. San Diego: Academic Press.

MAMMEN, E. and GEER, S. 1997. Locally adaptive regression splines. *Annals of Statistics*, 25(1): 387–413.

MEYER, Y. 1992. *Wavelets and Operators*. Cambridge: Cambridge University Press.

NEUBAUER, J. and VESELY, V. 2011. Change point detection by sparse parameter estimation. *Informatica*, 22(1): 149–164.

PAUL, S.K. 2011. Determination of Exponential Smoothing Constant to Minimize Mean Square Error and Mean Absolute Deviation. *Global Journal of Research In Engineering*: 11(3). Available at: https://engineeringresearch.org/index.php/GJRE/article/view/160

PROAKIS, J. 1995. *Digital Communications*. New York: McGraw-Hill, Inc.

RAJMIC, P., NOVOSADOVA, M. and DANKOVA, M. 2017. Piecewise-polynomial Signal Segmentation Using Convex Optimization. *Kybernetika*, 53(6): 1131–1149.

RIDEL, B., GUENNEBAUD, G., REUTER, P., and GRANIER, X. 2015. Parabolic-Cylindrical Moving Least Squares Surfaces. *Computers & Graphics*, 51: 60–66.

RUDIN, L. I., OSHER, S. and FATERNI, E. 1992. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1-4): 259–268.

SANKARAN, S. EHSANI, R. and ETXEBERRIA, E. 2010. Mid-infrared spectroscopy for detection of huanglongbing (greening) in citrus leaves. *Talanta*, 83(2): 574–581.

SIEGEL, A. F. 1982. Robust signal extraction for on-line monitoring data. *J. Statist. Plann. And Inference*, 122: 65–78.

DAVIES, P. L., FRIED, R. and GATHER, U. 2004. Robust signal extraction for on-line monitoring data. *Journal of Statistical Planning and Inference*, 122(1-2): 65–78.

STORLIE, C., BONDELL, H. and REICH, B. 2010. A locally adaptive penalty for estimation of functions with varying roughness. *Journal of Computational and Graphical Statistics*, 19(3): 569–589.

SVETUNOV, I. 2017. *smooth: Forecasting Using Smoothing Functions*. R package version 2.2.0. [Online]. Available at: https://CRAN.R-project.org/package=smooth [Accessed: 2018, February 15].

TAYLOR, B.A. and TIBSHIRANI, R. 2014. *genlasso: Path algorithm for generalized lasso problems*. R package version 1.3. [Online]. Available at: https://CRAN.R-project.org/package=genlasso [Accessed: 2018, February 15].

TIBSHIRANI, R., SAUNDERS, M., ROSSET, S., ZHU, J. and KNIGHT, K. 2005. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B*, 67(1): 91–108.

TIBSHIRANI, R. and TAYLOR, J. 2011. The solution path of the generalized lasso. *Annals of Statistics*, 39(3): 1335–1371.

TIBSHIRANI, R. 2014. Adaptive piecewise polynomial estimation via trend filtering. *Annals of Statistics*, 42(1): 285–323.

TSAY, R. S. 2005. *Analysis of financial time series*. 2nd Edition. Hoboken: Wiley.

TUKEY, J. W. 1977. *Exploratory Data Analysis*. Boston: Addison-Wesely.

UNSER, M. 1999. Splines: A Perfect Fit For Signal And Image Processing. *IEEE Signal Processing Magazine*, 16(6): 22–38.

WALCZAK, B.and MASSART, D. 1997. Noise suppression and signal compression using the wavelet packet transform. *Chemometrics and Intelligent Laboratory Systems*, 36(2): 81–94.

WAND, M.P. and JONES, M.C. 1995. *Kernel Smoothing*. London: Chapman and Hall.

WANG, X., DU, P. and SHEN, J. 2013. Smoothing splines with varying smoothing parameter. *Biometrika*, 100(4): 955–970.

WHITCHER, B. 2015. *waveslim: Basic wavelet routines for one-, two- and three-dimensional signal processing*. R package version 1.7.5. [Online]. Available at: https://CRAN.R-project.org/package=waveslim [Accessed: 2018, February 15].

ZHANG, B., GENG J. and LAI, L. 2015. Multiple change-points estimation in linear regression models via sparse group lasso. *IEEE Transactions on Signal Processing*, 63(9): 2209–2224.

Contact information

Martina Čampulová: martina.campulova@mendelu.cz