

1 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 ?? where time-constant thresholds were determined on a per station basis using the boxplot definition of outliers.

With the recommendation from ?] 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 ?] is the 'whole window - simple statistics' approach or 'two-sided median method' as described in ?] where deviations of measurements from a moving window filter are compared to a threshold statistic. We outline the method and our implementation below.

1.1 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 σ 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. ?] 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 robustness 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` [?].

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

[?] 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.

1.2 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 [?]. 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 [?] and a maximum window width of two third of a day.

1.3 Results

As an example, we illustrate results for PM_{10} measurements in Cyprus. [?] 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.

Figure 1.1 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 [?]. 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.

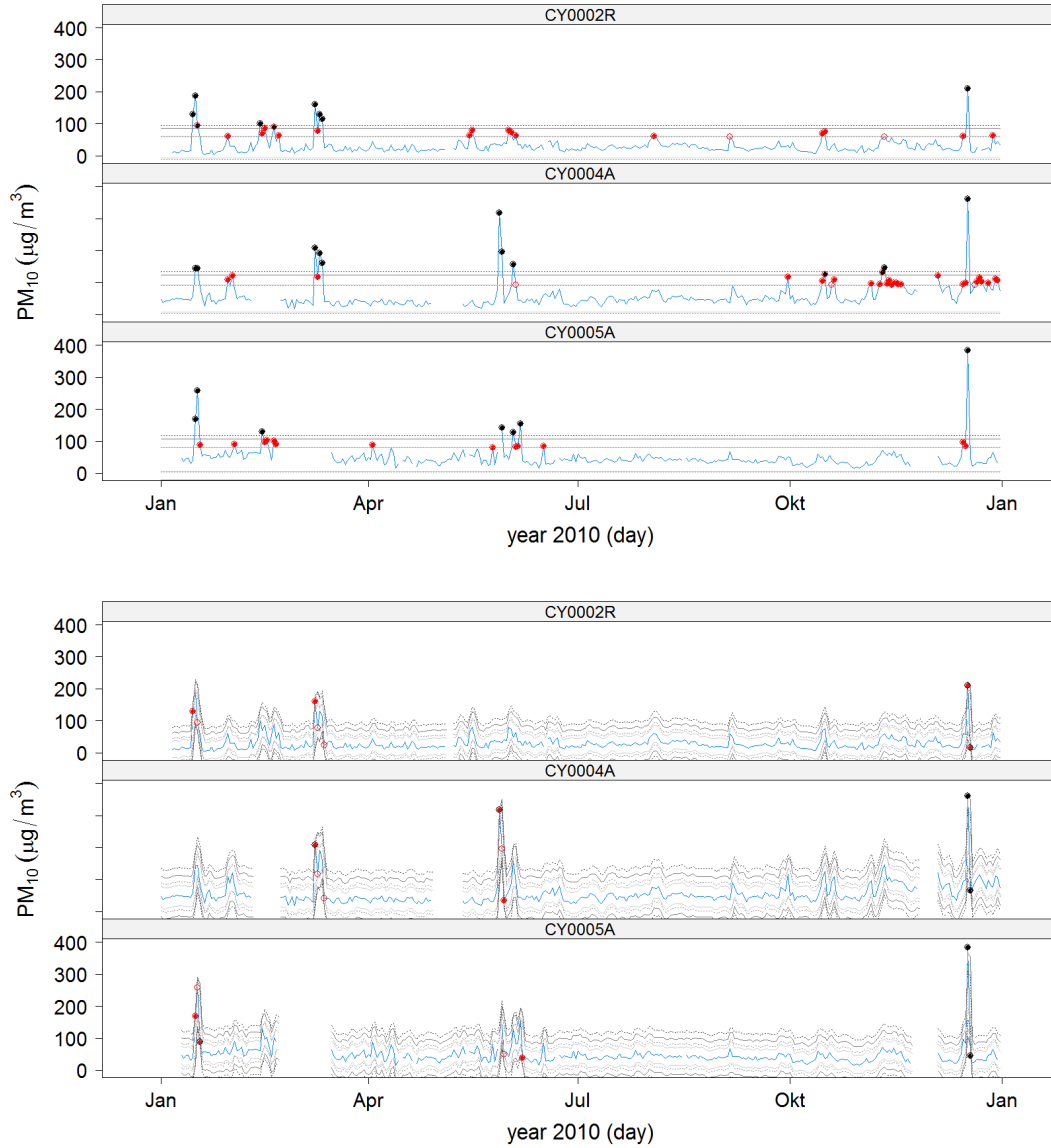


Figure 1.1: Original time series data, overview heuristics, no window (top) and extremely narrow window of width 3 (bottom).

In contrast to choosing a threshold rule, window size obviously matters. [Figure 1.2](#) and [Figure 1.3](#) 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 1.1](#), 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.

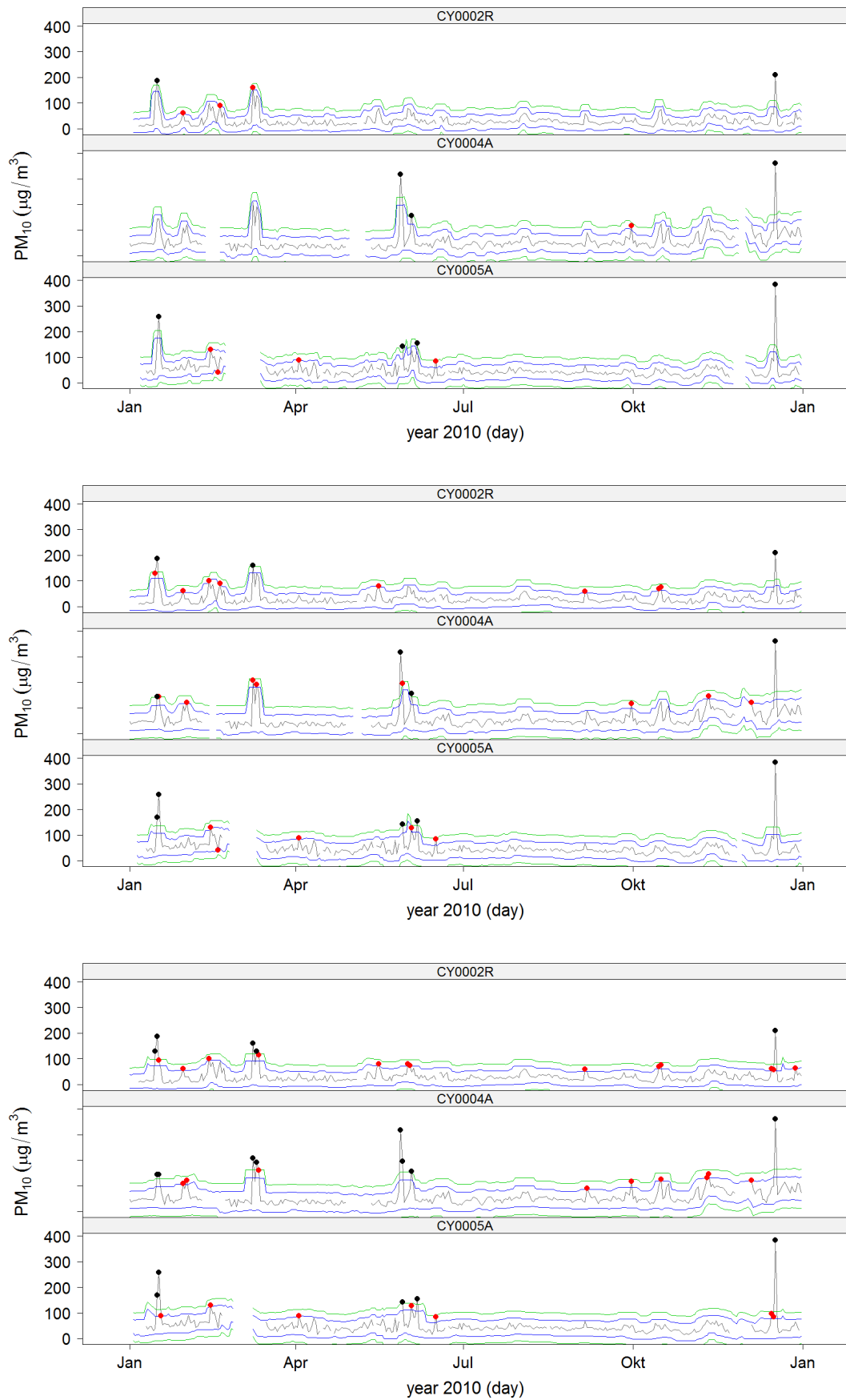


Figure 1.2: Original data, Tukey heuristic. From top to bottom: Window width = 7 days, 11 days and 15 days.

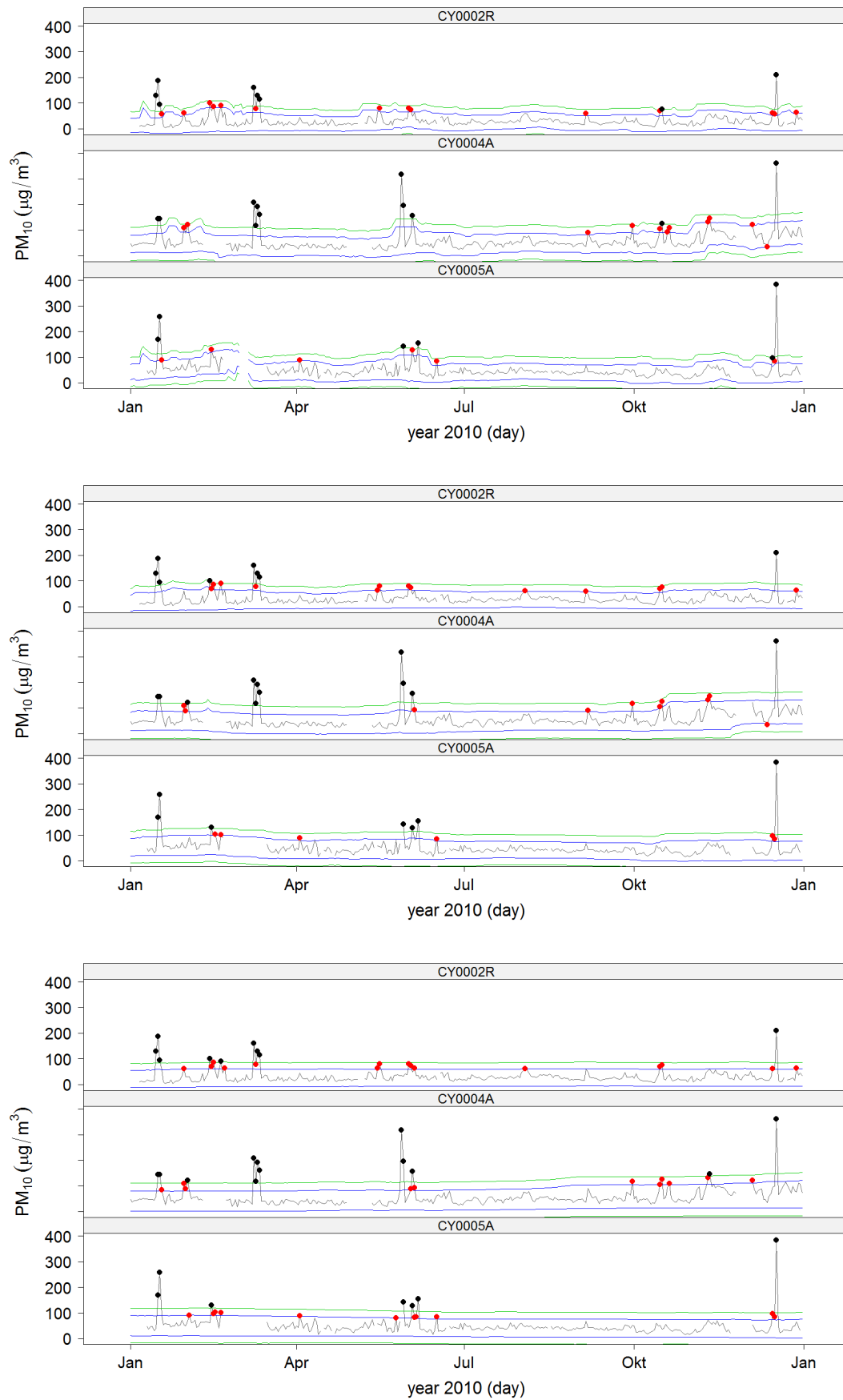


Figure 1.3: Original data, Tukey heuristic. From top to bottom: Window width = three weeks, eight weeks and eight months.

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 1.2, the Tukey and z-score case for transformed data in Figure 1.4. 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 ??, less observations are highlighted as outliers after transformation, but differences are not overly pronounced.

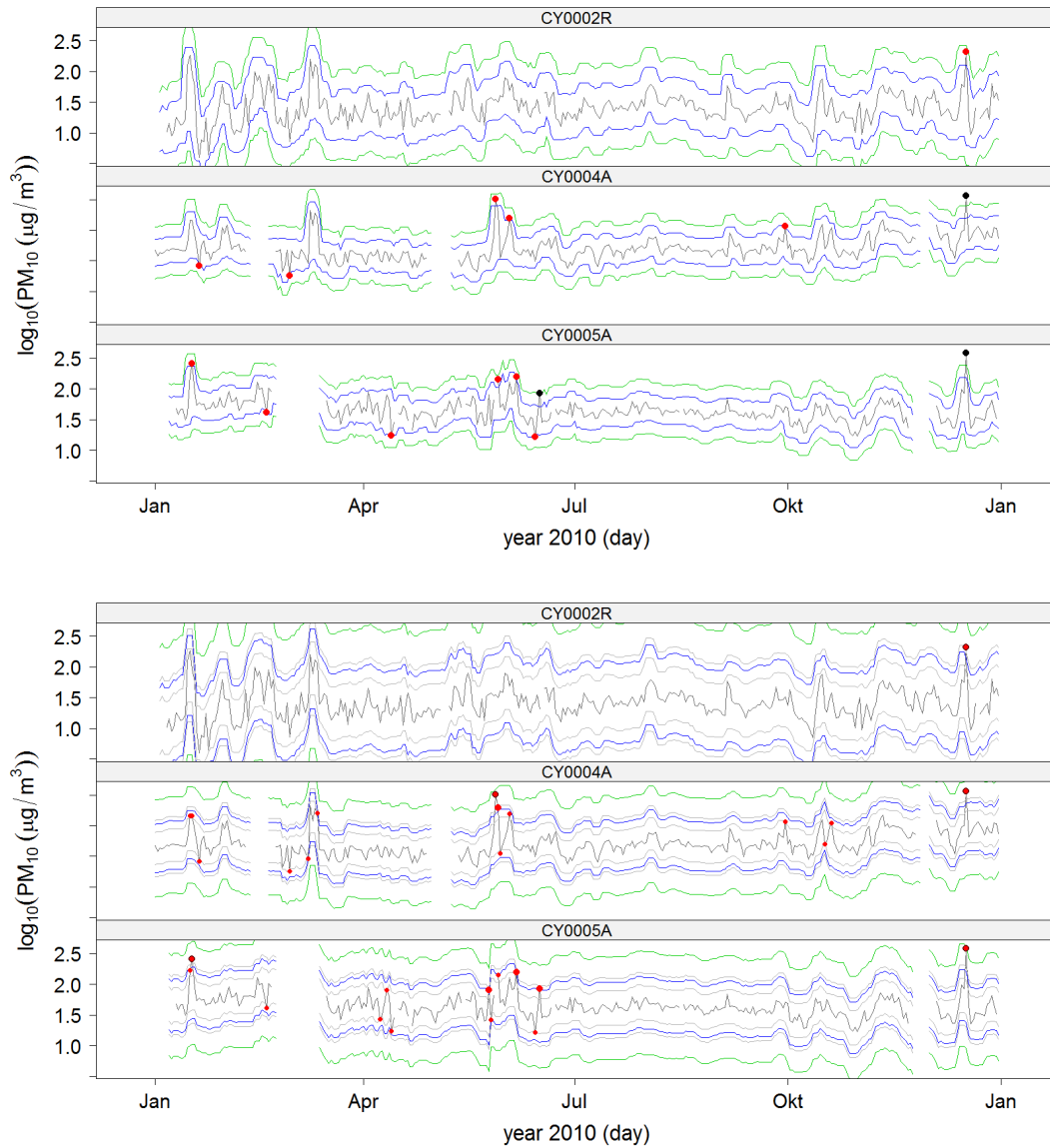


Figure 1.4: log10 transformed data, Tukey heuristic (top) and z heuristic (bottom), window width: 7 days.

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

assume 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 [?] 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 [?] where applicable. Fine-tuned parameters should be derived from simulated benchmark data. We comment on this point in the discussion in [?].

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 [?]), 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.