

# Detecting outlying observations and structural changes in European air quality data

ETC/ACM Technical Paper 2012/16

Task 1.0.1.2 - Subtask 5b

Mirjam Rehr, Edzer Pebesma, Benedikt Gräler  
mirjam.rehr@uni-muenster.de

May 22, 2013

University of Muenster

Institute for Geoinformatics

# Contents

<b>Contents</b>	<b>3</b>
<b>Summary</b>	<b>4</b>
<b>1 Introduction</b>	<b>5</b>
1.1 Aim . . . . .	5
1.2 Methods . . . . .	5
1.3 Data . . . . .	6
1.4 Software . . . . .	6
<b>2 Preliminaries</b>	<b>7</b>
2.1 Reading DEM and AirBase data into R . . . . .	7
2.2 Descriptive Measures and Graphs . . . . .	7
2.3 Transformations . . . . .	11
2.4 Time series decomposition . . . . .	12
<b>3 Exploratory Statistics for Outlier Detection</b>	<b>13</b>
3.1 Method and Parameterisation . . . . .	13
3.2 Approach . . . . .	14
3.3 Results . . . . .	15
<b>4 Exploratory Statistics for Break Detection</b>	<b>21</b>
4.1 Method and Parameterisation . . . . .	21
4.2 Approach . . . . .	23
4.3 Results . . . . .	24
<b>5 Discussion</b>	<b>29</b>
5.1 Outlier Detection . . . . .	29
5.2 Break Detection . . . . .	30
5.3 Further Remarks . . . . .	31
<b>References</b>	<b>33</b>
<b>Appendices</b>	<b>35</b>
<b>A Outlier Detection</b>	<b>36</b>
A.1 $O_3$ hourly data . . . . .	36
A.2 $SO_2$ hourly data . . . . .	38
A.3 $NO_2$ hourly data . . . . .	40
A.4 $CO$ hourly data . . . . .	42
<b>B Break Detection</b>	<b>44</b>
B.1 $PM_{10}$ hourly data . . . . .	44

B.2	$SO_2$ hourly data . . . . .	45
B.3	$O_3$ hourly data . . . . .	46

## Summary

This paper describes approaches for the detection of outlying observations and structural changes in European air quality data. The study is motivated by potential errors in measuring or reporting air quality time series. Typical artefacts that can occur are outlying observations and break points (structural changes). The responsible institutions are concerned with identifying such inhomogeneities both for causal research and quality improvement. We present continued and revised work based on Gerharz et al. [7] who recommend using moving window statistics for outlier detection and Kolmogorov-Zurbenko filters for break detection.

The aim of this study is to apply these methods to an extended set of data establishing their robustness and suitability, and to identify optimal parameters for the pollutants and temporal resolutions investigated. To this end so-called ground truth data, i.e. information about the presence or absence of inhomogeneities (at any point in time), is essential. This information, however, is not available. For break detection some events are known that might have altered time series from their time of introduction. Both from visual inspection and results from the Kolmogorov-Zurbenko adaptive filter this was only confirmed for one out of seven hypothesized breaks. Consequently, all suggestions we make in the end, next to theoretical considerations, could only be based on subjective judgements rather than on statistical measures.

Both outlier and break detection methods are of explorative nature and do not provide us with any probability statements about identified (i.e. potential) inhomogeneities. The methods are univariate and do not use any further information from neighboring stations or about covariates like station type (traffic or background) or weather conditions and might thus also label 'true' inhomogeneities that do not have any causes in the measuring device or reporting.

Instead of setting fixed thresholds for binary classification we suggest establishing ranks on the likeliness of being an outlier or break point as an alternative approach. Suspected inhomogeneities should be verified by domain experts and best thresholds worked out by operating experience.

For outlier detection we suggest to use the Tukey heuristic as it does not assume the data to follow a particular distribution. Over-detection seems to occur even with narrow window widths. We suggest to make use of local (running) measures of dispersion rather than global ones. Furthermore, it might be sensible to choose higher threshold values to avoid false positives. Assigning threshold (and other parameter) values was not possible due to the lack of ground truth data. Finally, we advise to combine relative outlier detection with absolute thresholds.

Our findings for break detection suggest that the method is rather robust against the choice of iteration parameters, while using narrow window widths clearly increases precision of locating the break. On the other hand, the narrower the window, the more time points get marked as break points. Thus, to avoid over-detection, higher thresholds should be adopted. We also recommend to apply deseasonalisation or adjustment to environmental conditions prior to analysis.

# 1 Introduction

## 1.1 Aim

The European Topic Center on Air pollution and Climate Change Mitigation (ETC/ACM) is concerned with investigating and enhancing the accuracy of the air quality data products they provide. Two typical inhomogeneities that can be found in measurement time series are outliers (potentially caused by measurement errors) and structural changes or breaks (possibly caused by changes in instrumentation or concentration unit). A broad range of methods exists for analysing time series data with or without inhomogeneities (e.g. Rao et al. [11]), but apart from the capability of procedures to handle uncleaned data, the error events themselves and their occurrences are of interest to ETC/ACM.

The aim of this working paper is to revisit the most promising methods recommended by Gerharz et al. [7], and to apply them to data from DEM database (raw data) [5] and AirBase (quality screened data) [6] for selected countries, components, measurement periods and time spans. We furthermore aimed at fine-tuning method parameters for the scenarios selected. As ground truth data turned out to be sparse and imperfect, we restricted this study to robustness testing instead of benchmarking within an extensive simulation study.

## 1.2 Methods

The first step in data analysis is understanding the data by looking at descriptive measures and plots. Exploration and hypothesis building follows quite naturally and directly from these first investigations and the questions at hand. It becomes clear that simple flagging of e.g. values above certain absolute thresholds or of series of constant values can hint at the phenomena of interest.

For outlier and structural change detection, explorative moving window approaches were applied. Moving window filters make use of observations' local neighbourhoods within a time series and own the window width and typically a threshold as parameters. Following the terminology of Gerharz et al. [7] we apply the 'whole window - simple statistics' approach as in Basu & Meckesheimer [2] for the outlier detection and the 'moving average filter' approach as described in Rao & Zurbenko [10] and Zurbenko et al. [15] - also known as the (adaptive) Kolmogorov-Zurbenko filter - for break detection. Both methods were adopted from the earlier work of Gerharz et al. [7] but newly implemented for this task.

Throughout this paper we consider only univariate methods, focussing on single time series and neither making use of time series from other (neighboring) stations nor of time series of other pollutants measured at the same station. Further, we did not include any covariate information.

All methods are taken to be retrospective as data are gathered from the countries once a year; for setting up a real-time surveillance system methods would need to be modified accordingly.

### 1.3 Data

We searched for outlying observations in the European Air Quality data from the DEM (Data Exchange Module) [5] database for the year 2010. The DEM database comprises original, raw monitoring data as delivered by the national representatives to the EEA in Aug-Oct 2011, that did not undergo any screening by the ETC/ACM. Countries considered are Romania (as average representative of quality of reporting), Switzerland (for its high quality of data reporting) and Cyprus (limited set of stations with its typical high  $PM_{10}$  concentrations showing large fluctuations). Pollutants examined are particulate matter  $< 10 \mu m$  ( $PM_{10}$ ) on a daily measurements basis; ozone ( $O_3$ ), sulphur dioxide ( $SO_2$ ), nitrogen dioxide ( $NO_2$ ) and carbon monoxide ( $CO$ ) are analyzed on an hourly basis.

For structural change detection, data were taken from processed and published AirBase data, version 6 (released Feb 2012)[6]. Components and measurement periods are  $O_3$  and  $SO_2$  on a hourly, and  $PM_{10}$  on an hourly and daily basis, countries selected are the Netherlands and Czech Republic in this case and the time interval considered comprises the years 2000 to 2010. The ETC/ACM keeps close contacts with the data suppliers in these countries and they are known for their objective and reliable feedback on data screening results. A selection of a few stations per country was entered into our study. This selection was based on the knowledge of events (e.g. changes in instrumentation) that might have caused structural changes.

### 1.4 Software

For the entire workflow (reading and organising data, descriptive analysis and explorative detection analyses) we used the R software (R: A Language and Environment for Statistical Computing) [9], making use of packages `RODBC` [12] for reading data from DEM, `caTools` [13] for moving window statistics and `kza` [4] for computation of the moving average filter for break detection. For report generation we benefited from the `knitr` [14] package. All implementations are provided as R (and `Rnw`) scripts, digitally available from [http://ifgi.uni-muenster.de/~epebe\\_01/ETC-ACM/subtask\\_1.0.1.2-5b](http://ifgi.uni-muenster.de/~epebe_01/ETC-ACM/subtask_1.0.1.2-5b), and are hence reproducible.

## 2 Preliminaries

In this chapter some technical notes on reading air quality data from DEM and AirBase databases are given, followed by a set of descriptive and explorative analyses. Finally we comment on the possibility to analyse transformed rather than raw measurements and on the potential benefits of time series decomposition prior to analysis.

### 2.1 Reading DEM and AirBase data into R

In DEM, for a given country, hourly and daily measurement values can be found in the tables `raw_data_hour` and `raw_data_day_report`, respectively. Additional information about station type (traffic, industrial, background, unknown) can be accessed from table `station_type` and spatial information about station location (latitude, longitude, altitude) is stored in the table `station`. Both are neglected in this study but could be considered in future analyses.

Data from AirBase are downloadable country-wise. In subdirectories ending in `_rawdata` data are organized station-wise. Files are named according to a scheme specifying among others station code, component code and measurement period. Additional information about station type and location can be accessed from tables `AirBase_country_v6_stations.csv`, again this information is neglected in this study but could be considered in future analyses.

R scripts explaining how to load and organize the DEM and AirBase data as a prerequisite for any further analysis are provided online at [http://ifgi.uni-muenster.de/~epebe\\_01/ETC-ACM/subtask\\_1.0.1.2-5b](http://ifgi.uni-muenster.de/~epebe_01/ETC-ACM/subtask_1.0.1.2-5b). After ODBC drivers have been specified, loading data subsets of interest from DEM (.mdb files) can be achieved by using package `RODBC` [12]. For reading data from AirBase (.csv files) no extra package is required.

### 2.2 Descriptive Measures and Graphs

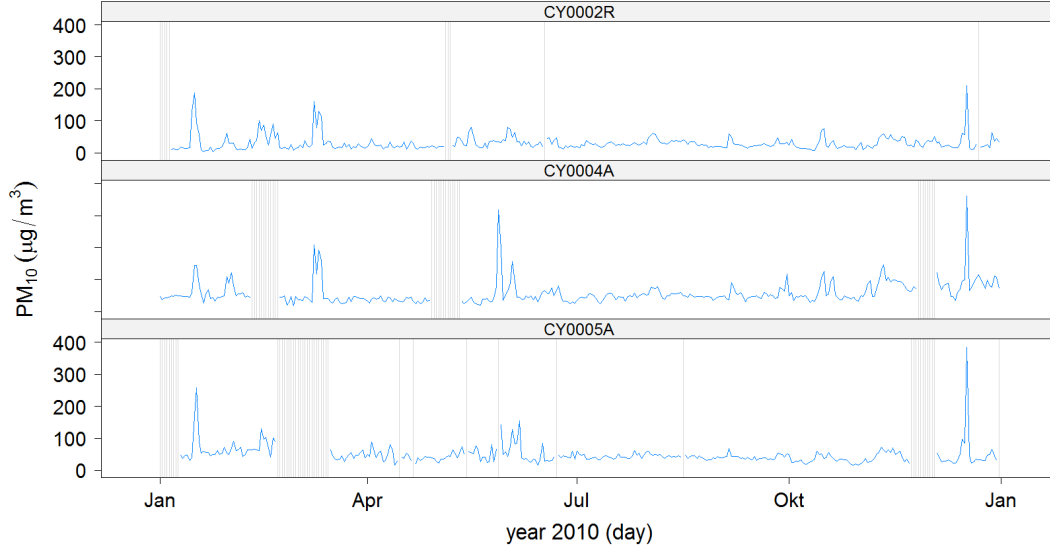
In the following, we distinguish between data from DEM database to test for outliers and data from AirBase to test for structural changes.

#### DEM data for outlier detection

The first graph to look at is a time series plot. [Figure 2.1](#) shows the 2010 DEM timeseries from the three  $PM_{10}$  measurement stations in Cyprus. We will use these data throughout the paper for illustration purposes.

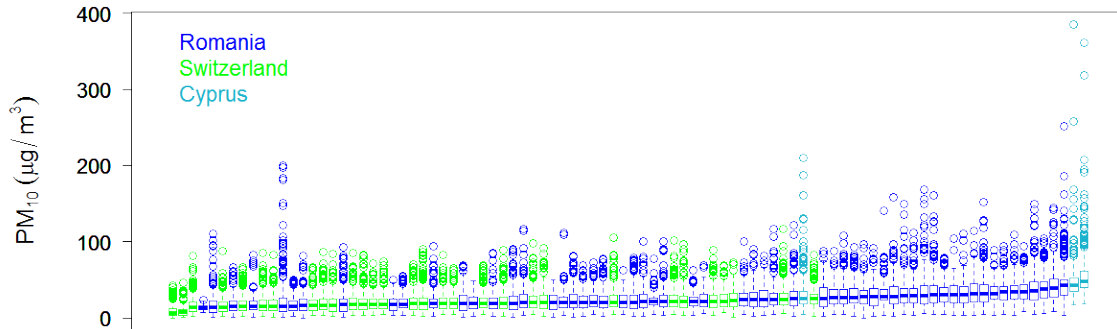
In [Figure 2.1](#) missing values (NA: not available) are shown in grey. We observe single missing values as well as longer intervals without any measurements. The time series do not show a

clear seasonality but at times increased pollutant concentrations are obvious and often coincide between stations, indicating correct measurements caused by a common factor e.g. by a holiday. No matter whether we can find plausible explanations of extreme measurements, in outlier detection we would like to judge if these extremes are abnormal in the sense that we would very rarely expect such observations.



**Figure 2.1:** Cyprus  $PM_{10}$  stations, missing values in grey.

Next, ignoring the serial dependence between measurements that is typical for time series, we look at the distribution of measurements. We show a plot of per station boxplots ordered by median. A boxplot basically depicts the 5-point-summary of the empirical distribution and shows outliers (as circles in Figure 2.2), defined as values more extreme than the so called whiskers, i.e. the most extreme data points still within 1.5 times the inter quartile range (IQR) apart from the quartiles. In Figure 2.2 a boxplot is given for every measurement station considered, with stations ordered by median. Colour coding is used to indicate the country.



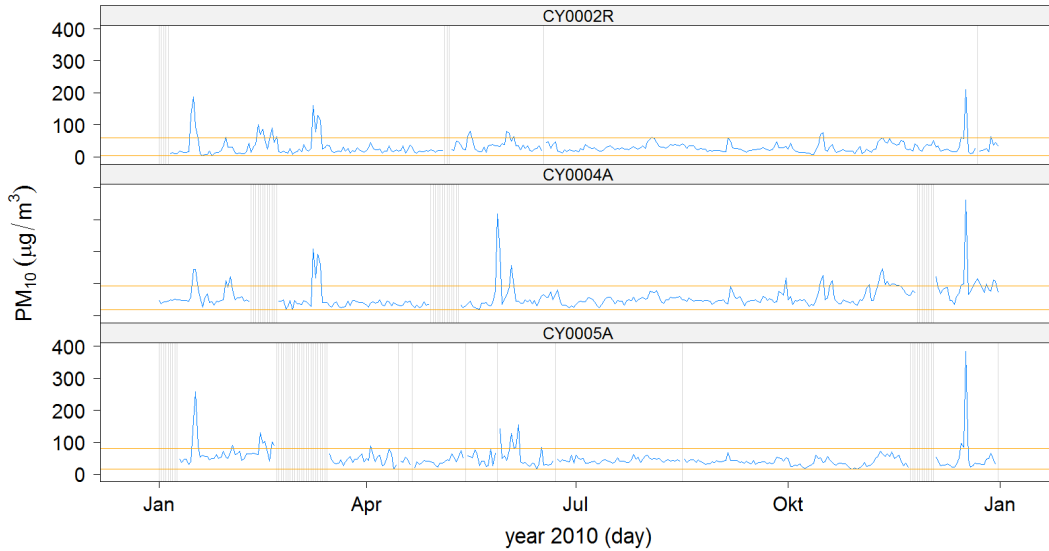
**Figure 2.2:** Boxplots per station, ordered by median.

Here, all observed outliers are high values and in general distributions are skewed to the right, but note that there is a lower boundary as measurements typically are non-negative. At all



three stations in Cyprus and a few stations in Romania we see outlying measurements of very high  $PM_{10}$  concentrations, and two stations in Romania without any boxplot outliers.

For our Cypriot example data Figure 2.3 shows the time series plot where thresholds of potential outliers as identified by the boxplots are indicated by orange lines.



**Figure 2.3:** Cyprus  $PM_{10}$  stations, missing values in grey, boxplot outlier thresholds in orange.

More statistics are shown in the following output where `low` and `high` indicate lower boxplot thresholds and upper boxplot thresholds, respectively. The number of boxplot outliers per station is called `number` and `percentage` is the percentage of boxplot outliers per station.

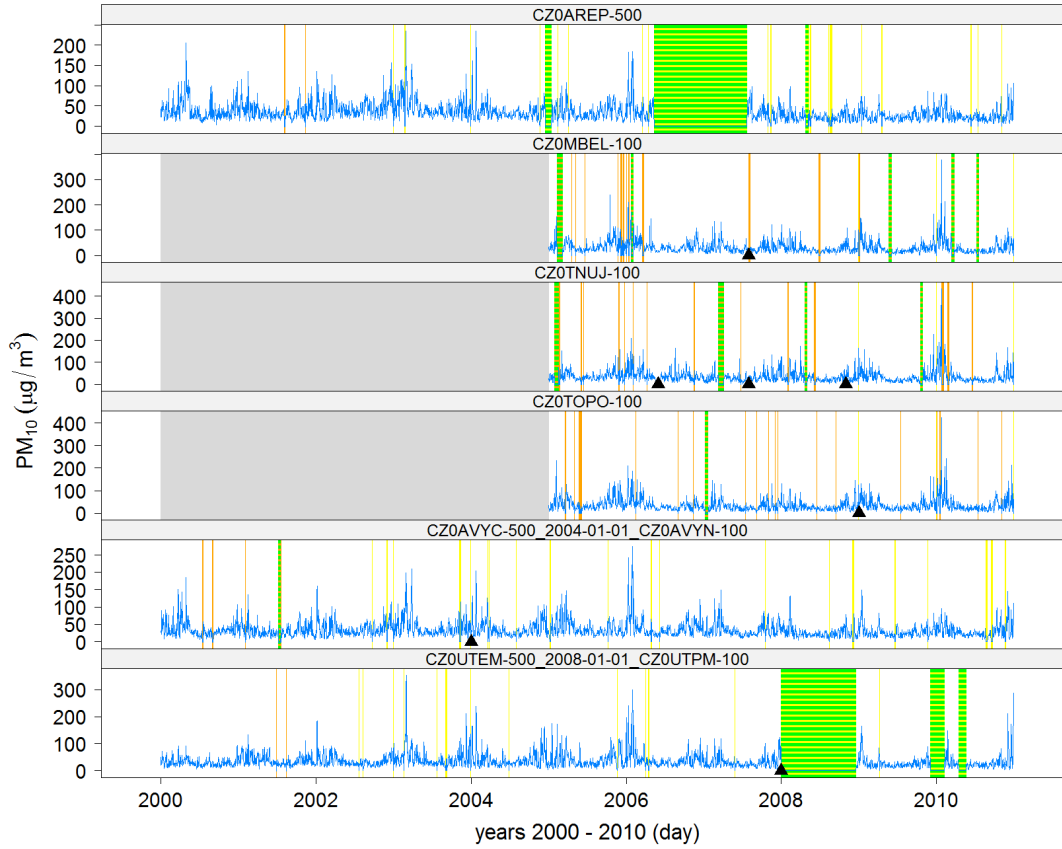
##	CY0002R	CY0004A	CY0005A
## low	4.000	18.100	16.500
## high	60.100	92.600	79.500
## number	24.000	36.000	20.000
## percentage	6.575	9.863	5.479

Summarizing, the boxplot is a good means for analysing empirical distributions especially when one is interested in outliers. No distributional assumption is made, but the autocorrelation of time series is neglected. To overcome this, a moving window approach will be applied in the next chapter.

### AirBase data for break detection

With AirBase data we were interested in break detection rather than outlier detection. Therefore, instead of looking at boxplot thresholds we investigated runs of repeated measurements as a special case of breaks.

Figure 2.4 shows daily  $PM_{10}$  AirBase time series for the years 2000 to 2010 from selected stations in Czech Republic. Hypothesized break points are marked by black triangles. We will use these data throughout the paper for illustration purposes.



**Figure 2.4:** Selection of  $PM_{10}$  stations from Czech Republic. Missing values are shown in grey, zero values in yellow, negative values in orange, and runs of equal values of length 10 or longer in green. Black triangles mark hypothesized break points.

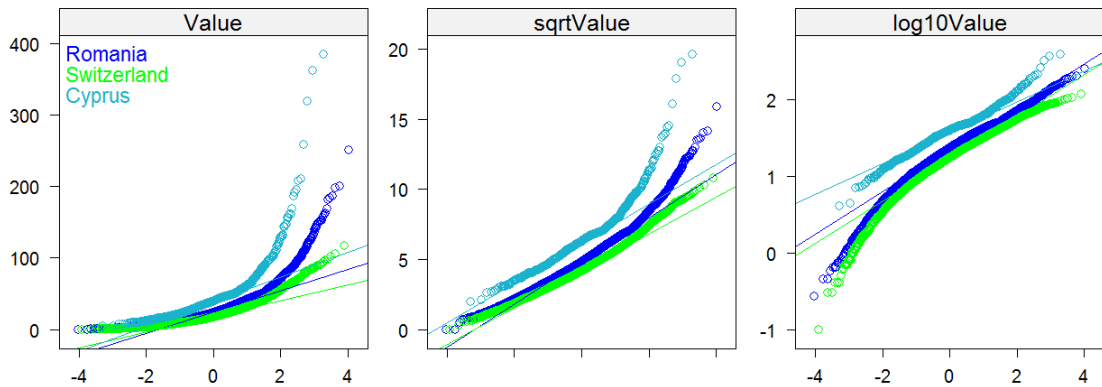
Changes of emission structure as well as of location in microscale are known to have happened for station CZ0AREP within the observation period but dates are not available. For stations CZ0MBEL and CZ0TNUJ changes of emission structure due to new bypass roads had happened in August 2007 and November 2008, respectively. In addition, for station CZ0TNUJ intensive construction activity is reported for the period June 2006 to July 2007. Changes of location in microscale apply also to stations CZ0AVYC (afterwards renamed to CZ0AVYN) and CZ0UTEM (afterwards renamed to CZ0UTPM), where these changes had happened with the beginning of 2004 and 2008, respectively. At station CZ0TOPO a change of the measuring device for  $PM_{10}$  had taken place in the beginning of 2009 and values before that date are said to be probably somewhat underestimated.

For stations CZ0AREP and CZ0UTEM/CZ0UTPM long periods of constant zero measurements stand out. As a simple detection method for such kinds of irregularities we marked all runs of equal values of length 10 or longer (shown in green in Figure 2.4). We also marked all zero and negative values (in yellow and orange, respectively) and it turned out, that all runs of constant values found within these time series concern zero and negative measurement values.

## 2.3 Transformations

Some heuristics in [Chapter 3](#) on outlier detection are based on the normality assumption. As we have seen above, values might be distributed quite asymmetrically and in fact the normality assumption is not fulfilled for neither of the variables considered, with the exception of Ozone. A solution to this problem may lie in transforming the measurement values.

Here we investigate the distribution of original data, their square root transforms and log transforms for the five different variables considered. This is done by comparing quantiles from the empirical distribution to quantiles from a theoretical normal distribution. [Figure 2.5](#) shows the Normal-Quantile plots for the  $PM_{10}$  data. There are obvious differences between the three countries, but note also that there are quite different numbers of observations behind these plots, in particular for Cyprus there are only three stations measuring  $PM_{10}$  and even less for some of the other variables.



**Figure 2.5:** Normal-Quantile plots of original and transformed variables for  $PM_{10}$  data.

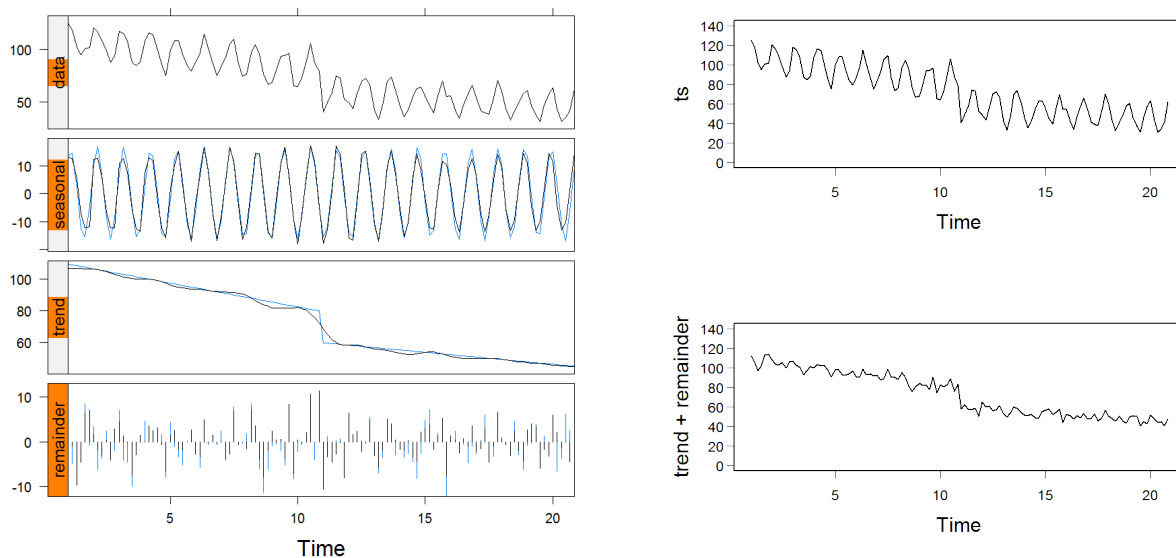
As we are more interested in outlying high values rather than in lower concentration measurements, and have to deal with the restriction of non-negativeness we pay less attention to deviations in the lower quantiles. Note also that negative values (and zero for log) cannot be transformed, leading to less observations in the lower tail.

In general, for untransformed data we see right-skewed distributions, pointed up by light left and very heavy right tail distributions with more extreme high values than expected under normality, i.e. outliers when assuming normality. The square root transform attenuates these effects while the logarithmic transform tends to reverse it with potentially less extreme observations at the upper end than expected under the normality assumption, and much more observations in the lower tail, thus left-skewed distributions.

For the DEM data it turns out that for Ozone the best choice is not to transform the data at all. For the other four pollutants values tend to comply better with normality for log transformed data, at least in the upper tail. The log transform is indeed a common choice in air quality data, where data are supposed to follow a log-normal distribution. Therefore we do not consider the square root transform any further, but use the log transform instead and compare with methods for untransformed values.

## 2.4 Time series decomposition

Structural changes as searched for in [Chapter 4](#) might be difficult to recognize by visual inspection. The reason for this is that measurements are often noisy and may exhibit various forms of periodicity (daily, weekly, yearly) and additionally a flexible trend over time. Time series decomposition aims at separating these different phenomena. A common assumption is that time series build up additively of a trend component, a seasonal (periodical) component and an error component. Decomposition of such a model can be achieved by the STL procedure ('A Seasonal-Trend Decomposition Procedure Based on Loess') suggested by Cleveland et al. [3] and implemented in the R function `stats::stl`. From the derived trend component, breaks should be considerably better visible than from the raw time series. On the other hand, this procedure requires new parameters that have to be set with care. [Figure 2.6](#) shows an example of such a decomposition with simulated data. The inserted break might easily be overlooked in the original time series, but is clearly visible in the deseasonalized series.



**Figure 2.6:** Left: Decomposition of simulated time series: true components in blue, STL decomposition in black. Right: Comparison of simulated time series before and after decomposition and removal of seasonal effects.

## 3 Exploratory Statistics for Outlier Detection

An outlier is an observation that strongly differs from what we expect to observe. Expected values can be estimated by measures of location, deviations by measures of dispersion. Based on these measures, rules for assessing outlying observations can be formed.

Given natural or technical limits that cannot be passed, and where measurements beyond these limits can clearly be identified as false recordings, one can simply use global static thresholds without the need to estimate anything. Typically, such limits are identical for all data, i.e. in our application for all stations from all countries, and at every point in time. However, outliers as defined above would not necessarily be detected. In contrast, a more data driven approach was presented in [Chapter 2](#) where time-constant thresholds were determined on a per station basis using the boxplot definition of outliers.

With the recommendation from Gerharz et al. [\[7\]](#) we now allow for thresholds to vary over time and deal with typical time series characteristics like trends and periodicity.

The outlier detection method recommended in Gerharz et al. [\[7\]](#) is the 'whole window - simple statistics' approach or 'two-sided median method' as described in Basu & Meckesheimer [\[2\]](#) where deviations of measurements from a moving window filter are compared to a threshold statistic. We outline the method and our implementation below.

### 3.1 Method and Parameterisation

A window filter is a statistic computed within a subset of the data, the window. The term is used mainly in time series analysis where the windows are time intervals. A moving window filter then computes the filter statistic for every interval of a given length (the window width) within the study period. Typically, the filter statistics are measures of location like the mean or the median, and are often called running mean and running median.

#### Outlier detection method

For each observation  $x_t$  in a time series the deviation  $|x_t - \text{filter}(t, q)|$  is computed, where  $w = 2 \cdot q$  is the window width and the filter statistic is a measure of location. The deviations are compared to a threshold  $b$  which can be  $b = f \cdot \sigma$ , with  $\sigma$  a measure of dispersion. Given the filter and the threshold rule, the method now depends on two parameters: half the filter window width  $q$  and a factor  $f$  for threshold computation.

Typically, in outlier detection the moving window filter is the running median. Gerharz et al. [\[7\]](#) decided to use the running mean instead as this speeded up their computations without losing much analytical performance. We argue here that the median should be preferred over the mean as we are concerned with flagging outlying observations, i.e. the statistic observations are compared to should not be influenced by outliers, but rather be stable. This kind of ro-

bustness against outliers can be assessed by the breakdown point. The breakdown point gives the maximum percentage of outliers a statistic can handle without breaking down, and can take values between 0 (extremely sensitive to outliers) and 0.5 (extremely insensitive to outliers). The mean has a breakdown point of 0 and is extremely sensitive to outlying observations, while the median is virtually the most robust measure of location with a breakdown point of 0.5. As we are concerned with computing time as well, in our implementation we make use of efficient moving window statistics computation implemented in function `runquantile` from R package `caTools` [13].

Again, the threshold statistics can be absolute values predetermined by the given application as is the case in Basu & Meckesheimer [2]. If no such inherent constraints exist (or are not well known), thresholds are derived from a variability measure  $\sigma$  of the given time series, e.g. the standard deviation ( $sd$ ) or - as we propose here - its robust analogue, the median absolute deviation ( $MAD$ ).

Gerharz et al. [7] found that for hourly  $PM_{10}$  measurements in AirBase the threshold factor should at least be  $f = 6$  while half window widths less than  $q = 10$  time units, i.e. less than 10 hours, performed best. It was hypothesized that parameters probably need not to be adapted for different pollutants. However, careful parameter fine-tuning was recommended for different measurement periods as the best parameters might be quite different for other temporal resolutions.

Indeed in our analysis  $PM_{10}$  measurements were only available on a daily basis. Hourly time series were considered for  $O_3$ ,  $NO_2$ ,  $SO_2$  and  $CO$ . The above mentioned findings will guide us in choosing parameters, but in the end no single true value can be designated as we are lacking ground truth data.

## 3.2 Approach

We based threshold selection on known heuristics and then combined selected thresholds  $b$  with a number of half window widths  $q$ . As we are lacking ground truth data to match, we compared the results (observations that were identified as outliers) in dependence of methods and parameters to each other.

We considered the following heuristics:

**2sd/3sd-rule (z-score)** Under the assumption of normally distributed data the interval  $[mean - f \cdot sd, mean + f \cdot sd]$  contains 95% of the data for  $f = 2$  and 99.7% of the data for  $f = 3$ . I.e. with normally distributed data we can expect the same count of (extreme) outlying observations for every two time series of the same length. If such data were contaminated with outliers, more extreme observations than expected would occur. All observations more extreme than 95% of the data would be judged as outliers, no matter what the true unknown percentage would be.

With  $f = 6$  we get the recommended version of Gerharz et al. [7]. We suggest to replace the mean by the median and the standard deviation by the MAD with normality correction ( $MAD_e = 1.483 \cdot MAD \approx sd_{normal}$ ) for a robust version.

**Modified z-score (Iglewicz & Hoaglin)** This is a similar approach that also assumes the data to follow a normal distribution. Unsuspicious observations lie within the interval  $[med - f \cdot MAD_e, med + f \cdot MAD_e]$  where  $f = 3.5$ . This heuristic is inspired by the z-score

above, and uses robust measures of location and dispersion by default.

**Tukey (boxplot like) heuristic** Outlier determination is conducted as in boxplots: observations that are at least  $f = 1.5$  times the (global) inter quartile range (*IQR*) away from the (running) quartiles are flagged as outliers, and extreme outliers when  $f = 3$ . Instead of the median (0.5 quantile) we investigate the 0.25 and 0.75 quantiles and use the *IQR* as a robust measure of variability. This method allows treating upward and downward outliers differently, which might be a sensible choice when measurement distributions are asymmetric.

We combined the heuristics with the following half window widths:

**Choice of half window width  $q$**  First, for demonstration of the method and for sensitivity analysis, we investigate the method behaviour for extreme values of  $q$ , i.e.  $q = 1$  and  $q = \text{half the length of time series (running median = global median)}$ .

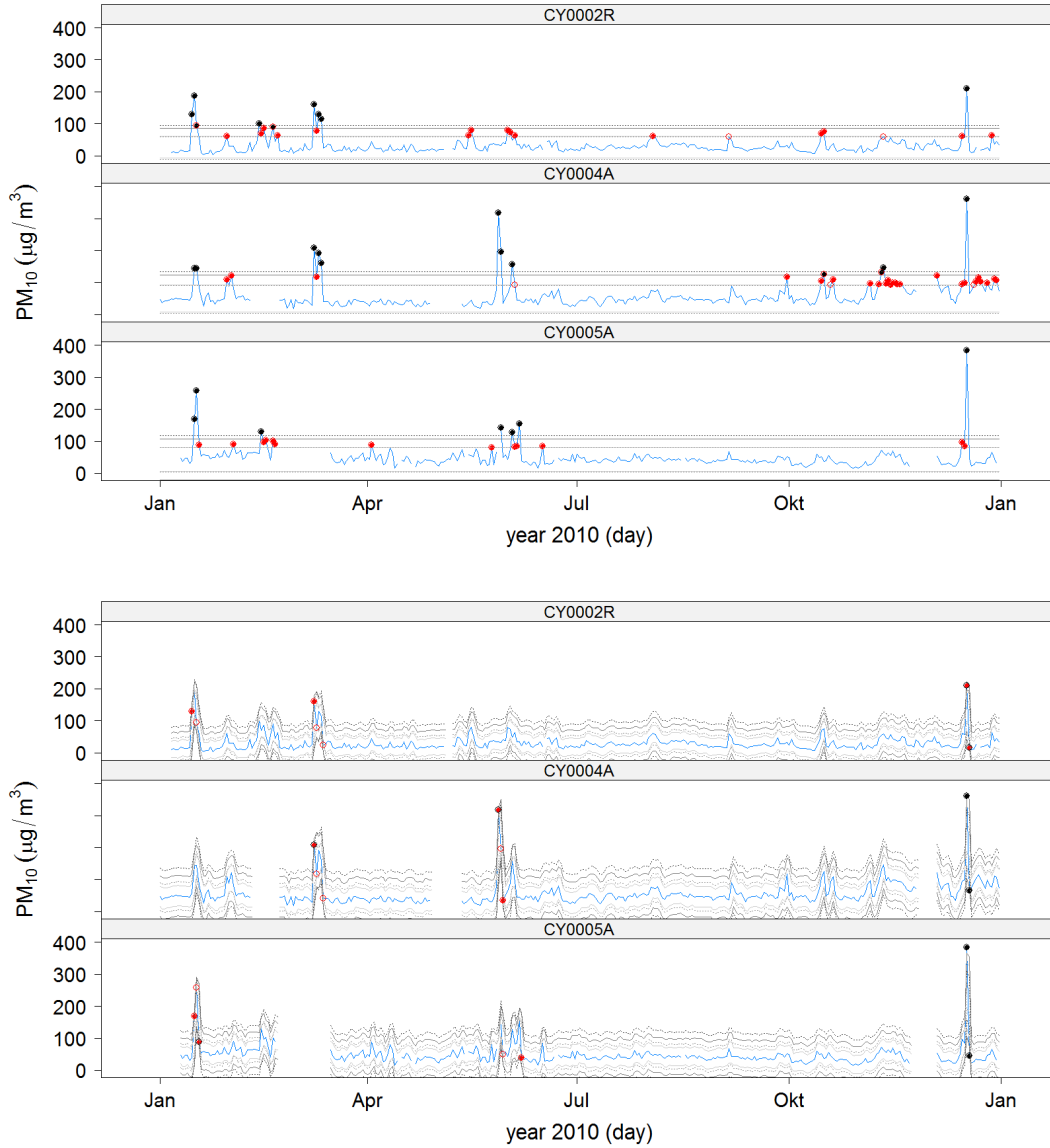
We consider different numbers for the half window width  $q$  for hourly and daily data. The rationale behind that is the following: the higher the short-term variability (as supposed in hourly data in comparison with daily data), the smaller the window has to be chosen in order to depict this variability. For example, with hourly data we might regularly measure lower values at night. If we choose a window width of only a few hours, outliers with high values at night time and low values at day time should be identified, whereas with wider windows high and low values might be flagged irrespective of the time of day.

For daily data, the  $q$ -values considered are 3, 5, 7, 10, 28 and 122, leading to window widths ranging from 7 days up to a maximum window of two third of a year. For hourly data, the  $q$ -values considered are 2, 4, 6 and 8, including the recommendation of  $q = 6$  from Gerharz et al. [7] and a maximum window width of two third of a day.

### 3.3 Results

As an example, we illustrate results for  $PM_{10}$  measurements in Cyprus. [Appendix A](#) contains more figures on outlier detection for stations from Cyprus. All other results can easily be reproduced and investigated using our scripts that are available online and can be found at [http://ifgi.uni-muenster.de/~epebe\\_01/ETC-ACM/subtask\\_1.0.1.2-5b](http://ifgi.uni-muenster.de/~epebe_01/ETC-ACM/subtask_1.0.1.2-5b).

[Figure 3.1](#) shows the thresholds derived from the different heuristics considering the two extreme cases for window size: In the top figure no window is defined, i.e. the whole study period is used. Apart from some minor computational differences the Tukey outlier thresholds should be identical to the boxplot outlier thresholds from [Chapter 2](#). In the bottom figure the window width is chosen to be three, i.e. only the preceding and the following measurement are involved at a time. The plotted thresholds are the 1.5 *IQR* (outlier, red dots) and 3 *IQR* (extreme outlier, black dots) Tukey thresholds (solid lines), and the ones from the 3 *MAD* (red circles) and 6 *MAD* (black circles)  $z$  heuristic (dotted lines). The 1.5 Tukey and the 3  $z$  thresholds are nearly concordant with each other. When doubling the factors, disagreements become more apparent.

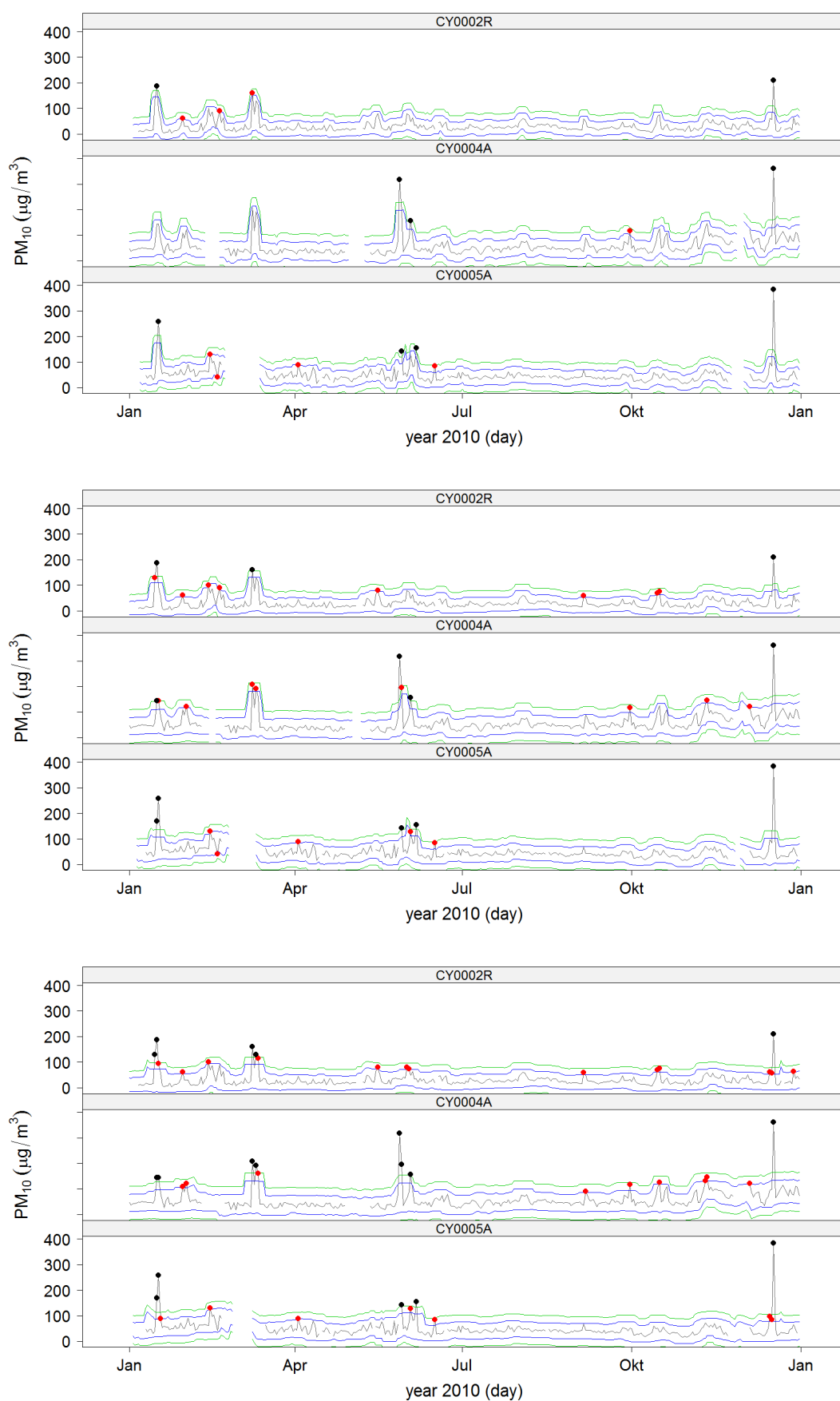


**Figure 3.1:** Original time series data, overview heuristics, no window (top) and extremely narrow window of width 3 (bottom).

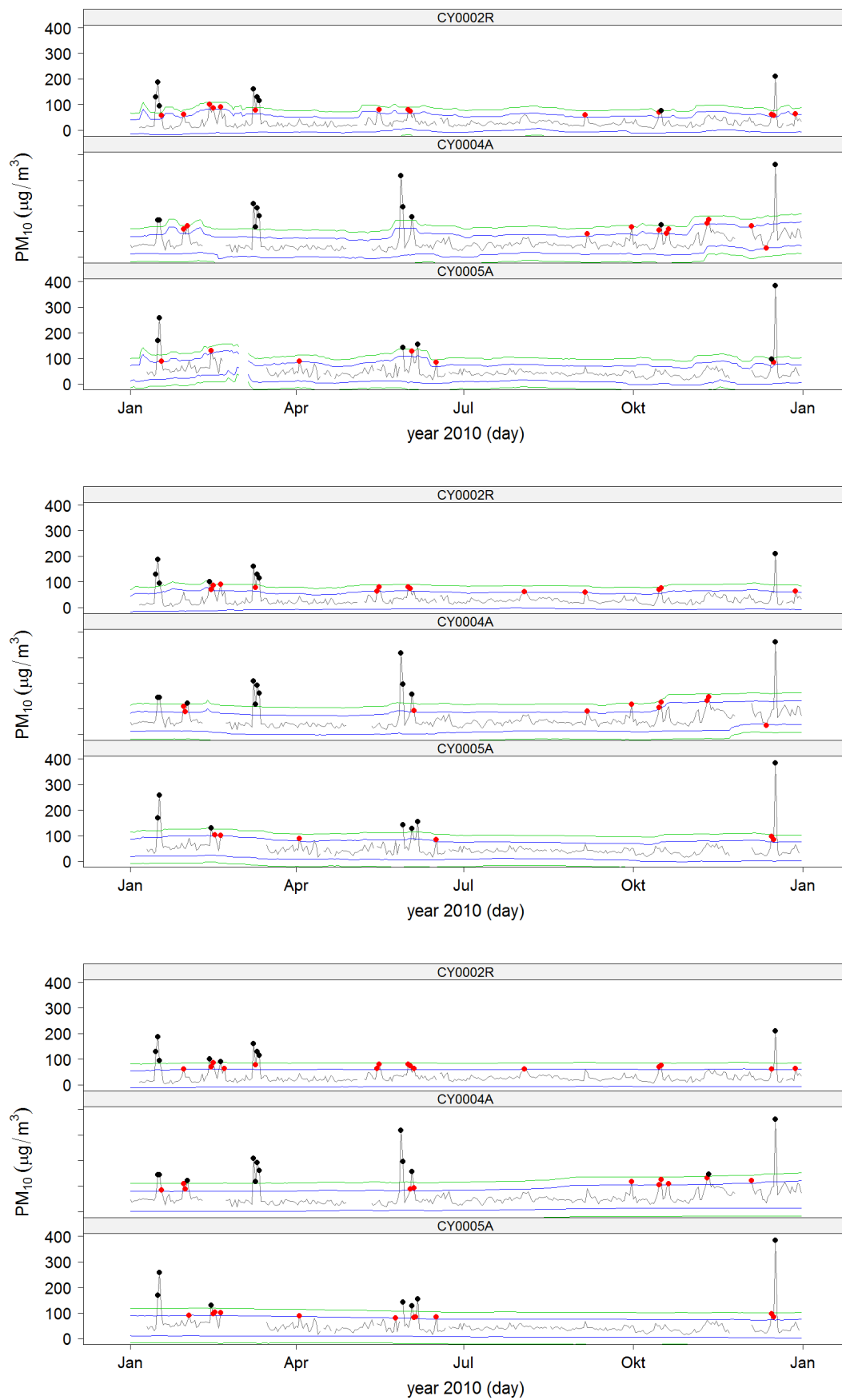
In contrast to choosing a threshold rule, window size obviously matters. Figure 3.2 and Figure 3.3 show window width dependence of outlier detection for  $PM_{10}$  daily original data measurements with thresholds derived from Tukey heuristics. Blue lines indicate 1.5 IQR thresholds, green ones the 3 IQR thresholds. Outliers are depicted by red dots, extreme outliers by black ones. The window widths considered are one week, 11 days, 15 days, three weeks, eight weeks and eight months as stated above.

Clearly, the larger the window the more observations are identified as potential outliers. For the extremely narrow window of width three (Figure 3.1, bottom) the outlier detection virtually becomes a break detection method, where after an extraordinary high measurement, say, a subsequent 'normal' measurement might get marked, too. This way, transient changes can be identified, too, and it will not necessarily be the highest (or lowest) value within this period, that gets marked. Using narrow windows outliers might be missed, while on the other hand large windows result in (potentially massive) overdetection.



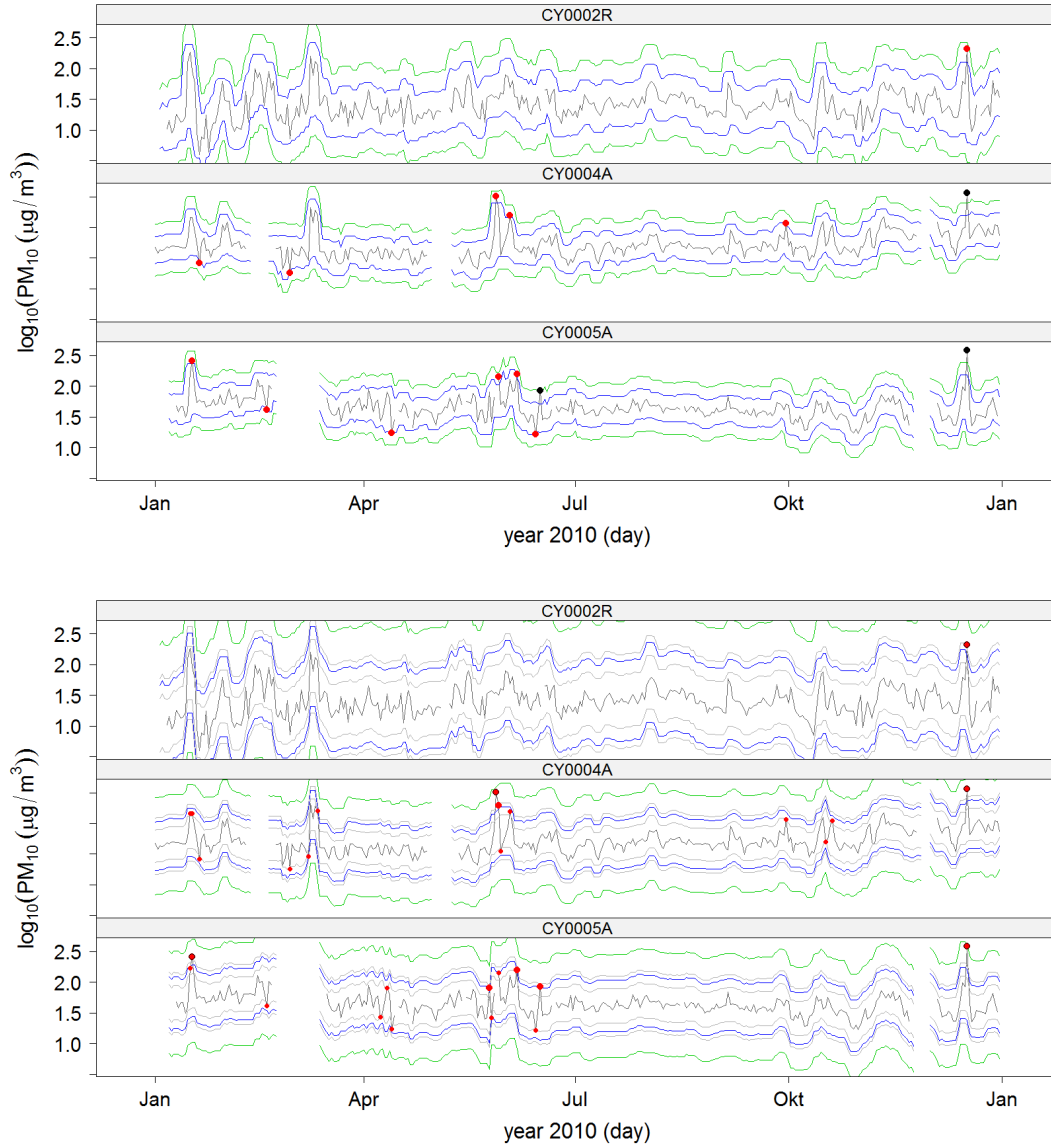


**Figure 3.2:** Original data, Tukey heuristic. From top to bottom: Window width = 7 days, 11 days and 15 days.



**Figure 3.3:** Original data, Tukey heuristic. From top to bottom: Window width = three weeks, eight weeks and eight months.

Next, we present one more threshold rule comparison for a window width of seven days. As opposed to the former plot this time we consider transformed data for the z rule (and also look at the Tukey rule with adapted IQR.) The Tukey case for original data is depicted in the top of Figure 3.2, the Tukey and z-score case for transformed data in Figure 3.4. The upper panel plot can be read as the preceding ones. The lower one uses 2 (grey lines), 3 (blue), 3.5 (grey) and 6 (green) as factors  $f$  for threshold generation. Outliers are marked by small red dots, big red dots, black circles and black dots, respectively. Not surprisingly, recalling results from the Normal-Quantile plots in Chapter 2, less observations are highlighted as outliers after transformation, but differences are not overly pronounced.



**Figure 3.4:** log10 transformed data, Tukey heuristic (top) and z heuristic (bottom), window width: 7 days.

In summary, we did not find strong differences between threshold rules. If raw (potentially asymmetric) data are used we recommend to apply the Tukey heuristic because it does not as-

sume the data to follow a specific distribution. The method can be adapted further in order to even better handle asymmetric data. For a set of distributions Banerjee & Iglewicz [1] derived optimal threshold factors in dependence of sample size. If approximately normally distributed data (possibly after transformation) were used, z heuristics can be applied just as well.

Regarding the choice of parameters (threshold factor and window width) we can only give ad hoc suggestions from comparing the results of the parameter combinations we investigated. For more profound results we recommend to follow findings in Banerjee & Iglewicz [1] where applicable. Fine-tuned parameters should be derived from simulated benchmark data. We comment on this point in the discussion in [Chapter 5](#).

Our suggestions for parameters to use with the Tukey or z statistics based on this analysis are as follows: With the Tukey heuristic a threshold factor of 1.5 was used to get outliers, a factor of 3 for extreme outliers. We recommend to use both thresholds or investigate even higher factors. Clearly, outliers as indicated by the higher factor threshold have to be taken more seriously. Equivalent remarks apply to the z heuristic, where we recommend factors of at least 3.5 for outliers and 6 for extreme outliers. These choices lead to similar but slightly less conservative results, i.e. less outliers, than with the Tukey heuristic.

We recommend choosing a rather narrow window width of 3 to 7 ( $q = 1$  to  $q = 3$ ) measurements. This needs little more computing time but reveals only sharp abrupt changes. Using the extreme case of  $q = 1$  it seems that both outliers and structural changes can be identified. If only outliers need to be detected, the window width has to be increased slightly, e.g. to  $q = 3$  or  $q = 5$ . Wider windows in combination with higher thresholds might also yield good and maybe even better results. We cannot judge this issue without any ground truth data to compare against. In general, false positives are more desirable than false negatives. Under-detection can be prevented by either increasing window width further or shrinking threshold factors. From our results and for the parameters chosen we do not consider under-detection an issue here, but suspect rather the opposite.

We notice that in parts of time series that exhibit higher measurements and higher variability (due to e.g. seasonality) substantially more observations were marked as outliers. We suspect that this effect could be prevented by using local running measures of dispersion rather than global ones as was done in this analysis.

The two-sided median method marks outliers with respect to measurements of single stations. For stations with constantly low measurements and little dispersion this can lead to many counterintuitive outliers. We observed this for example for  $NO_2$  data in Cyprus (check [Figure A.5](#)), where one station measures in an area with very low concentrations (but many observations were flagged) while the other one measures higher concentrations with also higher variability leading to very few observations flagged as outliers. It might therefore be sensible to specify appropriate thresholds for rural and traffic stations, too, or to combine relative outlier labeling (moving window method) with absolute threshold values.

All of the methods used here are exploratory, i.e. no formal tests are conducted. The results can only indicate suspect cases, i.e. potential outliers, without any probability or confidence statement, and without any assessment of causes.

## 4 Exploratory Statistics for Break Detection

A break or structural change in a time series is considered to have taken place if a certain characteristic of the time series is altered thereafter. Changes can be persistent over time or transient (i.e. two structural changes of the same amount or type but opposite directions). Typical causes for air quality data are changes in measurement instrumentation or in environmental conditions. Such changes can apply to the mean or minimum values, to the percentage of missing values or zero measurements and the like. We considered the latter ones in our descriptive analyses in [Chapter 2](#). In the following we focus on abrupt changes in the running median.

Regarding structural change detection, results in Gerharz et al. [\[7\]](#) are clear, but not as convincing as for the outlier case. The recommended method is the 'moving average filter', which we follow up by testing its performance and suitability on AirBase [\[6\]](#) data for both an extended period of 2000 to 2010 and a series of pollutants. Furthermore the aim is to proof its robustness and to refine and determine the optimum set of parameter values when applied on the various pollutants.

### 4.1 Method and Parameterisation

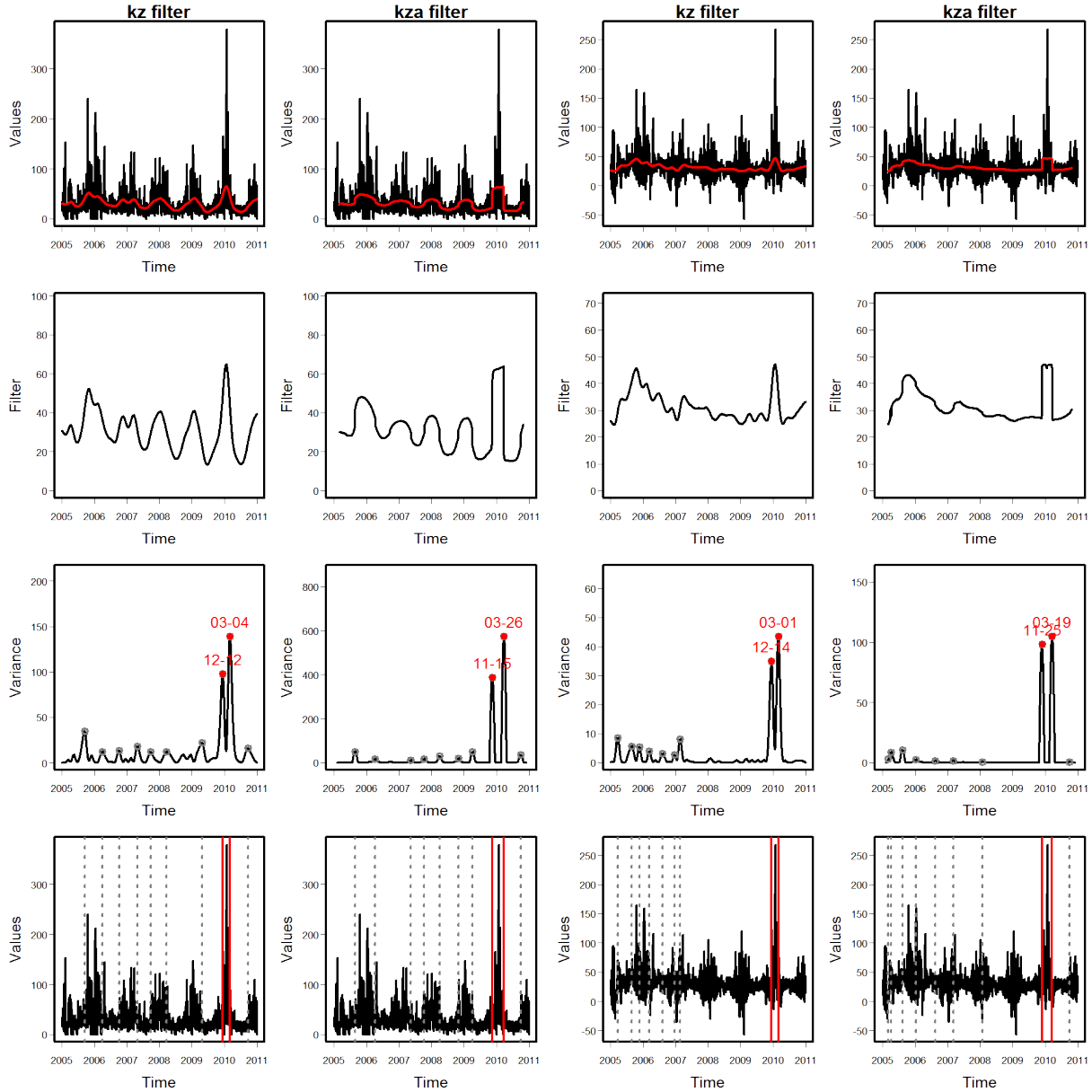
The method we consider is the Kolmogorov-Zurbenko adaptive filter, introduced by Zurbenko et al. [\[15\]](#), enhancing work by Rao & Zurbenko [\[10\]](#) in order to more easily detect abrupt changes and more accurately estimate the time of these events.

The idea behind the Kolmogorov-Zurbenko filter (Rao & Zurbenko [\[10\]](#)) is to smooth out short term variations from a time series by iteratively applying moving average filters. The smoothed time series reflects seasonal patterns, trends and potentially breaks, i.e. structural changes. Break points are characterized by increased variability in the surrounding window. Thus, for identifying break locations local maxima in variance of the smoothed time series can be investigated. However, abrupt discontinuities tend to be smoothed too, making it hard to identify them and to precisely determine their time of introduction. The procedure was thus refined by Zurbenko et al. [\[15\]](#) allowing for adaptive window width in dependence on the rate of change leading to a sharpening at break locations.

The method depends on two parameters: the half window width  $q$  and the number of filtering iterations  $k$ . As, especially for smaller window sizes, the method tends to over detection, a third parameter (threshold  $b$ ) is introduced. This threshold is typically a quantile of the variance time series corresponding to the adaptively smoothed time series. Only local maxima in variance above this threshold are considered to be indicators of suspected inhomogeneities.

The subsequent steps of change detection by the Kolmogorov-Zurbenko (adaptive) filter are illustrated in [Figure 4.1](#). The first two columns show the filters applied to an original time series, the columns to the right show the filters applied to the same time series after removing

the seasonal component. In each case, figures on the left show computation of the Kolmogorov-Zurbenko (kz) break points, while figures in the second column show respective plots for the adaptive version (kza). From top to bottom, first the raw time series is shown with the filter added in red, followed by a pure display of the filter, the variance of the filter where local maxima have been highlighted, and finally the raw time series with identified breaks.



**Figure 4.1:** Change detection with the Kolmogorov-Zurbenko (adaptive) filter. From top to bottom: time series with filter, pure filter, filter's variance and time series with detected break points.

The first two columns show the filters applied to the original time series from station CZ0MBEL, the columns to the right show the filters applied to the same time series after removing the seasonal component.

The example time series of daily  $PM_{10}$  measurements was taken from station CZ0MBEL. Parameters were a quantile threshold of 0.95 (and 0.75), window width of two months and three iterations for each of the filters.

In our implementation we made use of functions `kz` and `kza` from the R package `kza` [4], as well as function `runsd` from package `caTools` [13].

Parameter fine-tuning is complicated by the fact that ground truth data cannot be provided. For hourly data Gerharz et al. [7] recommend a single iteration for the Kolmogorov-Zurbenko filter (the adaptive version had not been considered for computational reasons), a half window width of 18 or 24 hours and restricting local maxima in variance to pass the 0.975 quantile. However, performance was not overly convincing, and parameter recommendation and testing was based on only a single AirBase time series, respectively, where furthermore breaks (and non-breaks) had been identified based on subjective judgement.

In the following paragraphs we first list the parameter values tested with selected time series from the Netherlands and Czech Republic where we had some information about events that potentially affected the time series. Finally we present exemplary results and draw conclusions.

## 4.2 Approach

Our approach is similar to the preceding section: We combine selected thresholds  $b$  with a number of half window widths  $q$  and the additional parameters of filtering iterations  $k$ . Again we are lacking ground truth data but were provided with some potential break dates for a set of stations where changes of the emission structure, of location in microscale or of the measuring device had taken place as outlined in Chapter 2. Thus, at these points in time and for the given station a break can be suspected and is hypothesized (but does not need to have taken place). We do not have any information whether any further breaks have happend at any other time points.

In our analysis we considered the following parameter values:

**Choice of threshold  $b$**  Zurbenko et al. [15] specify the expected number of detections  $E(D)$  in a clean time series without any breaks in dependence on the number of observations (length of time series)  $n$  and the method's parameters half window width  $q$  and number of iterations  $k$ :

$$E(D) = \frac{n}{2q\sqrt{k}}$$

For a given tolerance of false positives and a given length of the time series, method's parameters could be chosen accordingly. Decreasing the tolerance, however, leads to more computationally expensive iterations or increased window widths such that breaks within the windows become hard to distinguish and accuracy of break location deteriorates. Instead of choosing suboptimal parameters based on a tolerance for false detections, only the most probable, i.e. the largest, most evident breaks could be investigated by introducing threshold parameter  $b$ . We chose the local maxima in variance to pass the 0.75, 0.95, 0.975 and 0.995 quantiles.

**Choice of half window size  $q$**  The choice of the window width parameter should depend on the minimum size of breaks of interest, expressed as a multiple of the time series' standard deviation, (the smaller the breaks the wider the window has to be chosen) or of the desired accuracy of locating the break (the more accurate the narrower the window has to be chosen). As we are both interested in obvious breaks and higher accuracy, the half window width was chosen to equal six month at most for daily and two months for hourly data. A window width of two months in the case of hourly data is 24 times longer than in the case of daily data (as is the

time series in all). Short-time variability is smoothed out, while accuracy of locating a break is worsened by this factor, but is still within the same adequate range of days. Furthermore we can expect the same amount of false positive detections. The minimum window width for daily data was set to equal two weeks. For hourly data we decided to consider also windows of seven days, but no narrower as we were less concerned about short-time discontinuities than about breaks that are rather persistent and of relevance for some longer time period or the entire time series.

In detail, for daily data the  $q$ -values considered are 7, 15, 30, 60 and 90, leading to window widths ranging from two weeks up to a maximum window of half a year. For hourly data the  $q$ -values considered are  $3 \cdot 24$ ,  $7 \cdot 24$ ,  $15 \cdot 24$ ,  $30 \cdot 24$ , leading to window widths ranging from one week up to a maximum window of two months.

**Choice of iteration  $k$**  The smoothing effect of additional runs of the filters was investigated by choosing and combining iteration parameters  $k$ . We considered 3 and 10 iterations for the Kolmogorov-Zurbenko filter and 1, 3 and 10 iterations for the adaptive Kolmogorov-Zurbenko filter.

### 4.3 Results

We show results for daily  $PM_{10}$  measurements from selected stations in Czech Republic. [Appendix B](#) contains more figures on break detection for selected stations from Czech Republic. Complete results can easily be reproduced and investigated using our scripts that are available online and can be found at [http://ifgi.uni-muenster.de/~epebe\\_01/ETC-ACM/subtask\\_1.0.1.2-5b](http://ifgi.uni-muenster.de/~epebe_01/ETC-ACM/subtask_1.0.1.2-5b).

The example data were presented in [Chapter 2](#). Black triangles are used to mark a priori suspected breaks. In the time series plot in [Figure 2.4](#) a change is clearly visible only at the suspected break time point in station CZ0UTEM/CZ0UTPM, and another one might be hypothesized at station CZ0TOPO. All other suspected breaks are rather questionable. When investigating deseasonalized time series only the suspected break for station CZ0UTEM/CZ0UTPM would probably get marked as such by merely looking at the time series (i.e. subjective judgement). In fact, applying Kolmogorov-Zurbenko adaptive filters do not reveal any breaks at the hypothesized time points but do so for the one in station CZ0UTEM/CZ0UTPM. This one is a special case in the sense that it is the start of a series of null measurements of the kind that were investigated in [Chapter 2](#). We see these runs of equal (null) values as a special type of true transient breaks. Thus, we hope to find and precisely locate these break points with the Kolmogorov-Zurbenko adaptive filter.

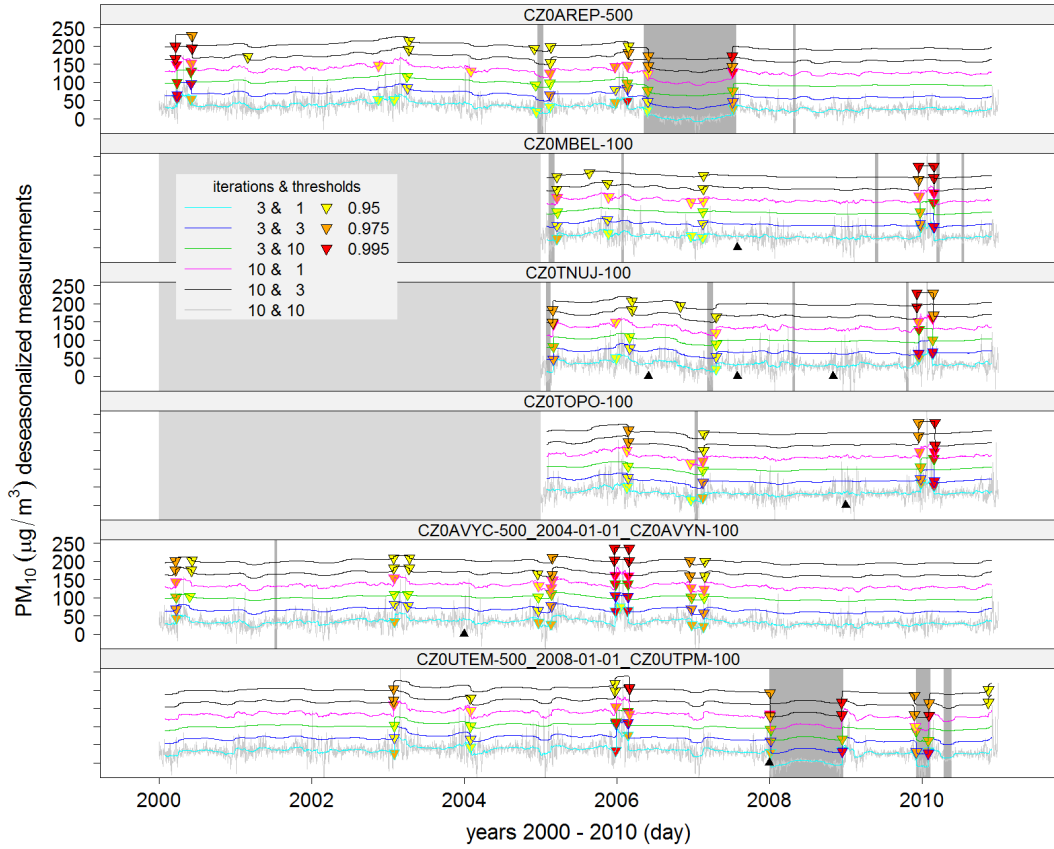
First, we comment on the rationale and the effect of deseasonalized data rather than original time series for use with Kolmogorov-Zurbenko (adaptive) filters. Zurbenko et al. [15] considered time series spanning multiple decades. To prevent from overdetection large window sizes had to be chosen, such that as a side effect annual seasonality was smoothed out in their examples. On the other hand the large window widths resulted in poor accuracy of locating the breaks. The AirBase data we are exploring here span eleven years at most and more accurate results are desired. Recall that the window width had been restricted to a maximum length of half a year for exactly this reason. [Figure 4.1](#) not only serves to demonstrate the steps in break detection but also shows differences in applying the method to original and deseasonalized time series.

We see that seasonality is preserved (together with trend and potential breaks) in the filter from the original time series. We observe a much higher amplitude in the winter 2009/10 in



comparison with other years which in both cases gets marked as a transient break. Considering less obvious breaks (above 0.75 quantile threshold, depicted by grey dotted lines) with original data some of the winter periods of other years get marked, too, which is not an issue for deseasonalized data. Instead some other minor discontinuities become visible here.

Even though the 0.75 threshold in combination with the other parameters investigated seems not to be high enough to prevent from overdetection (and for this reason and the sake of lucidity was excluded from any further plots) we recommend the use of deseasonalized data. We show another comparison in the end of this section when examining the impact of window widths.



**Figure 4.2:** Findings of the Kolmogorov-Zurbenko adaptive filter for varying numbers of iterations, applied to deseasonalized data and with a window width of one month.

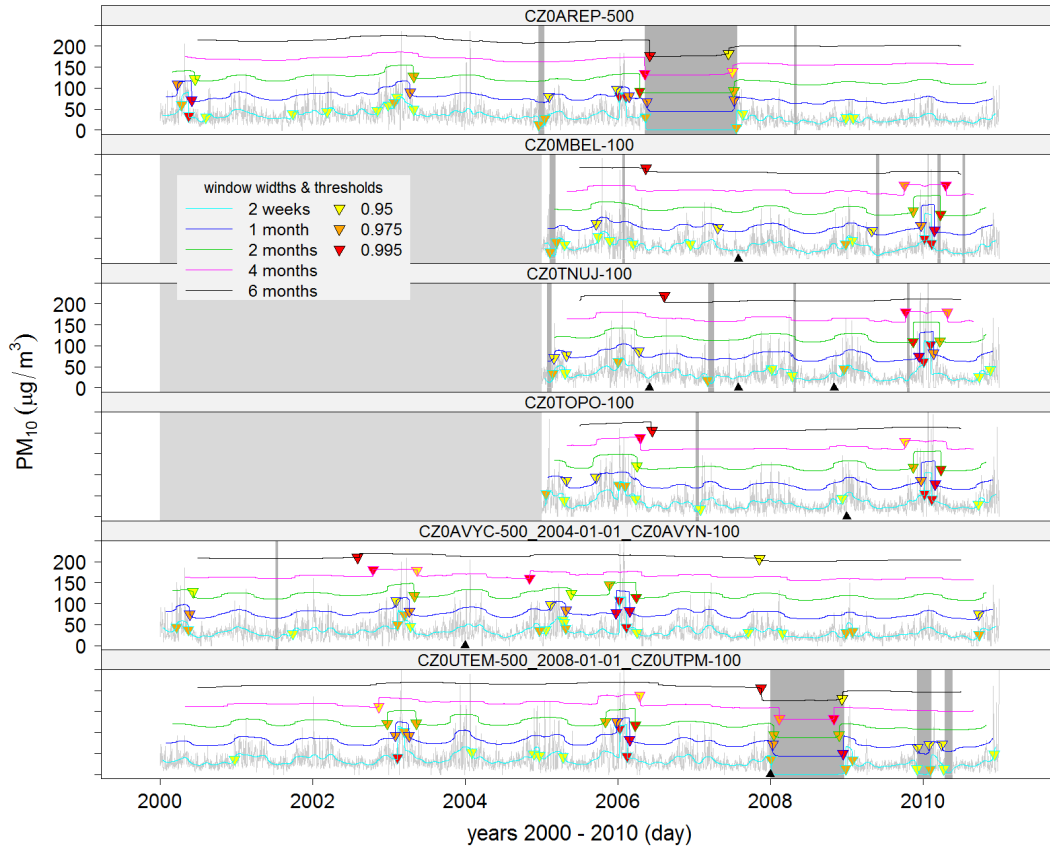
Next, we present results of comparing the Kolmogorov-Zurbenko adaptive filter with differing numbers of iterations both for the initial (kz) and the adaptive runs (kza). For a given window width of one month ( $q = 15$ ) in Figure 4.2 the deseasonalized time series of Czech example stations are overlaid with the filter using three kz and one kza runs. All other combinations considered are depicted with additional offsets for the seek of graph readability. Identified break points are shown by triangles and are color coded with respect to thresholds  $b$ . Clearly, the higher the threshold the more evident is the suggested break.

More iterations are expected to result in lower precision of locating the breaks. This effect seems to be too small to be observed in Figure 4.2. The smoothing effect of additional runs can be seen both for the kz and especially for the kza runs when comparing the filters. It should

lead to less false positives but might also dilute the power of detecting true positives to some extent. Without ground truth data we cannot judge on this. However, we observe that differing numbers of iterations seem not to result in big differences in the detected breaks. At least for the most evident breaks, results coincide quite well, indicating robustness of the method with regard to the number of iterations.

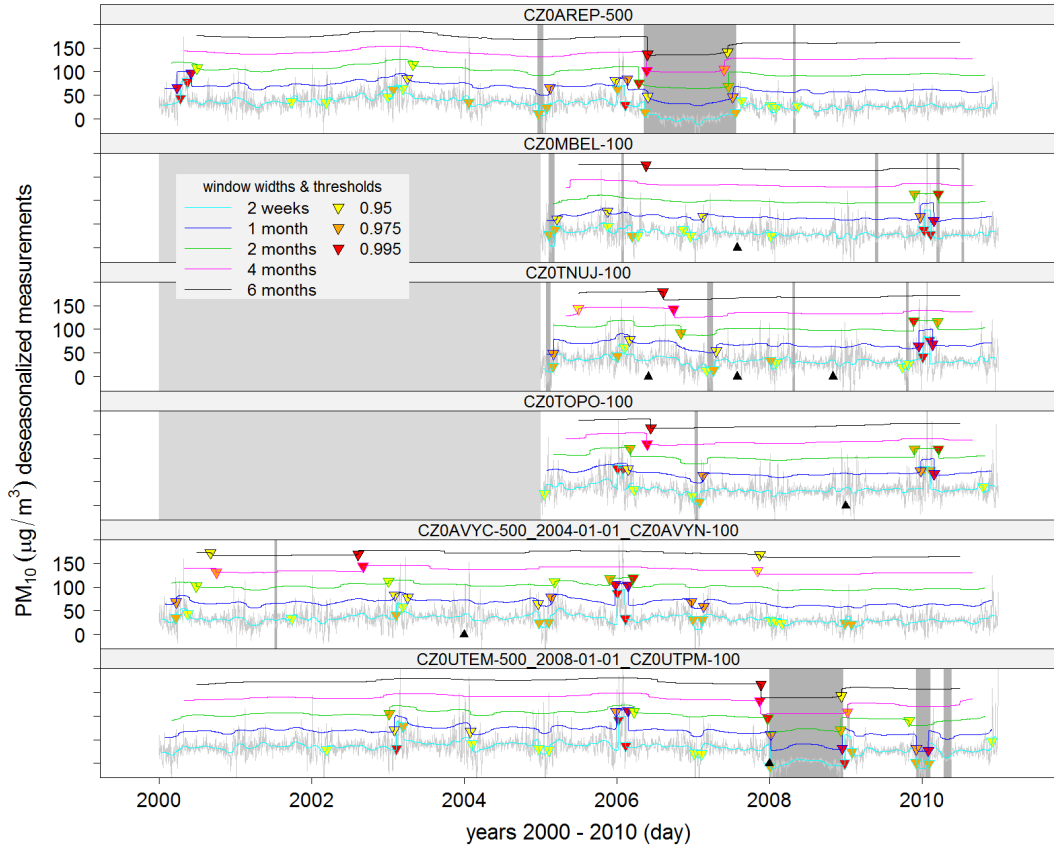
We recommend to use three iterations of filtering runs such that some smoothing is applied to the data and breaks are sharpened while false positives still can be accounted for either by window width or even better through the threshold parameter. As an advantageous side effect less iterations finally minimize computing time.

Last, we show results of investigating the effect of varying window widths. For given iteration parameters ( $k_{kz} = 3$  and  $k_{kza} = 3$ ) Figure 4.3 and Figure 4.4 show the original and the deseasonalized time series of Czech example stations, respectively. Again they are overlaid with the filters, using window widths of two weeks up to six months, in this case. Identified break points are shown by triangles and color coded with respect to the thresholds.



**Figure 4.3:** Findings of the Kolmogorov-Zurbenko adaptive filter for varying window widths.

Narrower window widths are expected to result in higher precision of locating the breaks on the one hand and in more false positive detections on the other hand. Both these effects are clearly visible from Figure 4.3 and Figure 4.4: considerably more detections occur for narrower windows as well as better localisation of transient breaks. We thus recommend to use narrow window widths in combination with high threshold values.



**Figure 4.4:** Findings of the Kolmogorov-Zurbenko adaptive filter for varying window widths when applied to deseasonalized data.

Window width comparisons for hourly data ( $PM_{10}$ ,  $SO_2$  and  $O_3$  measurements from Czech stations) can be found in [Appendix B](#). One week filters look rather rough and several (short transient) breaks are identified at least with low thresholds. At times it seems that for very narrow windows the break detection rather becomes an outlier detection method. For  $PM_{10}$  and  $SO_2$  data we regularly observed breaks across stations at the beginning of a year, especially for 2003 and 2006. For both pollutants, window widths of two weeks or one month seem to be adequate, just as for daily  $PM_{10}$  data. Ozone measurements exhibit very pronounced seasonality. We therefore highly recommend the use of deseasonalized data for this pollutant.

Summarizing our suggestions regarding parameter choice, we recommend to use three iterations each for the initial Kolmogorov-Zurbenko filter and the actual adaptive filter. We do not recommend to use window width longer than two months, but rather to choose narrow windows of widths up to e.g. one month and use rather high threshold values like the 0.975 and 0.995 quantiles when worried about false positives. We recommend the use of two or three threshold values to get an idea about how the relevance of detected breaks could be ranked.

When applying the Kolmogorov-Zurbenko (adaptive) filter with window widths that do not span the largest periodicity within the data, deseasonalized measurements should be used in order to reduce false positive break detections. Even with deseasonalized data some breaks, especially transient changes within periods of typically higher or lower measurements, might rather be due to physical reasons (e.g. particularly cold winters or particularly hot and dry summer seasons) than due to errors or changes in the measuring device. Transient changes in general can also be

seen as an accumulation of (potentially reasonable) outliers. In order to keep such false detections away, instead of deseasonalizing the data, prior to analysis, we recommend to adjust data with respect to the variables that seasonality is a substitute for (e.g. weather conditions). Consequently, natural expected breaks are accounted for and would not get marked as breaks any more. Another possibility would be a comparison between neighboring stations or all stations in a region with similar (weather) conditions. Figures 4.2, 4.3 and 4.4 show quite a number of such coincidences across stations, questioning their classification as undesirable breaks.

## 5 Discussion

The task in this research was twofold: Moving window statistics were applied to European air quality time series from the DEM database [5] with the objective of detecting outlying observations. The study period was set to the year 2010, and methods were applied to all stations from Romania, Switzerland and Cyprus measuring  $PM_{10}$  concentrations on a daily basis and concentrations of  $O_3$ ,  $SO_2$ ,  $NO_2$  and  $CO$  on an hourly basis. Then again Kolmogorov-Zurbenko adaptive filters were applied to European air quality time series from AirBase [6] with the objective of detecting structural changes. The study period was set to comprise the years 2000 to 2010, and filters were applied to selected stations from the Netherlands and Czech Republic measuring  $PM_{10}$  concentrations both on an hourly and a daily basis and concentrations of  $O_3$  and  $SO_2$  on an hourly basis.

### 5.1 Outlier Detection

In order to identify outlying observations in a time series in [Chapter 2](#) per station boxplots were examined. A boxplot is a simple explorative tool that visually summarizes distributional information but ignores the temporal dependence structure within the data. The moving window method investigated in [Chapter 3](#) - when based on the Tukey heuristic - represents a direct enhancement of the boxplot considering local dependencies. Further heuristics considered were the z score as in Gerharz et al. [7] in conjunction with robust statistics, and the modified z score.

**Parameters** Choice of threshold parameters was based on the above mentioned heuristics. As already indicated by the boxplots, at some stations the amount of observations labeled as outliers is substantial. To prevent against over-detection increasing these thresholds might be a sensible choice, as well as choosing narrow window widths. These findings coincide with observations in Gerharz et al. [7]. By considering more than one threshold rule, the assessed inhomogeneities can be ranked accordingly. Assigning optimal parameter values was not possible due to the lack of ground truth data to test against.

**Heuristics** The effect of the heuristic used in the moving window method (at least for our mostly unimodal and near-symmetric data) is negligible. We prefer using the Tukey heuristic as it does not impose any distributional restrictions on the data.

**Local vs Global Measure of Dispersion** In our analysis we made use of global measures of dispersion (global IQR and global MAD respectively) rather than using their local versions within the windows (i.e. running IQR and running MAD). This choice was recommended by Gerharz et al. [7] based on a limited set of time series. We noticed however that substantially more observations were marked as outliers in parts of time series that exhibit higher measurements and higher variability. An effect that we believe can be circumvented by applying local (running) measures of dispersion rather than global ones. Window widths then might be chosen to be somewhat wider than with global measures. Alternatively window widths for location and

dispersion measures could be allowed to differ, such that narrow windows for running quantiles can be combined with somewhat wider windows for running IQR or MAD.

**Station Type** Probably due to station type (traffic or background) some stations show quite high variability in combination with higher measurements, while other stations can be characterized by constantly low measurements. For the latter type higher thresholds need to be chosen to prevent against over-detection. Alternatively, we suggest to combine relative outlier detection with absolute thresholds in order to reject these kind of outliers.

**Reasonable Outliers** Not all 'true' outliers are false recordings. Outliers that coincide across (neighboring) stations point to other causes like extreme weather conditions or fireworks on public holidays. If such measurements should not be assigned as outliers some data adjustment based on covariate information must take place prior to running the moving window outlier detection.

**Log Transform** Air quality data often undergo some manipulation prior to analysis, typically a logarithmic transformation. This results in measurements that are exponentially shrunk towards zero, such that extremely high values might not appear that extreme any more. The log transform is thus said to have an outlier removing property. Another effect is that distributions that are positively skewed, i.e. skewed to the right, become more symmetric and thus more normally distributed. Clearly, for outlier detection in log transformed data, optimal parameters will differ from the ones for untransformed data. Typically, threshold factors will be lower and window widths wider than for original data.

**Recommendations** For outlier detection we suggest to use the Tukey heuristic with threshold factors of at least 3 (and 1.5) in the style of boxplot outlier labeling. Higher threshold factors are recommended to mitigate false positive labeling. Narrow window widths result in less detections and are recommended because of the fact that an outlier is a local property. We assume that mainly false-positives will thus be rejected in comparison to wider window widths. Given a global measure of dispersion we recommend using windows of three to eleven measurements. Optimal parameters might differ between pollutants and temporal resolutions but we suspect that station type is even more informative, as it strongly influences the variability and baseline measurements of the pollutant. It might thus be sensible to set different parameters for traffic and background stations or to combine outlier detection rules with absolute values that have to be passed. Recommendations for parameter values are based on subjective judgement and not on statistical measures as we are lacking ground truth data. We suggest to derive fine-tuned parameters from simulated benchmark data.

## 5.2 Break Detection

Structural changes or breaks within a time series can be of various types. One type of transient breaks was investigated in [Chapter 2](#). Runs of equal values of a certain minimum length indicate measurement errors and should better be recorded as not available. For our example AirBase time series of daily  $PM_{10}$  measurements from Czech Republic it turned out that all of these runs had either negative or zero values. In [Chapter 4](#), abrupt changes in the running median were analyzed by means of the Kolmogorov-Zurbenko adaptive filter.

**Parameters** We found that the method is rather robust with respect to the number of iterations of filter runs, both for the initial smoothing and the actual adaptive smoothing. Window width should be chosen with time series lengths, potential periodicity, the desired precision of

locating a break, and the risk of an increased false positive rate in mind. All else left equal, for longer time series window width has to be increased accordingly to keep the expected number of false positives and the precision of locating a break equal, too. Narrower windows result in both increased precision and increased number of detections. A threshold parameter was introduced to deal with over-detection. Increasing the threshold allows setting narrow windows and thus increased precision. By considering more than one threshold rule, the assessed inhomogeneities can be ranked accordingly. Assigning optimal parameter values was not possible due to the lack of ground truth data to test against. To prevent against labeling 'seasonal breaks' window width could be chosen wide enough to capture this periodicity.

**Deseasonalization** Alternatively, to gain precision, we advise to decompose time series prior to analysis and to use the Kolmogorov-Zurbenko adaptive filter with deseasonalized data (i.e. time series where the seasonal component has been removed). We applied the STL procedure [3] assuming yearly periodicity but without any fine-tuning. This step clearly deserves more attention should it be used on a regular basis.

**Reasonable Breaks** Some breaks might be caused by environmental conditions rather than changes or errors in the measuring device or reporting. If such data were available, data adjustment based on covariate information can take place prior to running the break detection filter in order not to flag these kind of breaks. Spatial information might be used as a surrogate. Indeed, identified breaks often coincided across various stations.

**Recommendations** For break detection our recommendation is to iterate each of the filters three times and to use narrow windows in order to increase precision of locating the break in combination with high threshold values in order to prevent over-detection. In general, we recommend to apply deseasonalisation or adjustment to weather conditions prior to analysis. We conclude that a window width of two weeks to one month in combination with high threshold values (0.975 and 0.995) are suited best. This holds for all pollutants and both hourly and daily data, assuming a time series length of eleven years. Recall however that hourly intervals consist of 24 times as many measurements than daily data when setting parameter  $q$ . For considerably longer or shorter time series some further parameter adjustments might be required. We state again that recommendations for parameter values are based on subjective judgement and not on statistical measures as we are lacking ground truth data, and that we suggest to derive fine-tuned parameters from simulated benchmark data.

## 5.3 Further Remarks

Without ground truth data, parameter fine-tuning on a sound statistical basis is not possible. All recommendations given here are either based on theoretical considerations or subjective judgement, mainly based on visual inspection of the data, and compared to the results of the inhomogeneity detections.

As ground truth data seems to be structurally absent for European air quality data, we suggest instead to derive fine-tuned parameters from simulated benchmark data. The challenge herein would be the design of such data. Time series should mimic the characteristics of real but clean time series, i.e. without any outliers or structural changes. The challenge would then be the assignment of inhomogeneities: What kind of inhomogeneities are of most interest? Which amount of change should still be detected? Do certain patterns (number and distribution within a time series) need to be investigated? With such data at hand, however, optimal parameters



could easily be determined. For this end the Jaccard coefficient was used by Gerharz et al. [7], although with a somewhat meaningless definition for the case of clean time series. As an alternative we recommend the use of the ROC curve (receiver operating characteristic) because it offers more insight, a meaningful interpretation, and can be presented graphically.

For both outlier and break detection methods, we observed that with extremely narrow windows it might be possible to detect also the other kind of inhomogeneity, respectively. Nevertheless, we suggest to maintain the designated methods for the different tasks.

In this paper we focused on the binary classification of outliers vs. non-outliers, structural breaks vs. absence of structural breaks. In the absence of ground truth data needed to calibrate these classifications, an alternative approach might be to label potential outliers or break points in a continuous way, and establish ranks on the likeliness of being an outlier or break point. Such a labeling could be used by network operators as a means to manually verify whether suspect cases need further labeling in the official data sets.

The methods considered are all univariate, focussing on single time series and neither making use of time series from other (neighboring) stations nor of time series of other pollutants measured at the same place. Further, no covariate information was included that could account for justified inhomogeneities. The methods could thus be improved, potentially at the cost of simplicity and applicability in an automated setting. Maybe of greater interest to station maintainers than to the EEA, instead of retrospective detection as in this paper, a real-time surveillance system could be set up, which would trigger an event whenever a newly introduced outlier or structural break was identified.



# References

- [1] Banerjee, S. & Iglewicz, B. (2007). A simple univariate outlier identification procedure designed for large samples. *Communications in Statistics: Simulation & Computation*, 36(2), 249 – 263.
- [2] Basu, S. & Meckesheimer, M. (2007). Automatic outlier detection for time series: an application to sensor data. *Knowledge and Information Systems*, 11, 137–154.
- [3] Cleveland, R., Cleveland, W., McRae, J., & Terpenning, I. (1990). STL: A seasonal-trend decomposition procedure based on loess. *Journal of Official Statistics*, 6(1), 3–73. Statistics Sweden.
- [4] Close, B. & Zurbenko, I. (2012). *kza: Kolmogorov-Zurbenko Adaptive Filters*. R package version 2.03.
- [5] EEA (2011). DEM 2010 raw monitoring data as delivered by the national representatives. Original data that did not undergo any screening by the ETC/ACM.
- [6] EEA (2012). AirBase: European Air Quality Database. Version 6. Data sources and processors: ETC/ACM.
- [7] Gerharz, L., Gräler, B., & Pebesma, E. (2011). *Measurement artefacts and inhomogeneity detection*. Technical report.
- [8] McLeod, A., Yu, H., & Mahdi, E. (2012). Time series analysis with R. In *Handbook of Statistics: Time Series Analysis: Methods and Applications*. Elsevier Science.
- [9] R Core Team (2012). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- [10] Rao, S. T. & Zurbenko, I. G. (1994). Detecting and tracking changes in ozone air quality. *Air & Waste*, 44(9), 1089–1092.
- [11] Rao, T., Rao, S., & Rao, C. (2012). *Handbook of Statistics: Time Series Analysis: Methods and Applications*. Handbook of Statistics. Elsevier Science.
- [12] Ripley, B. & from 1999 to Oct 2002 Michael Lapsley (2012). *RODBC: ODBC Database Access*. R package version 1.3-6.
- [13] Tuszynski, J. (2012). *caTools: Tools: moving window statistics, GIF, Base64, ROC AUC, etc.* R package version 1.13.

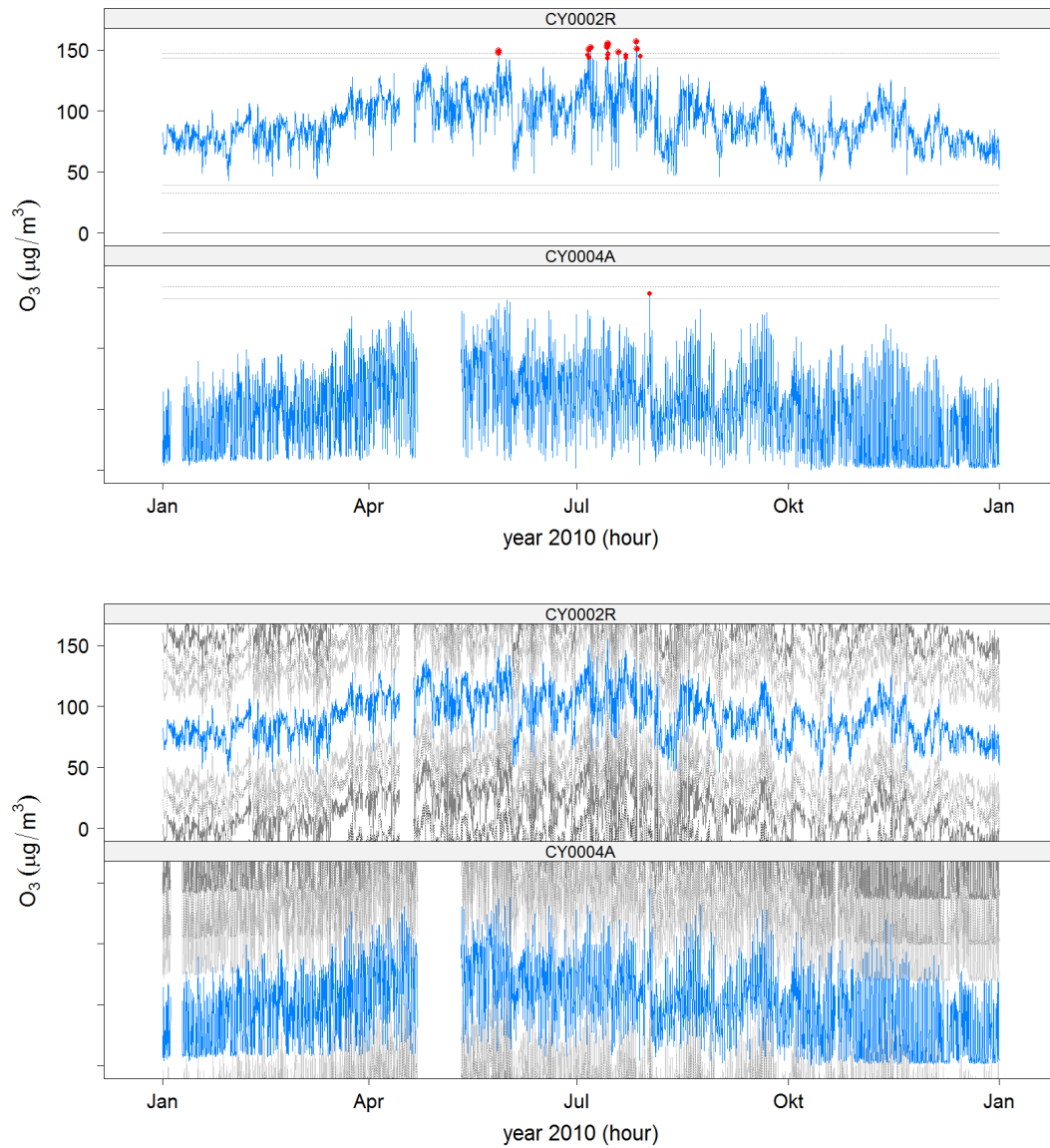
- [14] Xie, Y. (2012). *knitr: A general-purpose package for dynamic report generation in R*. R package version 0.7.5.
- [15] Zurbenko, I., Porter, P., Gui, R., Rao, S., Ku, J., & Eskridge, R. (1996). Detecting discontinuities in time series of upper-air data: development and demonstration of an adaptive filter technique. *Journal of Climate*, 9, 3548–3560.

# Appendices

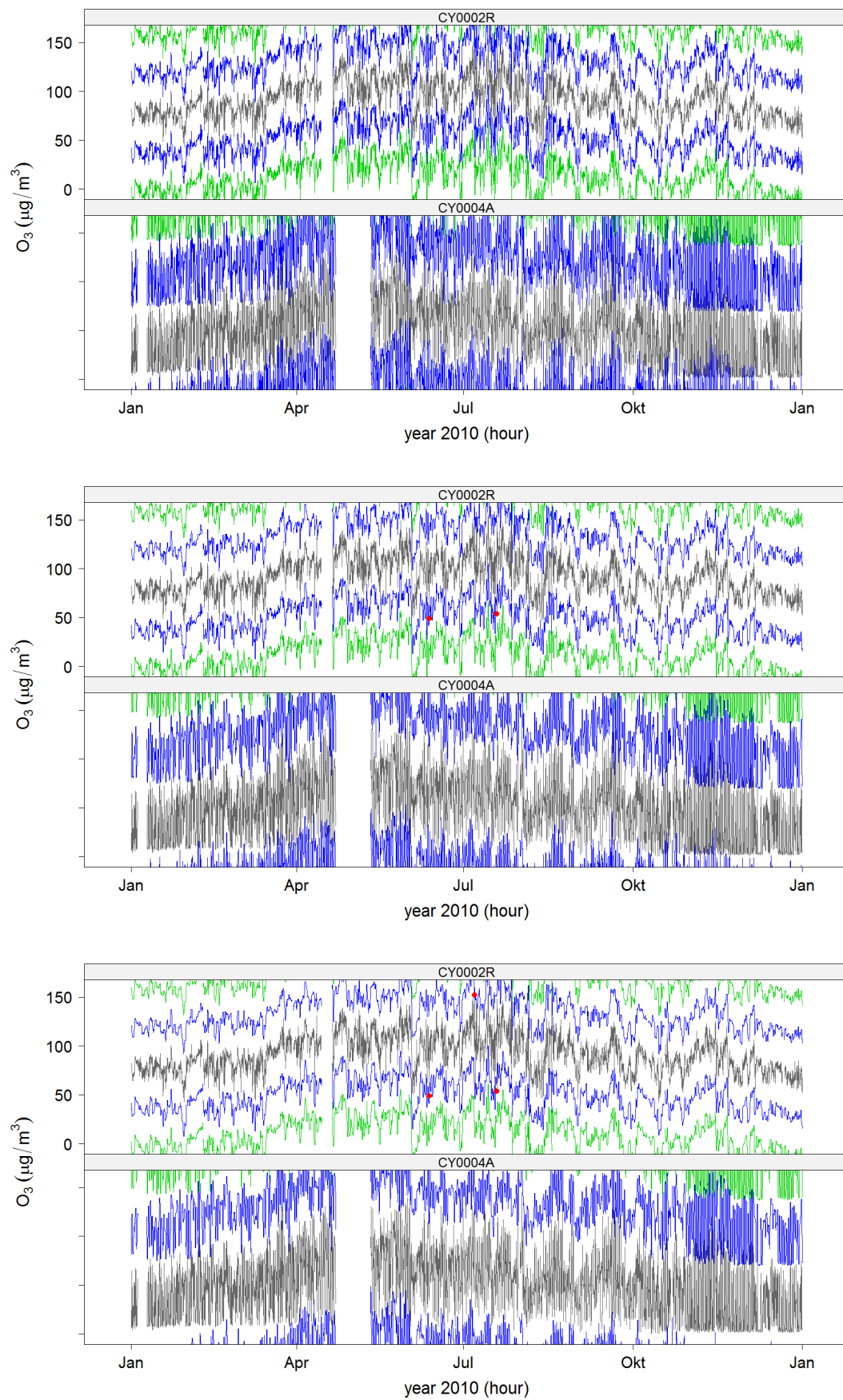
# A Outlier Detection

This appendix contains more figures on outlier detection for stations from Cyprus. We present figures with sensitivity assessment and window width comparisons for half window widths of 2, 4 and 6 hours.

## A.1 $O_3$ hourly data

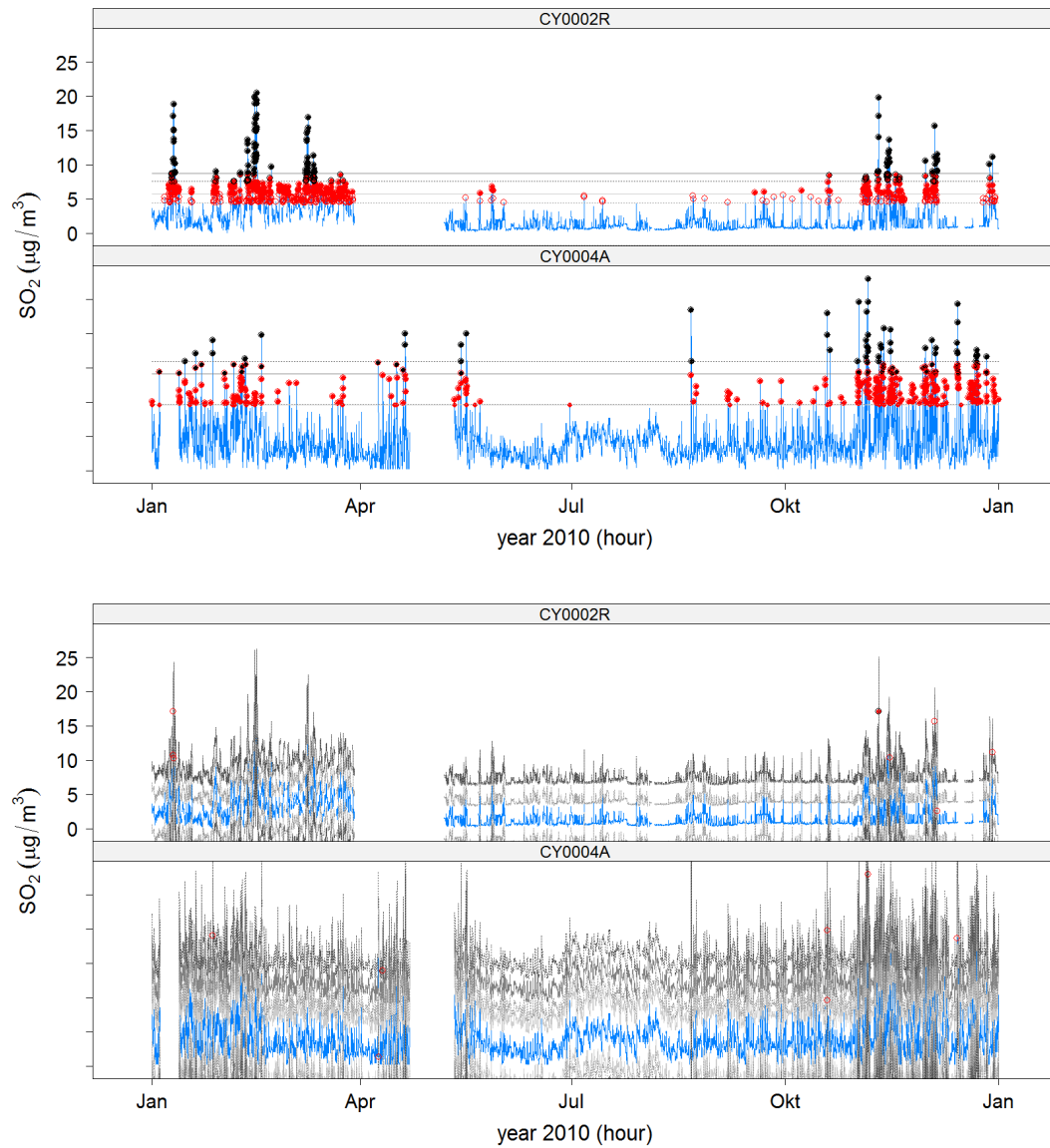


**Figure A.1:** Original time series data, overview heuristics, no window (top) and extremely narrow window of width 3 (bottom).

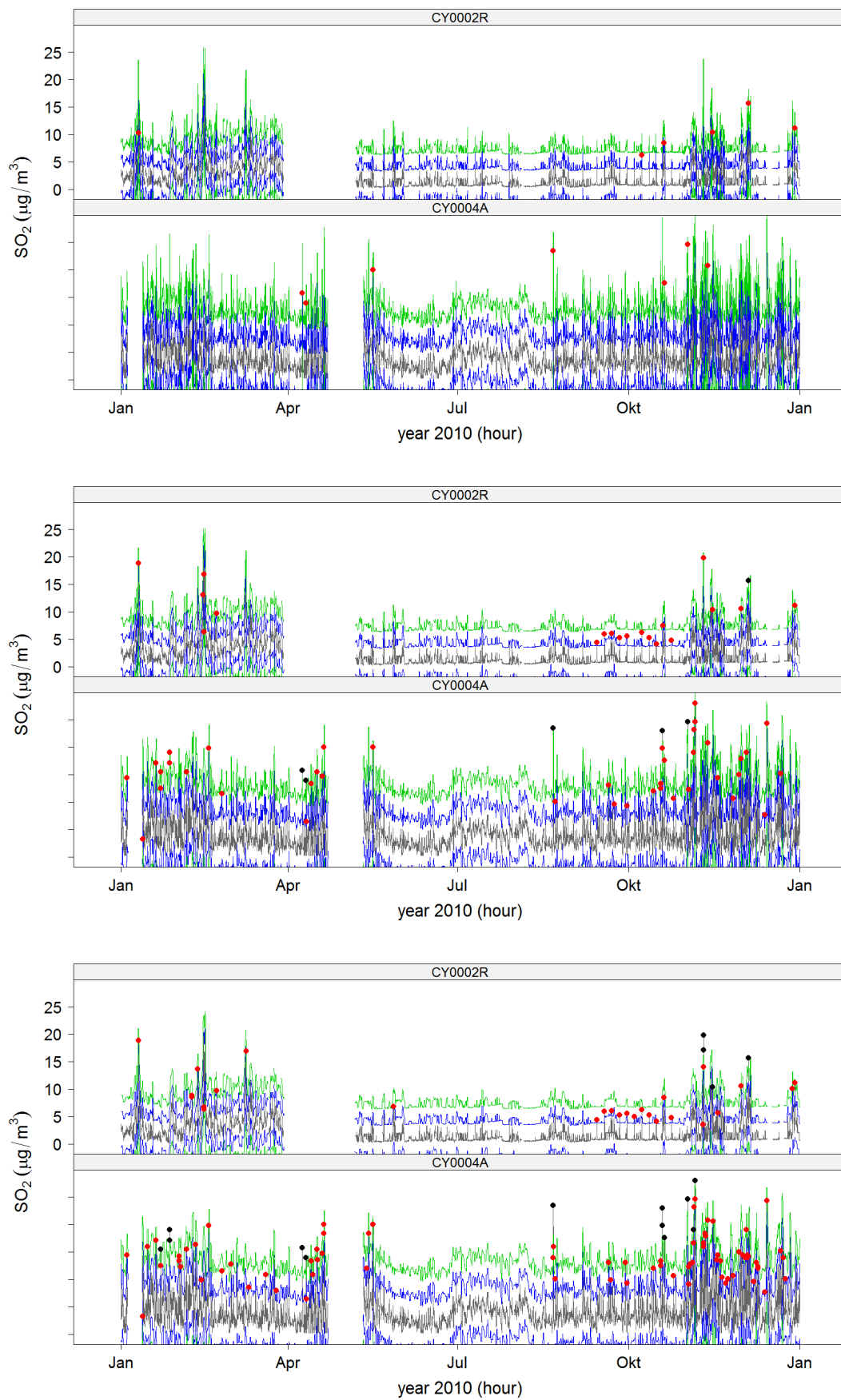


**Figure A.2:** Original data, Tukey heuristic. From top to bottom: Window width = 5 hours, 9 hours and 13 hours.

## A.2 $SO_2$ hourly data

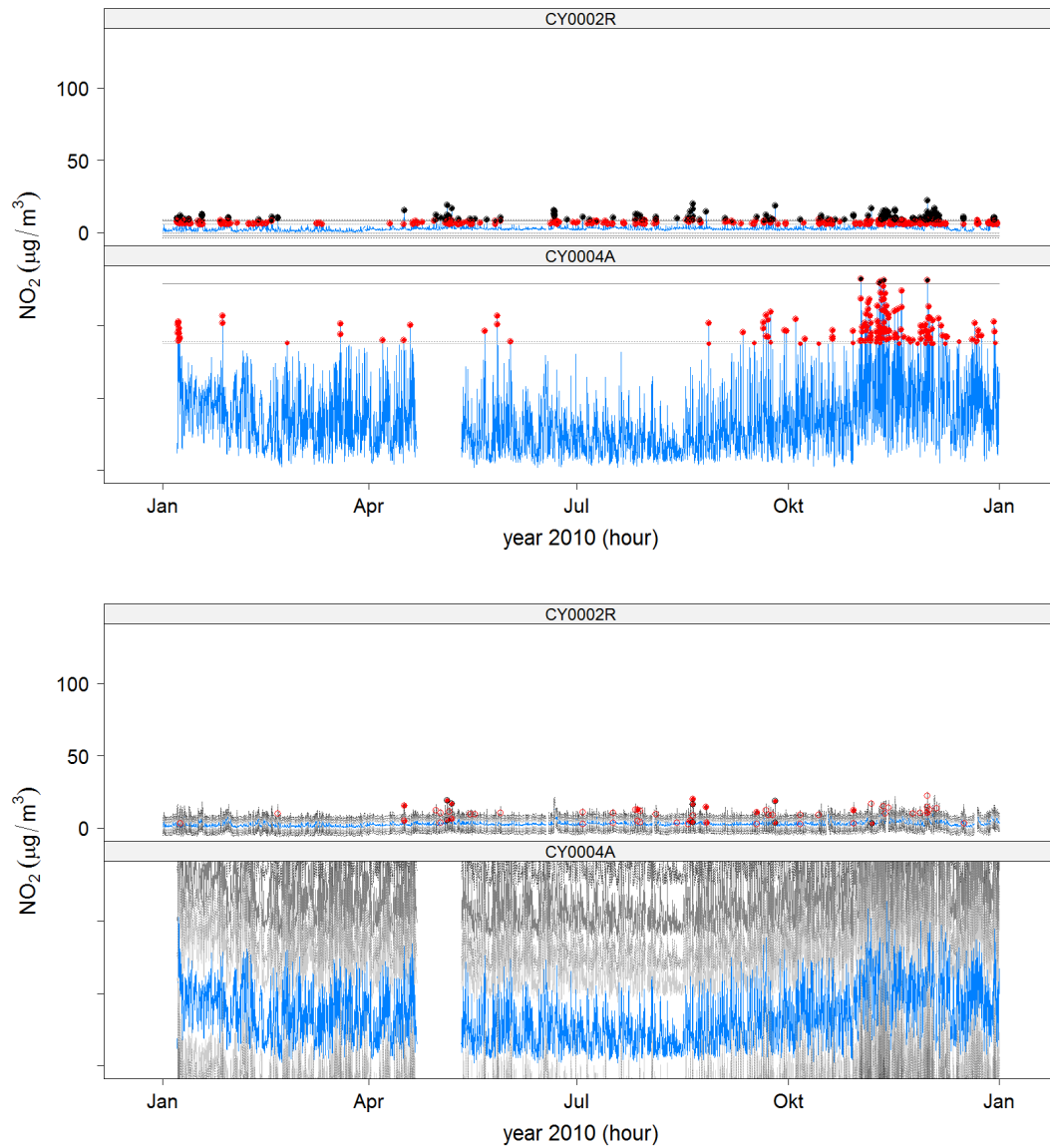


**Figure A.3:** Original time series data, overview heuristics, no window (top) and extremely narrow window of width 3 (bottom).



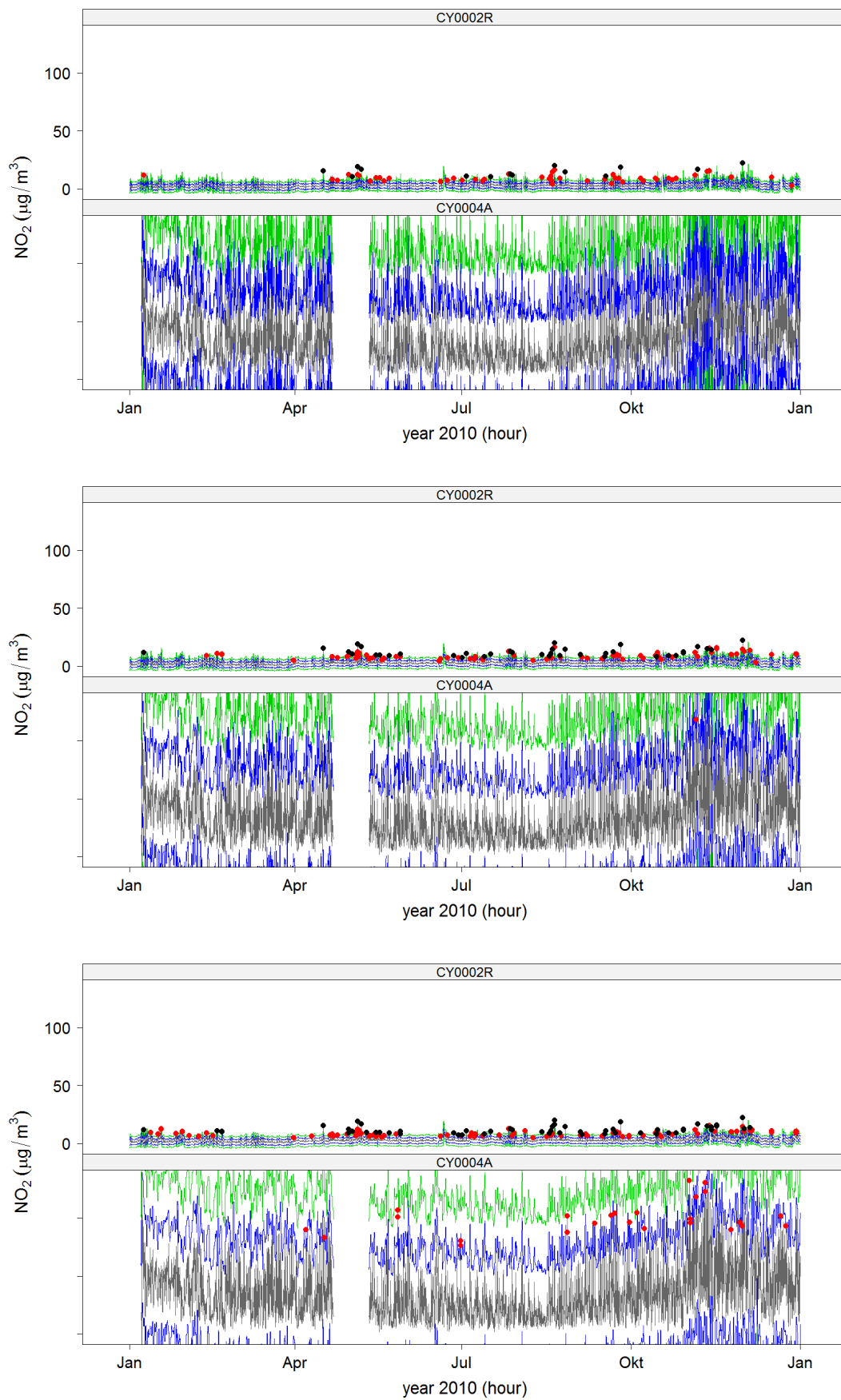
**Figure A.4:** Original data, Tukey heuristic. From top to bottom: Window width = 5 hours, 9 hours and 13 hours.

### A.3 $\text{NO}_2$ hourly data



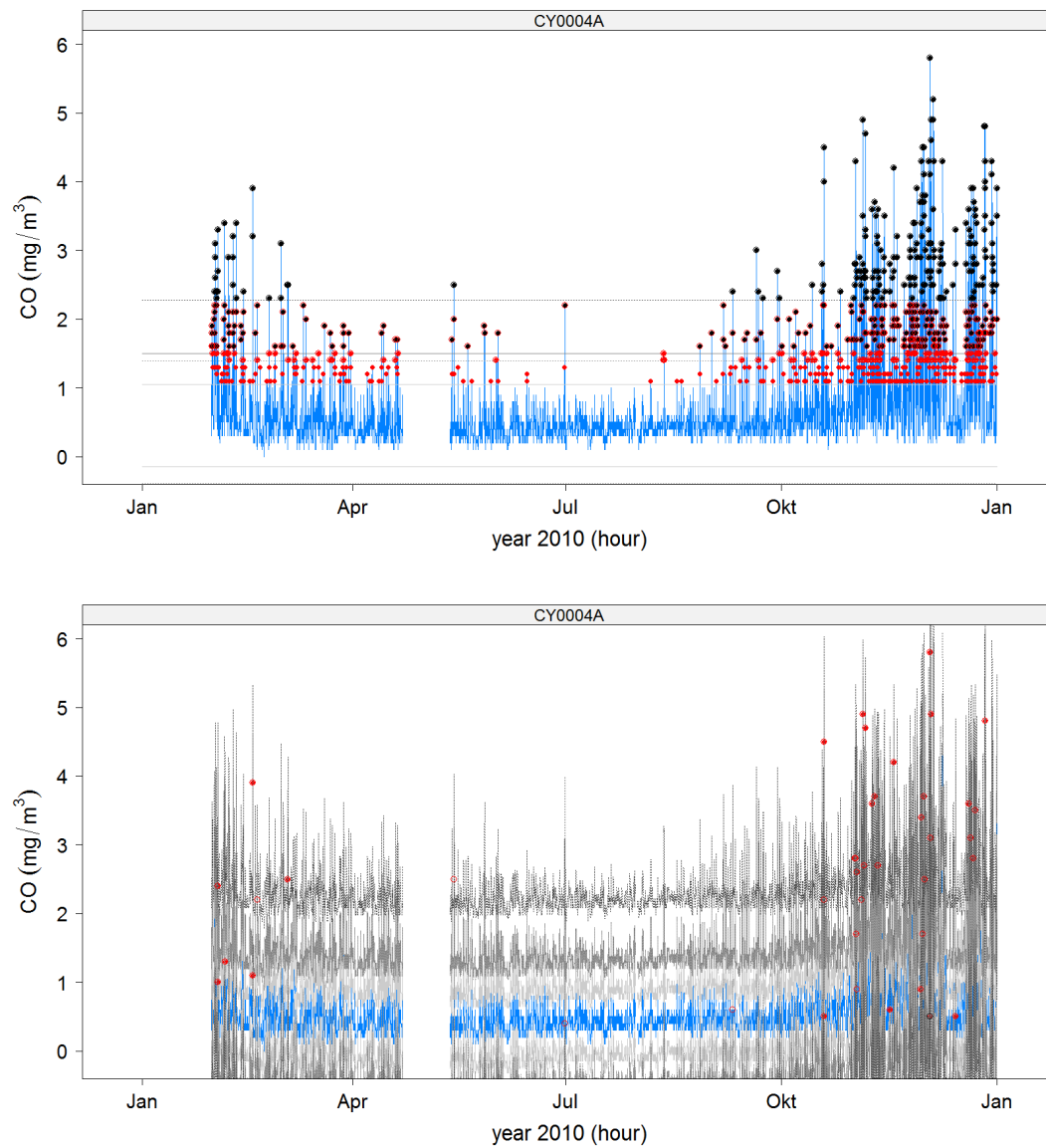
**Figure A.5:** Original time series data, overview heuristics, no window (top) and extremely narrow window of width 3 (bottom).



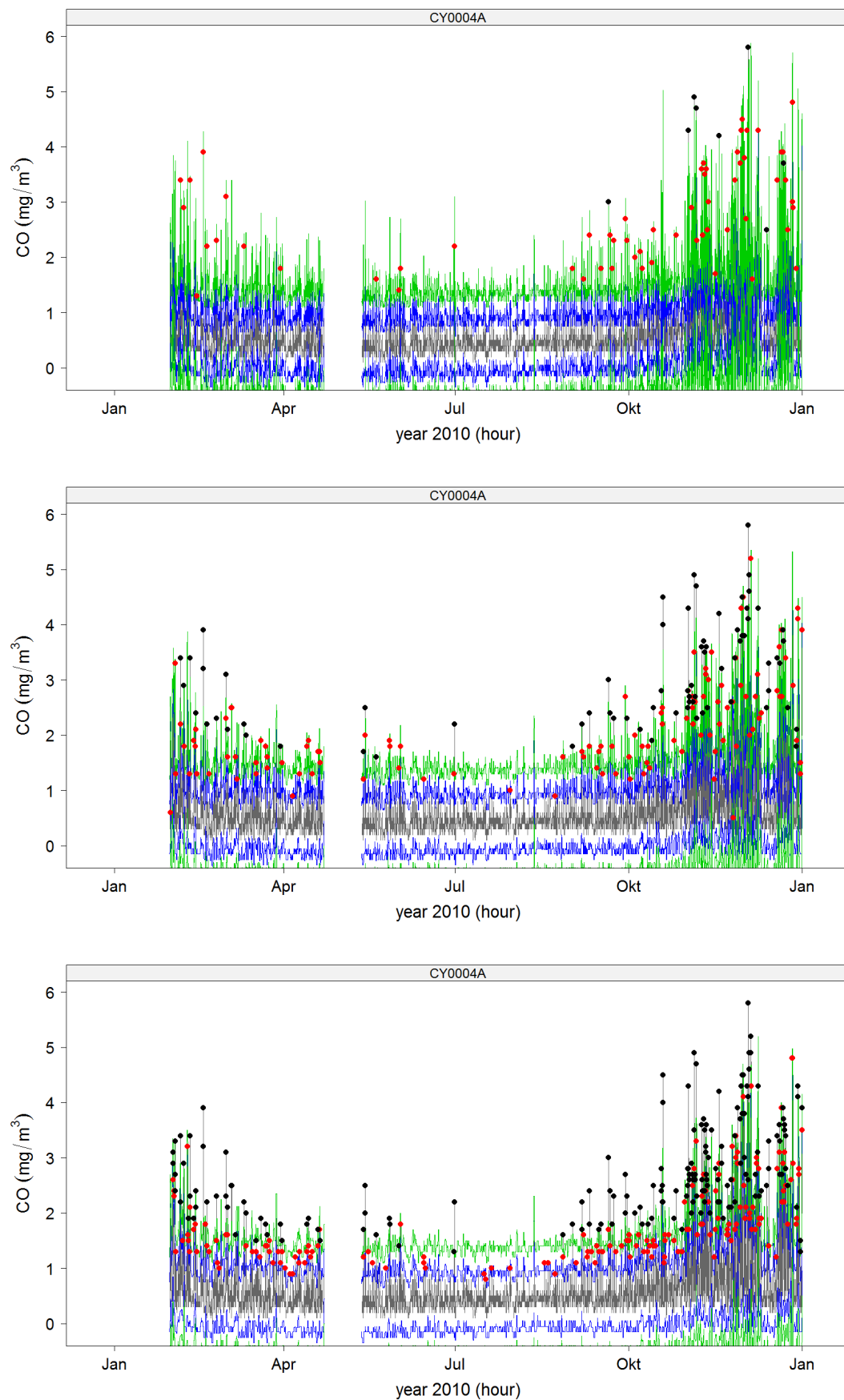


**Figure A.6:** Original data, Tukey heuristic. From top to bottom: Window width = 5 hours, 9 hours and 13 hours.

## A.4 CO hourly data



**Figure A.7:** Original time series data, overview heuristics, no window (top) and extremely narrow window of width 3 (bottom).

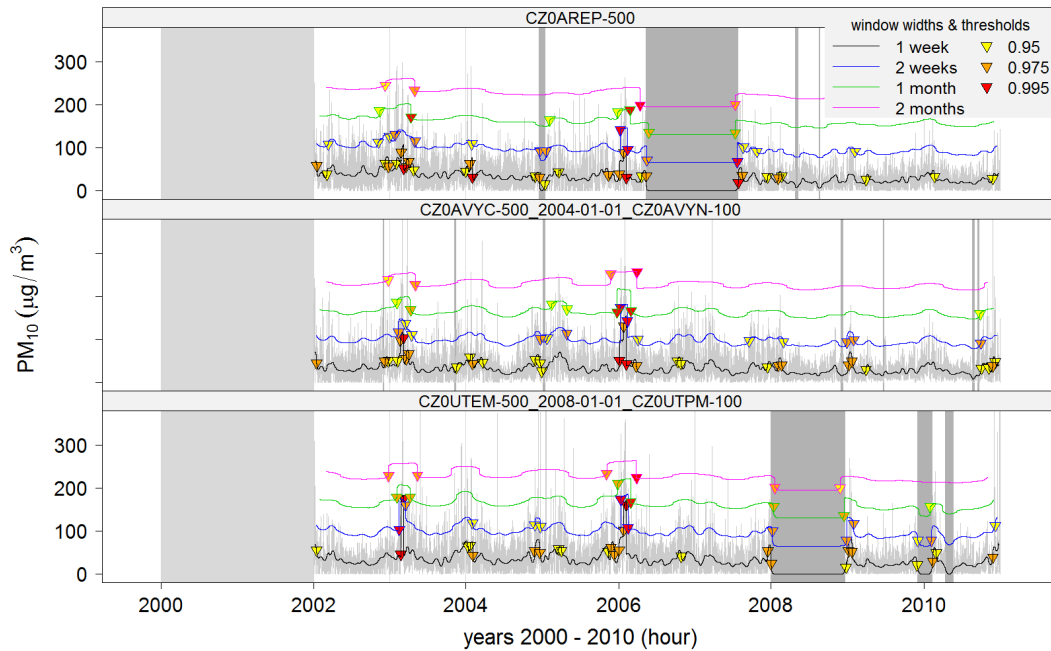


**Figure A.8:** Original data, Tukey heuristic. From top to bottom: Window width = 5 hours, 9 hours and 13 hours.

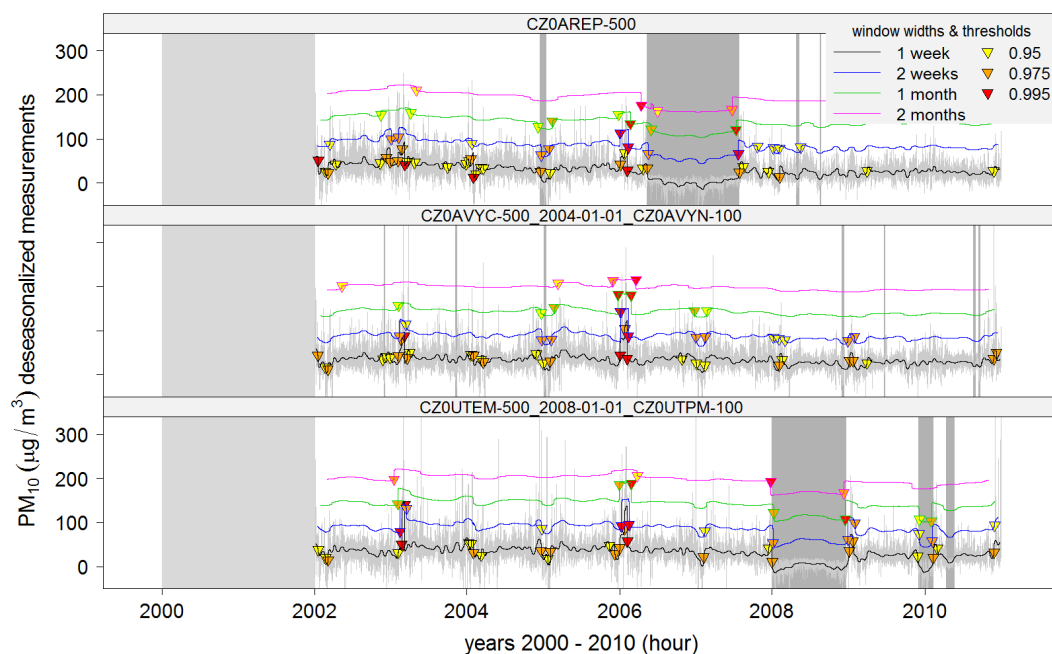
## B Break Detection

This appendix contains more figures on break detection for selected stations from Czech Republic. For hourly data only runs of equal values with a minimum length of 120 measurements, i.e. five days, have been marked in the figures. We present figures with window width comparison for original time series and deseasonalized data.

### B.1 $PM_{10}$ hourly data

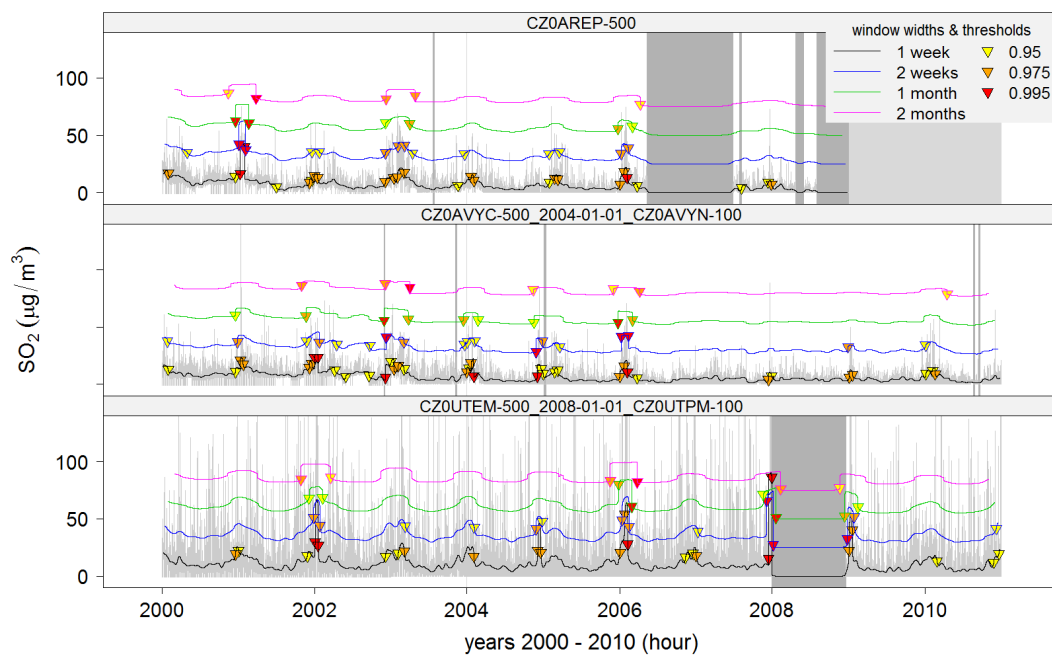


**Figure B.1:** Findings of the Kolmogorov-Zurbenko adaptive filter for varying window widths.

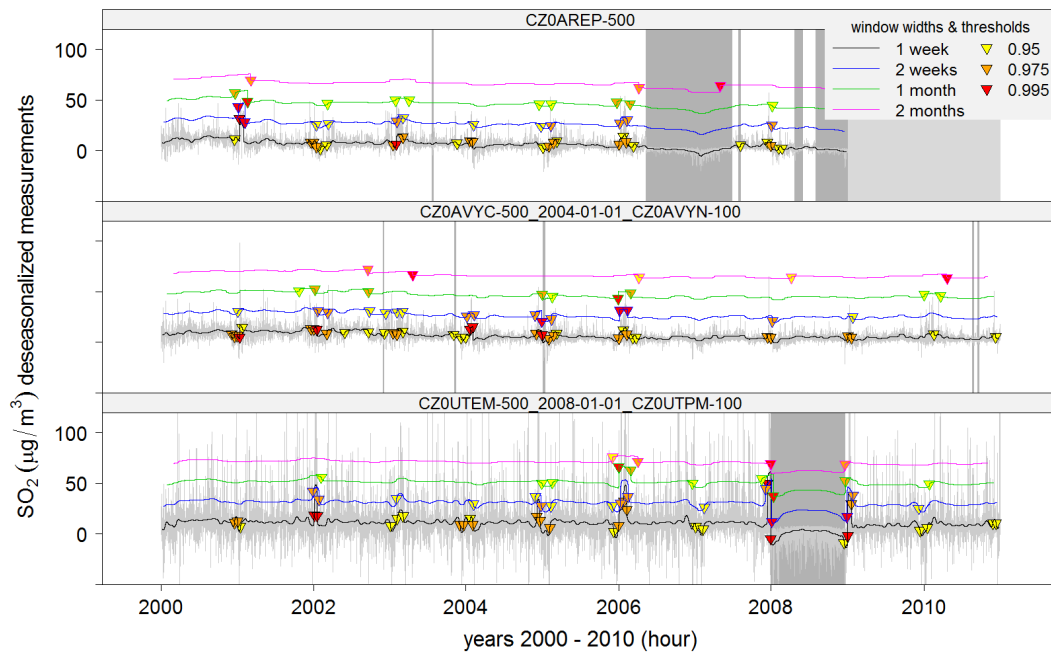


**Figure B.2:** Findings of the Kolmogorov-Zurbenko adaptive filter for varying window widths when applied to deseasonalized data.

## B.2 $SO_2$ hourly data

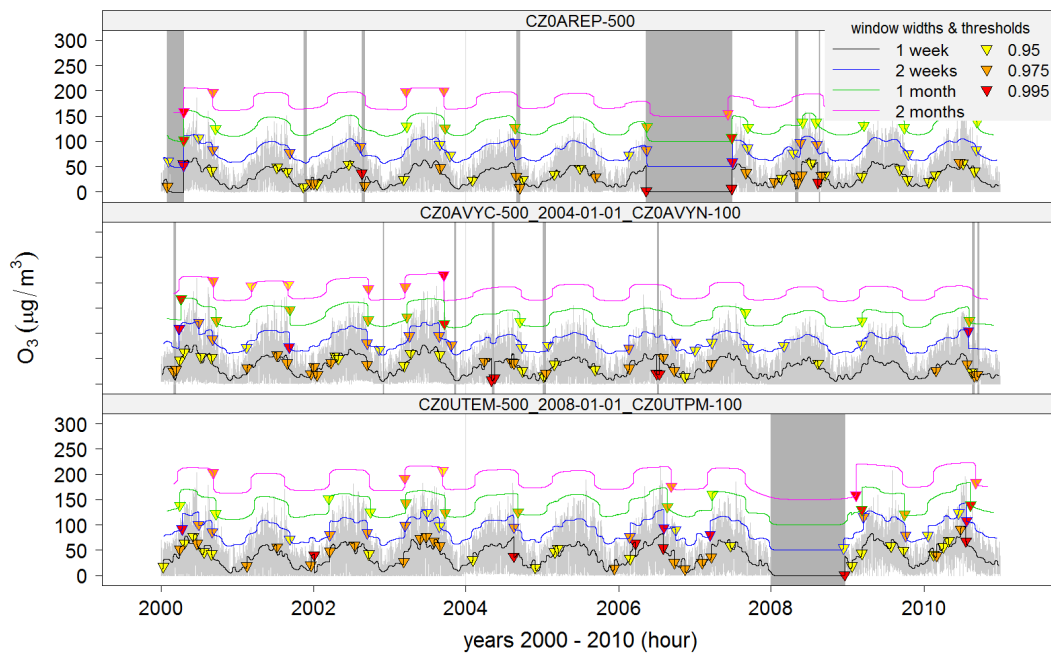


**Figure B.3:** Findings of the Kolmogorov-Zurbenko adaptive filter for varying window widths.

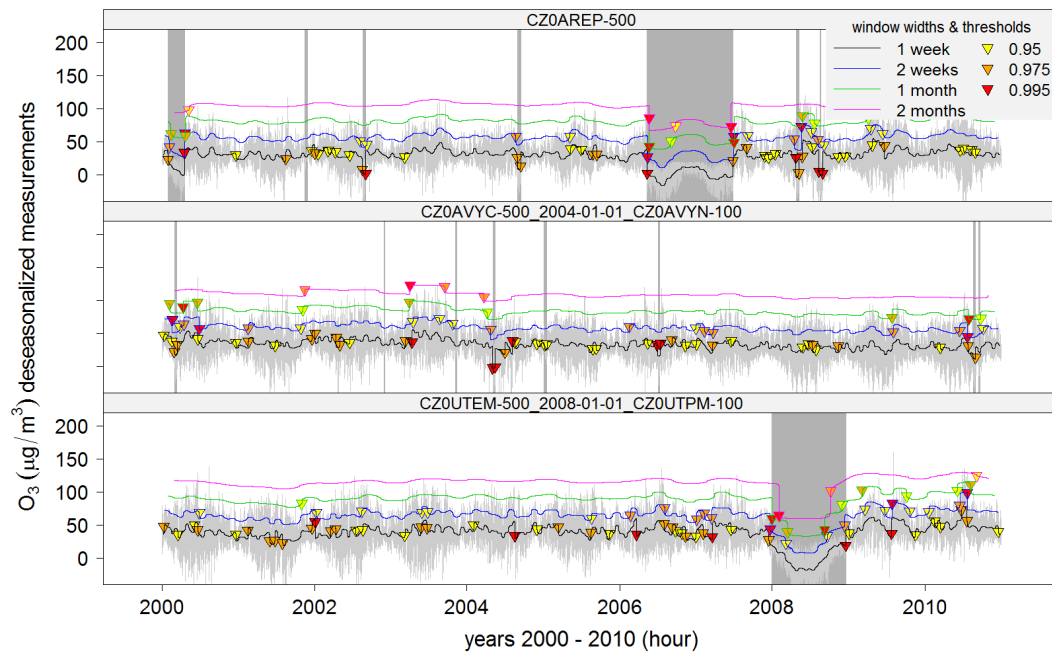


**Figure B.4:** Findings of the Kolmogorov-Zurbenko adaptive filter for varying window widths when applied to deseasonalized data.

### B.3 $O_3$ hourly data



**Figure B.5:** Findings of the Kolmogorov-Zurbenko adaptive filter for varying window widths.



**Figure B.6:** Findings of the Kolmogorov-Zurbenko adaptive filter for varying window widths when applied to deseasonalized data.