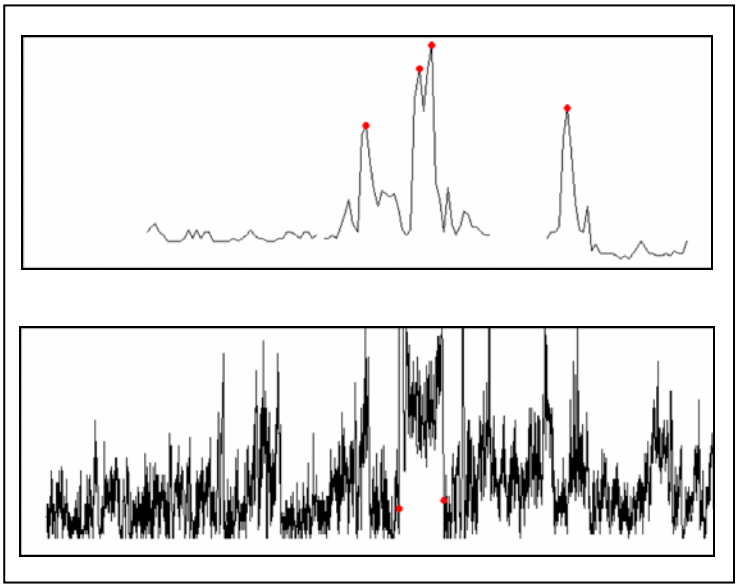
Measurement artefacts and inhomogeneity detection



ETC/ACM Technical Paper 2011/8

November 2011

Lydia Gerharz, Benedikt Gräler, Edzer Pebesma

European Topic Centre on Air Pollution and Climate Change Mitigation

The European Topic Centre on Air Pollution and Climate Change Mitigation (ETC/ACM)
is a consortium of European institutes under contract of the European Environment Agency
RIVM UBA-V ÖKO AEAT EMISIA CHMI NILU INERIS PBL CSIC

Front page picture:

Top: Example of a Type I outlier: monthly mean CO concentration time series 1997 - 2009 for a station with detected outliers (red dots) of type I (peaks, i.e., additive isolated outliers not affecting the time series; the isolated observations exceed a fixed upper and lower threshold level. Possibly caused by a measurement error); - This paper, Appendix A.1, Monthly average data, graph I.co.

Bottom: Example of a Type II.b outlier: daily mean PM_{10} concentration time series for Jan. – Jun. 2007 for Germany for a station with a detected temporal structure change or level shift (between red dots) classified as outliers of type II.b. (an innovative outlier that affects subsequent observations and as such the time series (Type II) with a class 'b' characteristic, i.e., an innovative outlier showing a transient change characteristic and as such a short-term effect in the time series. Possibly caused by wrong concentration units reported/registered or missing correction factors); - This paper, Appendix A.1, Hourly data, graph IIb.pm10.

Author affiliation:

Lydia Gerharz, Benedikt Gräler, Edzer Pebesma: Institute For Geoinformatics (IfGI), University Of Muenster

DISCLAIMER

ETC/ACM Technical Paper 2011/8
European Topic Centre on Air Pollution and Climate Change Mitigation
PO Box 1
3720 BA Bilthoven
The Netherlands
Phone +31 30 2748562
Fax +31 30 2744433
Email etcacm@rivm.nl
Website http://acm.eionet.europa.eu/

INSTITUTE FOR GEOINFORMATICS, UNIVERSITY OF MUENSTER

Measurement artefacts and inhomogeneity detection

ETC/ACM Task 1.0.2.2 - Subtask 3

ETC./ACM Technical Paper 2011/8

Lydia Gerharz, Benedikt Gräler, Edzer Pebesma

lydia.gerharz@uni-muenster.de

11.11.2011

Content

Summary

The aim of the presented study was the review and evaluation of methods for statistical detection of inhomogeneities in air quality measurement time series in AirBase. For air quality time series, two different types of inhomogeneities were identified (I) outliers and (II) structural changes or breaks.

A literature study was carried out to identify existing methods for statistical inhomogeneities detection. From the reviewed techniques, five simple stochastic methods were selected and implemented in R to be tested with air quality data. The selected methods are autoregressive lag-2 model, moving window (whole window) test, moving window (two-sided window) test, lag-1 differences and moving average filter.

Test time series with labelled inhomogeneities from AirBase for monthly and hourly data were prepared to evaluate the implemented methods. The selected methods were tested for their performance (number of correct detections) and their robustness (sensitivity of parameters for different data sets). The performance was measured using the Jaccard's coefficient φ which penalises false positive and false negative detections. Under maximisation of φ optimal parameters for each method could be estimated. Robustness was estimated by applying the methods on a validation data set with the parameters estimated beforehand using the test data sets.

For the outlier (type I) detection the moving window (whole window) test showed very good results for the performance as well as for the robustness tests. Results for the structural changes (type II) were not as clear as for the outliers. The most promising method was the moving average filter with a tendency to over-detection. However, the visual checks of the method's results indicated that the detection method of the extremes in variance of the window averages (which is the indicator for a structural change) could be improved which might help to reduce the over-detection problem. Therefore, further tests are recommended to develop a more robust method which can be used for automatic inhomogeneity detection.

1 Introduction

1.1 Aim

Inhomogeneities and measurement errors are inherent to air quality measurement time series. In the AirBase database, data quality is ensured by several checking procedures. However, some outliers may still pass first checks and propagate into the derivative products like daily or monthly averages. The aim of this working paper is to provide a categorisation of different outlier types occurring in air quality time series, to review existing methods on their capability to capture those outliers, and to test a set of pre-selected methods on different types of time series from the AirBase database. Finally, a recommendation of simple and robust methods for outlier and inhomogeneity detection in air quality time series and an outlook on the possibility of implementing an (free, open source) outlier detection tool will be made based on the results of the prototypical implementation and tests of these methods.

1.2 Scope

The data in AirBase comprises a large set of different pollutants, measurement devices, station types, temporal resolution and thus various types of possible outliers. We do not aim to give a comprehensive analysis of all existing outlier detection methods nor on a solution for all different air quality data types existing in AirBase. This report aims at providing an overview of different outlier detection methods and focuses on those using local statistics and robust parameterisation, which will be used for automatic detection of outliers in single (uni-variate) PM_{10}, O_3, CO and NO_2 measurement time series. Methods were implemented prototypically and provided in scripts for the R statistical language (Ihaka & Gentlemen, 1996). R provides an open source environment for statistical analysis of data. Parameters will be derived and tested for a number of different pollutant time series with varying temporal resolution.

2 Inhomogeneities

2.1 Inhomogeneities in air quality measurement time series

From the statistical perspective, an inhomogeneity is an observation or a number of observations which show deviations from the general pattern of the time series. This does not always mean it is a true outlier from the air quality perspective. For example, the PM_{10} measurements during New Year's Eve usually show extremely high values around midnight due to the fireworks. As this happens only once a year this is truly an outlier from the time series from a modeller's perspective, but it is not a wrong measurement. Therefore statistical outlier detection is only the first step to identify potentially "suspicious" values which has to be followed by a thorough analysis of the possible causes.

Within meteorological and air quality measurement time series, different types of outliers can occur. Typically these outliers can have various reasons. Different methods are needed to detect the different types of outliers. For example, Tsay (1988) distinguishes between additional and innovational outliers. Whereas additional outliers affect a time series only at the point in time they occur, innovational outliers change the structure of the time series from the time on they occur. In air quality measurement time series the first type can occur due to measurement errors whereas the second one can be caused for example by changes in instrumentation.

In this report, based on the analysis of time series from AirBase, we distinguish between these two main types as "Outliers" and "Structure changes" whereas the latter one includes different definitions of inhomogeneities:

I. Outliers:
   Additive outlier → only one observation, does not affect rest of the time series (Peak)

II. Structure changes/Level shifts:
   Innovation outlier → affects time series (subsequent observations)

   a. Persistent change → Break (abrupt change, e.g., change of instrumentation)

      a.1. Change in mean

      a.2. Change in minimum

   b. Transient change → short-term effect on the process

Table 1: Categorisation of inhomogeneities in air quality time series.

Type	Description	Possible cause
I	Outliers	measurement error
IIa1	Break in mean	change/replacement in monitors, data transfers, change in reporting units (e.g., initially µg/m3 , later ppb)
IIa2	Break in minimum	change in treatment of data below detection limit
IIb	Transient change	wrong concentration unit, missing correction factor

Table 1 gives an overview of the inhomogeneity types identified in air quality time series and possible reasons causing these outliers. Of course, these definitions depend on the temporal resolution of the data. For example, outliers are more often found in time series with a higher temporal resolution, such as hourly data, and are averaged out when moving to coarser resolutions. Structural changes on the other hand are more regularly found in coarser resolved time series (see next section for examples).

2.2 Examples from AirBase

Inhomogeneities occur at different temporal resolutions and for different pollutants. Thus examples are taken from AirBase for monthly and hourly averages of CO, NO_2, SO_2 and PM_{10}. The data shown here was also part of the test data used in the method evaluation (see appendix A.1 Test data plots).

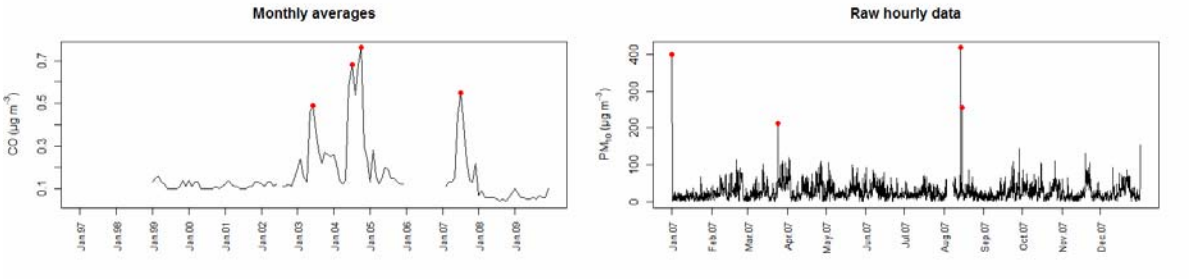


Figure 1: Type I outlier examples for monthly CO and hourly PM_{10} data.

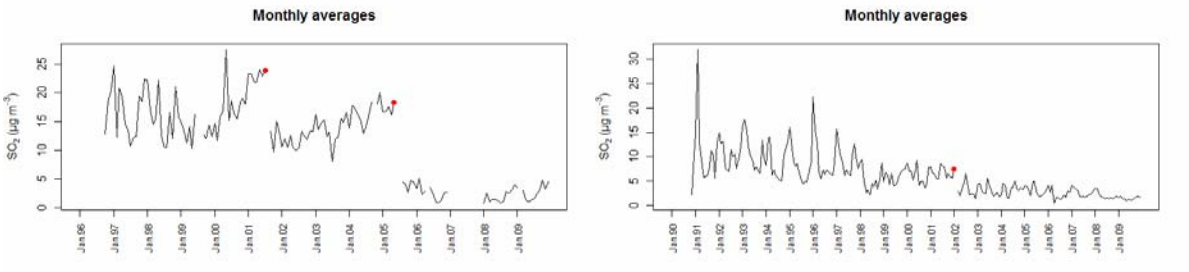


Figure 2: Type IIa1 permanent change (break in mean) examples for monthly SO_2 data.

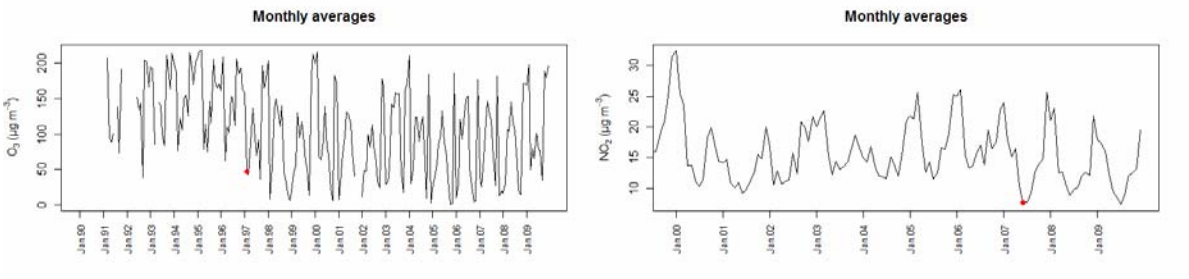


Figure 3: Type IIa2 permanent change (break in minimum) examples for monthly O_3 and NO_2 data.

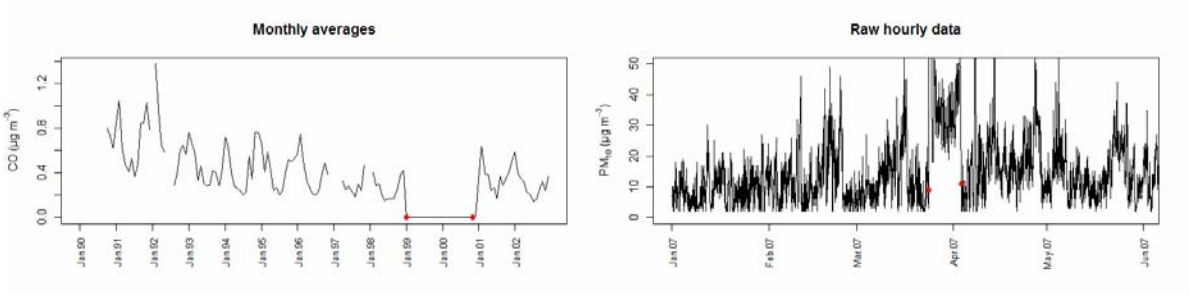


Figure 4: Type IIb transient change examples for monthly CO and hourly PM_{10} data

3 Current inhomogeneity checks in Airbase

Raw data uploaded annually to AirBase undergoes beforehand the quality control of the responsible member state. If the member states use the Data Exchange Module (DEM) for submitting the data, additional quality checks are performed:

Outliers (type I) checks by:

- Exceedance of fixed upper and lower threshold levels

Structural changes (type II) checks by:

- Yearly mean calculated from the submitted data is zero
- Yearly mean calculated from the submitted is negative
- Yearly mean is more than three times lower or higher than previous yearly mean

These checks for just that reporting year qualify all concentrations above or below a threshold value as being “questionable”. The upper and lower threshold values are pollutant dependent but kept constant over the years and, in general, similar for all countries and station types. This simple and unrefined filter is efficient in detecting extreme isolated outliers (type I) and some flagrant structural extremes of type II, such as concentrations in wrong units or inserted default codes in the concentration field indicating for example ‘missing value’ or ‘value below detection limit’. Many inhomogenieties and component specific behaviour in measurement patterns will be missed.

When suspicious data is detected via these checks, a report is send back to the member state with request for correction. The member state corrects data or accepts identified outliers which are then uploaded to AirBase.

4 Review of methods for detecting inhomogeneities

A plethora of different methods for detecting outliers and breaks in time series are available in literature. Inhomogeneities and outliers in times series are usually detected by statistical models or approximations of the "general" shape of the curve to detect deviations from this shape. Those models use statistical properties, for example the autoregression between values at different times, to describe the time series. Outliers can be defined as deviation from the shape described by the statistical properties. Such statistical outliers do not necessarily have to be true outliers as exceptional measurements in air quality can be caused by specific conditions, such as extreme emission events. A general way to model time series is to decompose the series up into different components (see for example Chatfield, 2004):

- Trend component, could be linear or non-linear
- Periodic component, e.g., daily or seasonal patterns usually approximated by sinus or cosinus functions
- Random component, i.e. white noise

The shape of the curve can be estimated globally by a parametric model (see 4.1/4.2) or locally within a window surrounding the value under investigation (4.3). Besides investigating a single time series at a time, multivariate methods use time series of nearby stations or different parameters at the same station to detect outliers (4.4). However, not all methods could be investigated and tested here. As the detection of inhomogeneities should be performed unsupervised, i.e. automatic, not every method in literature is equally suited for our purposes. In this report we will focus on simple, statistical methods for detecting outliers and structural changes. Other methods will be presented and briefly discussed in chapter 4.5.

4.1 ARIMA time series model

Detecting and removing inhomogeneities is important for robust estimation of parameters for time series models used ,e.g., for prediction. One commonly used model type is the autoregressive integrated moving average (ARIMA) model (Chatfield, 2004). The ARIMA model consists of an autoregressive and a moving average component. The autoregressive (AR) component of order p is:

$$y_t = \varepsilon_t + \sum_{i=1}^{p} a_i y_{t-i}$$

The observation y at time t can be estimated by the sum of the previous observations within lag p and weighted with a_i plus an additional error term ε_t. The moving average (MA) model describes the observation at time t as the sum of errors within lag q weighted by b_j:

$$y_t = \sum_{j=1}^{q} b_j \varepsilon_{t-j}$$

The ARIMA model combines both models to one with the order p and q (orders of its AR and MA components):

$$y_t = \varepsilon_t + \sum_{i=1}^{p} a_i y_{t-i} + \sum_{j=1}^{q} b_j \varepsilon_{t-j}$$

4.1.1 Modelling inhomogeneities

Numerous studies have dealt with the problem of detecting inhomogeneities using ARIMA models. Such approaches use a parametric, deterministic or stochastic, function to model the outliers, while the function is independent from the underlying ARIMA model fitted to the time series (Tsay, 1988). Thereby additional outliers (type I) as well as innovational outliers (structural changes, type II) can be considered and detected by hypothesis tests.

Methods based on the ARIMA model require parametric fitting of the model to complete times series or subsets of time series which requires considerable knowledge and prior analysis of the time series. For example, the time series needs to be cleaned from trend and seasonal components before the ARIMA model can be fitted (Fox, 1972). Thus the automated, unobserved inhomogeneity detection using this type of models remains difficult.

4.1.2 Autoregression model (AR2)

A simplified version of the ARIMA model uses only the autoregressive properties of the time series. For a small lag in a time series of e.g. air quality measurements the values usually show a strong autocorrelation as the changes in air quality parameters are not completely random and happen slowly.

Zhang et al. (2011) used an AR model of order 2 to detect outliers in time series of air temperature and relative humidity. They found the detection useful in application of wireless sensor networks. As the method only uses retrospective data it can be used in real-time settings to detect outliers.

4.2 Distribution-based inhomogeneity detection

In the distribution based inhomogeneity detection outliers are defined as outliers of a parametric distribution function representing the time series observations. Usually these methods require a fitting of a distribution to the whole time series. Grubb's test for outliers (Grubbs, 1950) offers a simple way of outlier detection but requires the assumption of a normal distribution. An outlier is detected by calculating a test statistic G which is the maximum deviation of an observation from the mean of the time series divided by the standard deviation of the time series. If G exceeds a certain threshold derived from the distribution function, the null hypothesis of no outliers is declined and the respective value is identified as outlier in the time series. By repeating this procedure, one outlier at a time can be identified until all are removed from the time series.

Another approach was followed by Van der Loo (2010). He used linear regression of Quantile-Quantile-plots of the measurements and the cumulative distribution function. In the extreme values package in R he offers two methods to identify outliers: (I) by determining the threshold above which only a certain number of observations is expected and (II) by deriving a test statistics if the extreme value could be drawn from the designated distribution. However, the implementation is not able to deal with data point with invalid or not available (NA) values in the time series and was therefore excluded from further analysis.

4.3 Moving window filters

A widely used group of methods to check for inhomogeneities are based on examining the local neighbourhood of each point in time, i.e. all neighbouring values within a certain time

window. The advantage of these methods is that no assumptions about stationarity of the time series have to be made. For an appropriate window size the influence of a global trend and periodic component can be ignored. Besides the window size and threshold values for inhomogeneity detection, no additional parameters need to be fitted. The optimal size of the window depends on the change rate within the time series and might be constant or adaptive. In the case of the AirBase measurement data the window size will likely be different for different temporal resolutions, possibly also for different pollutants.

4.3.1 Whole window – Simple statistics (MW)

Basu & Meckesheimer (2007) demonstrated the use of window statistics on flight data recorder measurements on altitude and roll angle. In their method, the deviation of a measurement from the median of the surrounding measurements was used as detection methods for outliers (type I) in the sensor data. Alternatively to the median, the mean can be used although it is more sensitive to outliers.

For the whole window statistics thresholds for the deviation of a measurement from the window mean/median $\Delta_{t,\mu}$ need to be estimated. In Basu & Meckesheimer (2007) absolute values depending on the phenomenon were chosen. Ideally, the thresholds are derived from the time series itself, e.g., by its variability. This can be done by using either the local standard deviation within the window or the global standard deviation estimated for the whole time series. Thereby values are classified as thresholds if they fall outside the range of +/- the standard deviation σ_y times a factor f of ,e.g., 3. The standard deviation can be estimated locally within the window or globally for the whole time series. An outlier test would then look like:

$$if\,(\Delta_{t,\mu} < -f\sigma_y \mid \Delta_{t,\mu} > f\sigma_y) = true$$

4.3.2 Two-sided window – Simple statistics (MW2)

A method to detect structural changes is to compare statistics between the lower, i.e. containing previous measurements, and the upper, i.e. containing future measurements, half of a window surrounding the value. Depending on the type of structural change, different statistics might be useful, e.g., a change in mean (type IIa1) could be assessed by comparing the means in both window halves. Potential test statistics are:

- T-Test on significant differences of means in both window halves
- Testing against thresholds of differences between both window halves in
    - Means
    - Medians
    - Minima
    - Maxima
    - Quantiles
    - Variance

While for the T-Test a level α of significance is required, the thresholds for the other methods can be derived in a similar way as for the whole window approach but using a factor for the global standard deviation of the values as threshold.

4.3.3 Lag-1 differences (LAG1)

Instead of looking at the measurements, the first derivative (lag 1 differences) can be used to allow conclusions about outliers. Horálek et al. (2004) presented a methodology using the distribution of lag-1 differences $\Delta_{t,t-1} = y_t - y_{t-1}$ between the measurements within a window to detect inhomogeneities. Therefore a double exponential (Laplace) distribution is fitted to the lag-1 differences per window and upper and lower quantiles derived from this distribution are tested against thresholds. Three error types can be distinguished:

Type a

One difference falls outside the quantiles of probability α or $1-\alpha$. This corresponds to an extreme growth or decline within the window (structural change, type II).

$$if\,(\Delta_{t,t-1} < q_{\alpha} \mid \Delta_{t,t-1} > q_{1-\alpha}) = true$$

Type b

Two subsequent differences fall outside the same quantiles of probability β or $1-\beta$. This corresponds to an ongoing increase or decline in values (structural change, type II).

$$if\,(\Delta_{t,t-1}\ \&\ \Delta_{t+1,t} < q_{\beta} \mid \Delta_{t,t-1}\ \&\ \Delta_{t+1,t} > q_{1-\beta}) = true$$

Type c

Two subsequent observations fall outside different quantiles of probability χ and $1-\chi$. This corresponds to an increase and decrease or decrease and increase of values, i.e. negative or positive peak, within one interval (outlier, type I).

$$if\,((\Delta_{t,t-1} < q_{\chi}\ \&\ \Delta_{t+1,t} > q_{1-\chi}) \mid (\Delta_{t,t-1} > q_{1-\chi}\ \&\ \Delta_{t+1,t} < q_{\chi}) = true$$

4.3.4 Moving average filter (MA filter)

Rao & Zurbenko (1994) used a moving average filter as a low pass filter to smooth upper-air measurement time series. Applying the moving average filter iteratively separates seasonal pattern and the trend from the short term variations that are averaged out. Thus, structural changes are preserved and cleaned from random components. For the smoothed time series the variance per window is calculated. At break (type II) locations local maxima for the variance remain. Detecting these maxima can be tricky, as the variance varies continuously with peaks of different size. It is critical to determine which peak size allows the conclusion of a true structural change. This can be done by, e.g., first separating the highest percentage of variance values and then detecting in each region the peaks in variance (see Figure 5 for illustration).

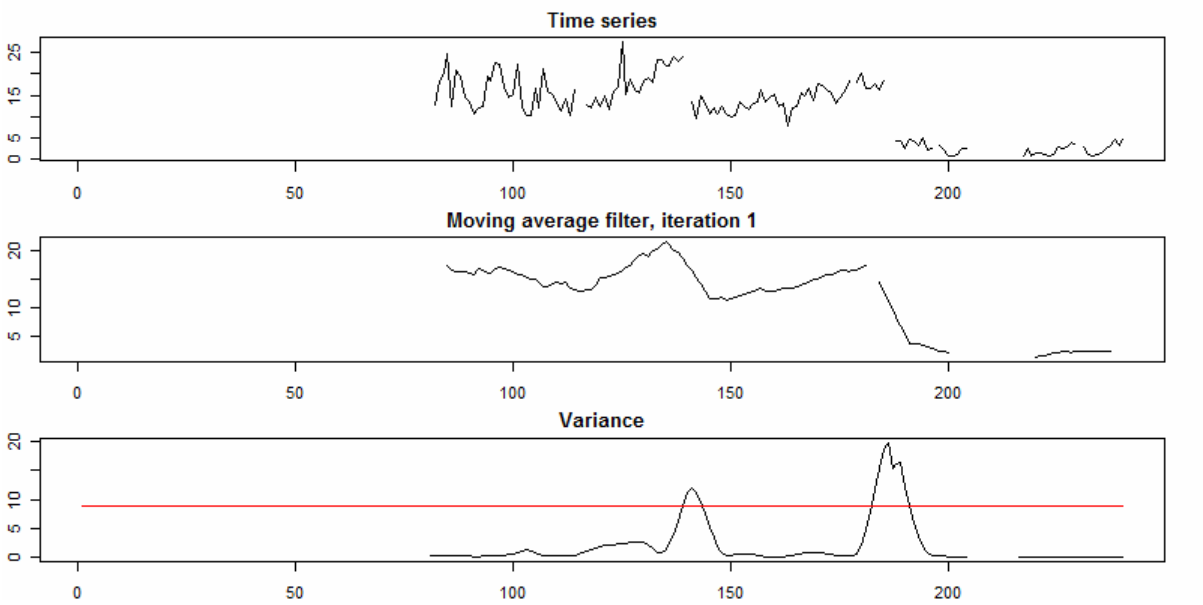


Figure 5: Raw and, filtered time series with the variance of the filtered value. Local maxima in variance are detected by cutting of the upper percentage (red line) and estimate the maximum within each region.

In Zurbenko et al. (1996) the moving average filter approach was extended to an adaptive version. The adaptive filter dynamically adjusts length of the window depending on the rate of change (the more change, the smaller the window size). Therefore the data first needs to be smoothed by the non-adaptive moving average filter. Then the values can be used to calculate the rate of change (as differences between smoothed data points) to estimate the size of the adapted filter. If the differences are decreasing the half window size behind the data point is reduced, if differences are increasing, the half before the data point is reduced. This adaptive version requires additional computing time but possibly yield better results if the rate of change is highly variable within the time series.

4.4 Multivariate methods

Instead of taking only the actual time series into account, using additional data from the same or neighbouring stations can be helpful to separate true inhomogeneities due to, e.g., measurement errors from regional effects. Although these techniques were not implemented for this report, a review of methods is presented here for consideration in future work.

4.4.1 Reference time series

For climate data, i.e. annual averages, a common method is building reference time series from nearby stations. The aim is thereby to determine which stations will have least inhomogeneities and are comparable to the time series of interest. Alternatively to using neighbouring measurements of the parameter, covariate time series of other parameters or pollutants from the same station can be used for similar analyses if they are correlated.

Potter (1981) tested each time series against a mean series created from the nearby stations. By repeating this for each station he could determine which stations had inhomogeneities compared to the majority of the other stations. In the double mass curve analysis (Kohler, 1949) cumulative sums of candidate and reference time series (mean of several stations) are plotted together. Breaks in the line indicate structure change. Alternatively the residuals between candidate and reference time series can be tested for randomness in the residuals. The standard normal homogeneity test (SNHT, Alexandersson, 1986) detects non-randomness

which is an indicator for outliers or breaks. As another simple approach, multiple linear regression can be performed on the candidate series using nearby reference series as independent variables to predict the candidate series (Vincent, 1998). If the regression residuals are still auto-correlated, the time series is divided into segments. Thus, the time series is divided into homogeneous segments, separated by structural changes.

4.4.2 Spatio-temporal analysis

Spatio-temporal analysis of measurement time series takes neighbouring time series together with their distance to the candidate time series into account. Spatial or spatio-temporal interpolation can help to identify outliers (type I). As shown by Gräler et al. (2011) cross-validation of interpolations leaving one station out can reveal outliers.

4.5 Other methods

Numerous other methods for inhomogeneities exist but were excluded from further analysis as they are not well suited for simple, automated application. Peterson et al. (1998) gives a comprehensive overview of further inhomogeneity detection methods for climate data.

One group that should be mentioned here is imputation methods. Imputation aims to substitute missing values removed as outliers and can therefore also be applied to detect inhomogeneities by leaving out suspicious values. One commonly used method is interpolation of the missing value using nearby valid measurements. To detect outliers, each value can be interpolated and the deviation between the interpolated and measured value be compared to a defined threshold. Junninen et al. (2004) summarise a number of imputation methods from simple ones like Nearest Neighbour and regression-based imputation to more complex ones like self-organising maps and neural network multi-layer perceptron techniques.

5 Method evaluation

From the methods identified in the literature review a set of simple methods was selected for implementation and evaluation with AirBase data. The focus during this initial selection was on methods which require few parameterisation and can be run automatically. The further evaluation tests focused on

- Performance, i.e., correctly detected outliers
- Computing time, especially for large amounts of data, like hourly measurement time series
- Robustness, to allow prior parameterisation and application on new data sets without new calibration

Performance can be measured in numbers of correctly detected outliers (true positive and true negatives), over-detected outliers (false positives) and under-detected outliers (false negatives) as shown in Table 2. A higher performance requires minimisation of false negatives and false positives. A simple parameter to measure for the performance is the Jaccard's coefficient φ as used by Basu and Meckesheimer (2007):

$$\varphi = \frac{1}{1 + \left(\frac{fp + fn}{tp} \right)}$$

Thus, the coefficient will always be positive with a maximum of 1. Very small values for φ (~ 0) indicate bad performance of the method while the perfect detection of outliers would lead to $\varphi = 1$. Note that in the coefficient the number of true negatives is not included as this number is usually much higher than the other counts and would dilute the influence of these. In this version of the coefficient fp and fn get the same weight, which can be adapted if either under- or over-prediction is considered more important.

Table 2: Overview of performance measures for outlier detection.

	Method – not outlier	**Method – outlier**
Ground truth – not outlier	Clean data points – true negatives (tn)	Overdetected outliers – false positives (fp)
Ground truth – outlier	Underpredicted outliers – false negatives (fn)	Outliers – true positives (tp)

In this chapter first the data used for evaluation is introduced. Then the initial implementation of the methods with the necessary parameters will be described. Next, selected time series from AirBase with identified outliers were used to estimate the optimal parameters for the implemented methods under maximisation of Jaccard's coefficient. Finally, the robustness and the computation time per method are estimated for validation data sets using the optimised parameters.

5.1 Data

5.1.1 Synthetic data

Following the approach of Zurbenko et al. (1996) synthetic data produced under controlled conditions was used to initially test the method implementations. The data was created from a linear trend, seasonal pattern (sinus function) and random noise sampled from a Gaussian distribution. Type I inhomogeneities were added by randomly selecting values to which the mean of the trend was added or substracted (see Figure 6). Type II inhomogeneities were added with size 0.1-0.5 σ of the white noise component (see Figure 7).

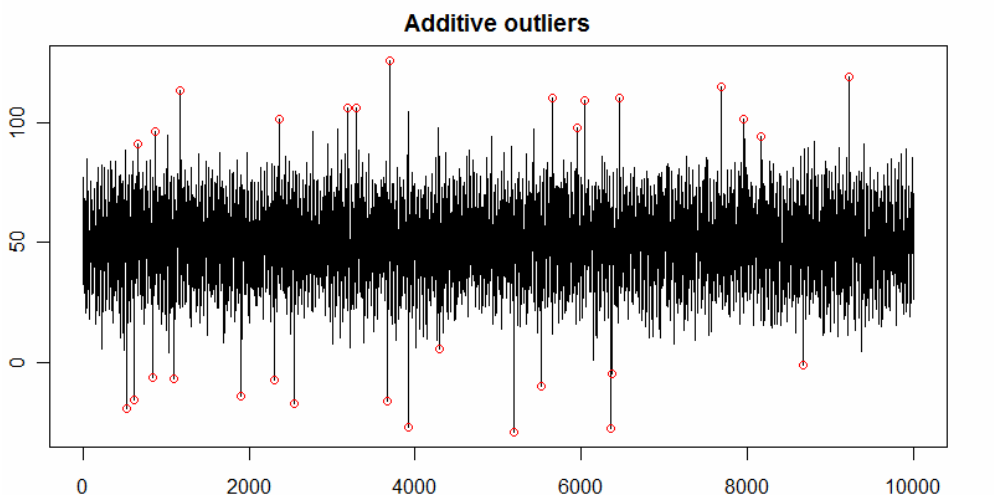


Figure 6: Synthetic data with type I inhomogeneities.

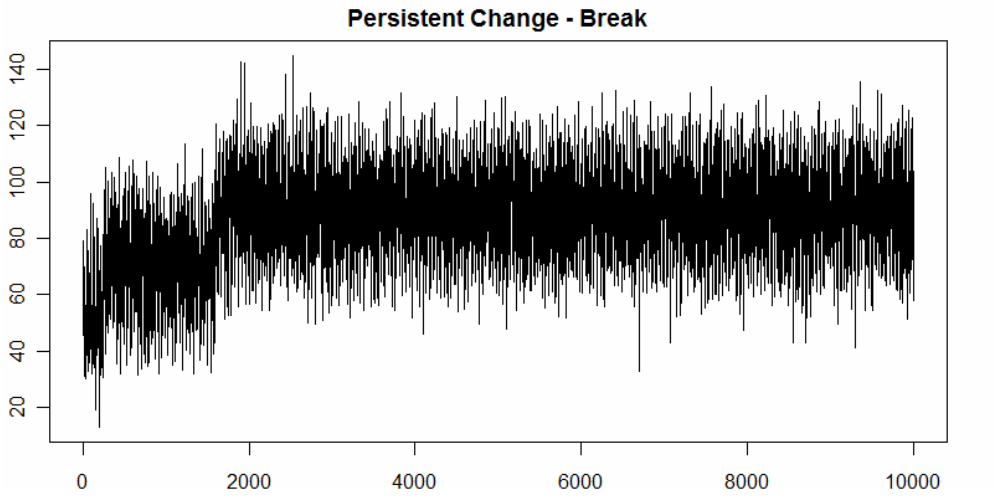


Figure 7: Synthetic data with type II inhomogeneities.

5.1.2 AirBase data

As test data sets time series of different pollutants and temporal resolution were selected from AirBase to estimate the optimal parameters per method. Specifically data available was:

- Monthly data of CO, NO_2, O_3, SO_2 (1990/97-2009) for Europe
- Hourly data of PM_{10} (2007) for Germany

From these data sets, time series with inhomogeneities of type I, IIa1, IIa2, IIb and clean time series as reference, depicted as type III, were selected as shown in Table 3. For each time series the locations of the inhomogeneities was recorded to allow the calculation of φ. Plots of the time series are in appendix A.1 Test data plots.

For each time series or part of a time series (e.g. within a window) the number of missing values was estimated. Only if > 75 % of the data was available, the respective statistics were calculated. Therefore, outliers in regions with many missing values might be missed by some of the methods, depending also on the size of the window used for the analysis.

Table 3: Time series from AirBase selected as test data.

Time series ID	Outlier type	Pollutant	Temporal resolution
I.co	Outlier	CO	Monthly
I.so2	Outlier	SO_2	Monthly
I.pm10.1	Outlier	PM_{10}	Hourly
I.pm10.2	Outlier	PM_{10}	Hourly
IIa1.so2.1	Permanent change in mean	SO_2	Monthly
IIa1.so2.2	Permanent change in mean	SO_2	Monthly
IIa1.so2.3	Permanent change in mean	SO_2	Monthly
IIa2.o3.1	Permanent change in minimum	O_3	Monthly
IIa2.o3.2	Permanent change in minimum	O_3	Monthly
IIa2.no2	Permanent change in minimum	NO_2	Monthly
IIb.co	Transient change in mean	CO	Monthly
IIb.pm10	Transient change in mean	PM_{10}	Hourly
III.co	Clean data series	CO	Monthly
III.o3	Clean data series	O_3	Monthly
III.no2	Clean data series	NO_2	Monthly
III.pm10.1	Clean data series	PM_{10}	Hourly
III.pm10.2	Clean data series	PM_{10}	Hourly

Additionally, for the tests on performance and robustness a different collection of validation data sets was selected. Therefore for each type I, II and III one monthly and one hourly time series was chosen, resulting in 6 validation time series.

5.2 Method implementation

Each selected method was implemented as a function in the R statistical environment (Ihaka & Gentlemen, 1996). The functions take the time series and additional parameters like window size and threshold values and return the IDs of detected outliers. Furthermore a number of helper functions were implemented to, e.g., get the values within a moving window filter which could be reused by a number of methods. The implemented outlier detection methods were:

AR2 (AR lag-2 model)

- Name: ar.outliers
- Parameters: raw time series, lag, sd_factor (for threshold)

MW (Moving window – whole window)

- Name: mw.outliers.sd
- Parameters: raw time series, half window size, difference function (mean, median), sd_factor (for threshold), differences per window (alternatively if moving window differences have been calculated beforehand)
- Name: mw.outliers.sd.local
- Parameters: raw time series, half window size, difference function (mean, median), sd_factor (for threshold), differences per window (alternatively if moving window difference have been calculated beforehand)

MW2 (Moving window – two-sided window)

- Name: mw2.outliers.sd
- Parameters: raw time series, half window size, difference function (mean, median, variance, minimum, maximum, quantiles), sd_factor (for threshold), values per window (alternatively if moving window values have been calculated beforehand)
- Name: mw2.outliers.t.test
- Parameters: raw time series, half window size, p-threshold (test probability), values per window (alternatively if moving window values have been calculated beforehand)

LAG1 (Lag-1 differences)

- Name: lag1.outliers.all
- Parameters: raw time series, half window size, distribution type (normal, Laplace), probabilities (for type a, b and c errors), type of error (a, b, c), moving window lags (pre-processed, alternatively)
- Requirements: package VGAM (for the Laplace distribution)

MA filter (Moving average filter)

- Name: kz.outlier
- Parameters: raw time series, half window size, number of iterations, adaptive (Boolean if filter should be adaptive), quantile (for selecting upper values and detect extremes), plot (boolean if results should be plotted)

Jaccard's coefficient φ

To evaluate the performance of the method the calculation of the Jaccard's coefficient was implemented as well:

- Name: jaccard
- Parameters: raw time series, detected outliers (by method), true outliers (as labelled), tolerance (number of ids the detected outliers might be shifted compared to the true outliers), plot (boolean if results should be plotted)

All implementations were tested with the synthetic data for their performance. Pre-test results showed rather poor results for the AR2 method in correctly detecting outliers. Furthermore the MA filter in the adaptive setting showed long calculation times. Therefore, the adaptive version of the moving average filter was excluded from the performance tests on the AirBase data sets.

5.3 Parameter optimisation

Each method was tested with the selected type I or type II time series depending on the error types the method could detect. The type III time series without outliers were used in each method as reference check for over-detection, i.e. the methods should detect here no outliers. For the parameters per method, ranges and a step size within these ranges were defined. Each time series was tested for each combination of parameters in the method by running the functions in loops. For each run the used parameters and the resulting value for φ were stored to evaluate the optimal combination of parameters afterwards.

The evaluation took place for monthly and daily data separately. For methods where different versions of the methods were available, e.g., using different statistics, distribution types, threshold methods or number of iterations, the version with the highest number of the overall maximum for φ was selected. It was desirable to select only one version of the method to be used for further tests. For the selected version the average φ values over all window sizes (except for AR2) and threshold values were calculated respectively. Those parameters leading to the upper 5 % of the φ averages were considered as optimal.

Using the averages of the Jaccard's coefficient allowed the identification of an optimal range of window size and threshold values instead of a single value. Optimal parameters were given per time series, for all time series and for all inhomogeneities (type I/II) time series together.

In this section, for each method the used parameter ranges for the tests are given first. The resulting optimal values per time series and combined analysis using average φ from all time series (or grouped by outlier type I and II) are listed in the Tables 4 to 8. For well performing methods level plots of φ per window size and threshold value are shown for the combined analysis. In these plots the threshold values (thr) are given on the x-axis and the window sizes (q) are given on the y-axis. For each combination of threshold and window size the resulting value for φ is given as grey shade. Areas with high values for φ are depicted darker than areas with low φ indicating bad performance. Averages of φ over rows and columns were used to identify the optimal values for threshold and window size, respectively as given in the tables. In contrast to the table values, the plots give an idea of the robustness of the results for neighbouring values. For instance, if the areas in the plot change slowly it means that slight changes in the parameters do not change the performance considerably. In contrast, random distributions of lighter and darker areas with immediate changes lead to the conclusion of non-robust parameters where small differences can lead to large changes in the performance. Plots per time series are given in appendix A.2 Parameter optimisation plots per time series.

For the robustness tests the combined results of the time series containing inhomogeneities (excluding the clean data series) per method were used. This was necessary as the results for the clean data series showed often a larger range and thus larger insensitivity especially if the method performed bad and detected no outliers at all. This lead to a large number of high φ values for clean data series which influenced the combined analysis negatively.

5.3.1 AR2

Parameter range

Threshold values

Standard deviation factor: Min 1.5, Max 20, Step Size 0.5

Results

Table 4: Parameter optimisation for AR2 method.

Data set	Maximum φ	Threshold value
I.co	0.17	5-6
I.so2	0.5	8-11
III.co	1	13.5-20
III.o3	0	-
III.no2	1	13.5-20
All monthly	**1**	**14.5-17.5**
I.pm10.1	0.005	4.5
I.pm10.2	0.02	13
III.pm10.1	0	-
III.pm10.2	0	-
All hourly	**0.02**	**13**

Plots were omitted here as only one parameter was left and the results were rather poor for all time series.

5.3.2 MW

Method versions

Statistics used: mean, median

Threshold method: Standard deviation global (*Sd global*) , standard deviation local (*Sd local*)

Parameter range

Window size: Min 1, Max 50, Step size 1

Threshold values

Sd factor local: Min 1.5, Max 6, Step size 0.5

Sd factor global: Min 1.5, Max 6, Step size 0.5

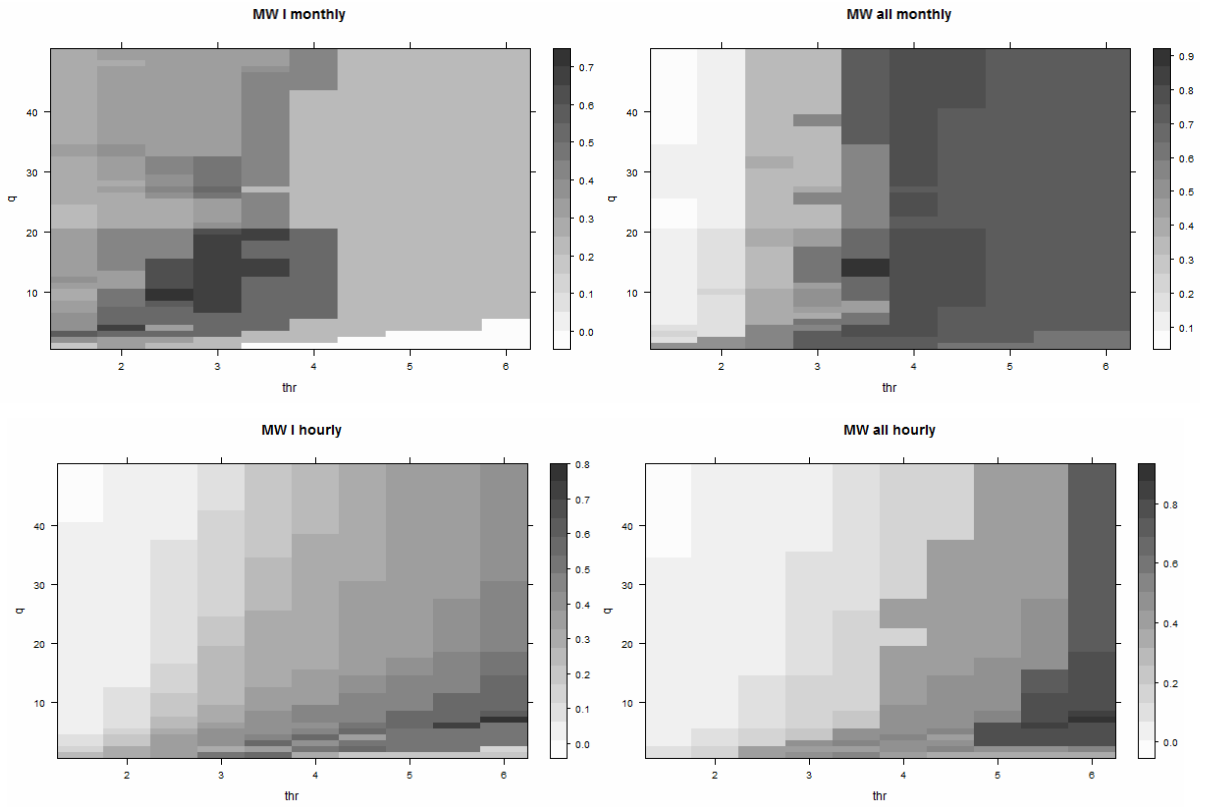
Results

Table 5: Parameter optimisation for the MW method.

Data set	Maximum φ	Statistic	Threshold method	Threshold value	Window size
I.co	0.75	Median	Sd global	1.5	31, 32, 41-50
I.so2	1	Median	Sd global	3.5, 4	8-20
III.co	1	Mean	Sd global	3.5-6	1-3 (all)
III.o3	1	Mean	Sd global	2.5-6	5-50 (all)
III.no2	1	Mean	Sd global	3-6	1-3 (all)
I monthly	**0.7**	**Mean**	**Sd global**	**3**	**9-15**
All monthly	**0.9**	**Mean**	**Sd global**	**4**	**1-4, 13-15**
I.pm10.1	1	Mean	Sd global	5.5-6	6,7
I.pm10.2	0.6	Mean	Sd global	3.5-6	1-14
III.pm10.1	1	Mean	Sd global	5-6	all
III.pm10.2	1	Mean	Sd global	5-6	all
I hourly	**0.7**	**Mean**	**Sd global**	**5.5, 6**	**6,7**
All hourly	**0.8**	**Mean**	**Sd global**	**6**	**2-6**

Differences between mean and median results are very small. As mean is calculated slightly faster, this is proposed as optimal parameter.

Plots



5.3.3 MW2

Method versions

Statistics used: mean, median, minimum (*Min*), maximum (*Max*), variance (*Var*), quantiles 0.05, 0.25, 0.75 and 0.95 (*Q05, Q25, Q75, Q95*)

Threshold method: Standard deviation global (*Sd global*), T-Test probability

Parameter range

Window size: Min 10, Max 50, Step size 1

Threshold values

Sd factor global: Min 0.2, Max 1.1, Step size 0.1

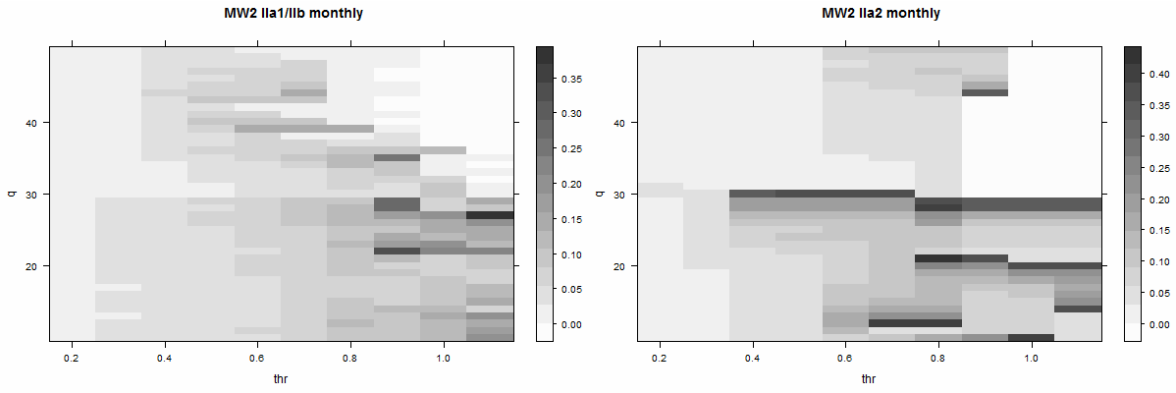
T-Test probabilities: Min 0.001, Max 0.01, Step size 0.001

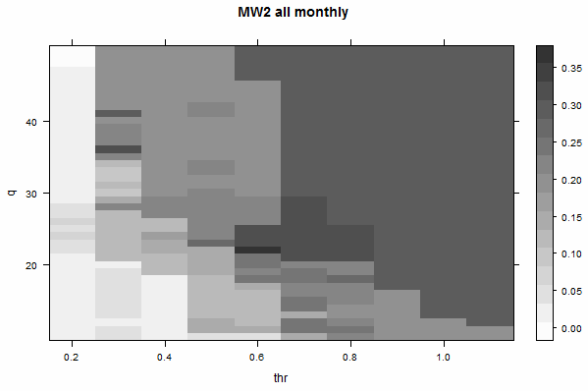
Results

Table 6: Parameter optimisation for the MW2 method.

Data set	Maximum φ	Statistic	Threshold method	Threshold value	Window size
IIa1.so2.1	0.5	Q75	Sd global	0.7	35, 39, 40
IIa1.so2.2	1	Q75	Sd global	0.9	10, 13, 22
IIa1.so2.3	1	Q75	Sd global	0.9	27-29
IIa2.o3.1	1	Q05	Sd global	1	12, 19, 22
IIa2.no2	1	Q25	Sd global	0.8	28-30
IIa2.o3.2	1	Min	Sd global	1.1	28-30
IIb.co	0.5	Mean	Sd global	1.1	11-12, 26-27
III.co	1	Q05	Sd global	0.7-1.1	45-50
III.o3	1	Var	Sd global	0.7-1.1	21-50
III.no2	1	T-Test	T-Test	0.001-0.002	10-15, 21-26, 33-37, 45-50
IIa1/IIb monthly	**0.3**	**Q75**	**Sd global**	**0.9**	**22, 26-29**
IIa2 monthly	**0.4**	**Min**	**Sd global**	**0.8**	**20, 27-30**
All monthly	**0.3**	**Var**	**Sd global**	**1.1**	**36, 41, 46-48**
IIb.pm10	1	Var	Sd global	1	12,14
III.pm10.1	0	-	-	-	-
III.pm10.2	0	-	-	-	-

Plots





5.3.4 LAG1

Method versions

Distributions used: Normal, Laplace

Threshold method: a, b, c

Parameter range

Window size: Min 2 (for hourly data 10), Max 50, Step size 1

Threshold values

a: for monthly data - Min 0.00005, Max 0.00015, Step size 0.00001, for hourly data - Min 0.001, Max 0.01, Step size 0.001

b: Min 0.001, Max 0.01, Step size 0.001

c: Min 0.001, Max 0.01, Step size 0.001

Results

Table 7: Parameter optimisation for the LAG1 method.

Data set	Maximum φ	Distribution	Threshold method	Threshold value	Window size
I.co	0.25	Normal	a	all	all
I.so2	0	-	-	-	-
IIa1.so2.1	0	-	-	-	-
IIa1.so2.2	0	-	-	-	-
IIa1.so2.3	0	-	-	-	-
IIa2.o3.1	0	-	-	-	-
IIa2.o3.2	0	-	-	-	-
IIa2.no2	0	-	-	-	-
IIb.co.1	0	-	-	-	-
III.co	1	Normal/Laplace	a	all	all
III.o3	1	Laplace	a	all	all
III.no2	1	Normal/Laplace	a	all	all
All monthly	-	Normal/Laplace	**a**	**all**	**all**
I.pm10.1	0	Normal/Laplace	a	all	all
I.pm10.2	0.02	Normal/Laplace	a	all	all
IIb.pm10	0.25	Normal/Laplace	a	all	all
III.pm10.1	0.02	Normal/Laplace	a	all	all
III.pm10.2	0.00	Normal/Laplace	a	all	all
All hourly	-	Normal/Laplace	**a**	**all**	**all**

Plots were omitted here as the results were poor for all time series.

5.3.5 MA filter

Method versions

Number of iterations: 1-4

Parameter range

Window size: Min 2 (hourly data: 10), Max 50, Step size 1 (hourly data: 2)

Threshold values

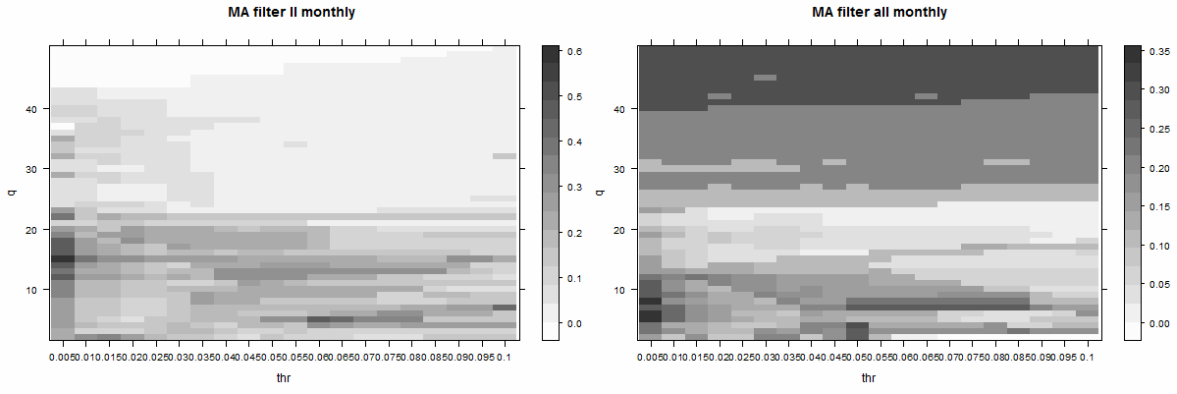
Maximum quantiles: Min 0.005, Max 0.1 (hourly data: 0.05), Step Size 0.005

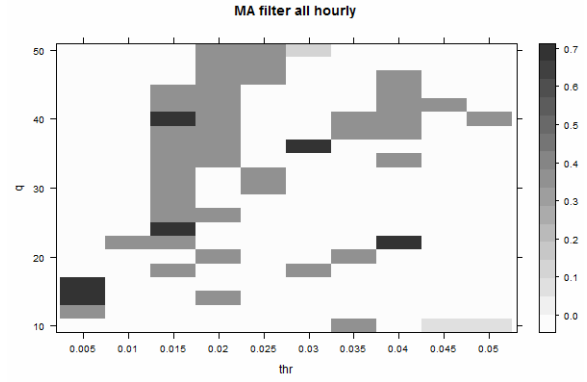
Results

Table 8: Parameter optimisation for MA filter method.

Data set	Maximum φ	Iterations	Threshold value	Window size
IIa1.so2.1	1	4	0.005	7
IIa1.so2.2	1	1	0.045	15
IIa1.so2.3	1	1	0.005	4
IIa2.o3.1	1	3	0.005	3
IIa2.no2	1	1	0.005	32
IIa2.o3.2	1	1	0.005	22
IIb.co	0.67	3	0.005	3, 7-8
III.co	1	4	0.05-0.065	25-50
III.o3	1	4	0.045-0.05	43-50
III.no2	1	4	0.01-0.07	28-50
II monthly	**0.63**	**1**	**0.005**	**1-20**
All monthly	**0.3**	**4**	**0.01**	**43-50**
IIb.pm10	0.5	1	0.025	18, 24
III.pm10.1	1	4	0.015, 0.04	40
III.pm10.2	1	4	0.015	36, 40, 44, 46
All hourly	**0.7**	**4**	**0.015**	**22, 36-42**

Plots





5.4 Performance and robustness tests

Based on the optimal parameters evaluated in section 5.3, performance tests for each method were carried out on validation data sets for type I, II and III for monthly and hourly resolution (i.e., one time series per type). Therefore the optimal parameters for each method as determined in the previous sections was tested with parameter values taken from estimated optimal ranges (see Table 4-8). The parameter combinations with the best results are shown in Table 9 and Table 10. The resulting values for φ serve as indicators how robust the methods worked with the same parameters for different data sets. The computation time was measured to compare which methods are better suited to process large data sets in a considerable time frame. These values are only for a first orientation as the method implementations in R were not optimised on computation time and might be improved for this purpose. Running time was estimated with the R 64bit version running on a Windows 7 system installed on an Intel® Core™ 2 Duo P9600 (2.53GHz, 1066MHz, 6MB L2 Cache) processing unit with 4 GB main memory.

Table 9: Computing time and results for φ per method with optimised parameters for monthly data.

Parameter	**AR**	**MW** mean, sd global	**MW2 IIa1/b** Q75, sd global	**MW2 IIa2** min, sd global	**LAG1** Laplace	**MA filter** 1 iteration
Window size	-	14	28	28	25	15
Threshold value	15.5	3	0.9	0.8	0.0001, 0.0005, 0.0025	0.005
Average computation time [s]	0.34	0.01	0.22	0.22	0.05	0.01, 0.03
φ (I)	0.33	1	0	1	0	0
φ (II)	1	1	0.036	0.037	0	1
φ (III)	0	1	0	1	1	0

Table 10: Computing time and results for φ per method with optimised parameters for hourly data.

Parameter	**AR2**	**MW** mean, sd global	**MW2** Var, sd global	**LAG1** Laplace	**MA filter** 4 iterations
Window size	-	6	12	25	24
Threshold value	13	6	1	0.0001, 0.0005, 0.0025	0.0015
Average computation time [s]	14.07	0.55	1.30	1.83	1.66
φ (I)	0	1	0	0	0
φ (II)	0	1	0	0	0.5
φ (III)	0	1	0	1	0

6 Recommendation and conclusions

6.1 Summary results per method

AR2

The autoregressive model showed very poor results. One reason might be that an order of 2 is too small, and only short-term variations are identified as outliers, whereas outliers persisting over larger periods were missed. As the computation took relatively long and probably would slow down with increasing lag size, this method is not recommended for the aim of this study.

MW

The Moving window method yielded very good results for all data sets with a minimum of 0.6 for the maximum Jaccard's coefficient φ (Table 5). The results were generally better for larger thresholds and for the hourly data for small window sizes (<10). With a very short running time and very good results on the validation data sets the moving window method using the deviation from the window mean seems to be a robust and useful method for outlier detection.

MW2

The two-sided moving window method showed good results for the monthly data sets but failed for the hourly raw data. As both clean data sets got φ values of zero there is a clear over-prediction in outliers by the method. Furthermore, the plots in appendix A.2 Parameter optimisation plots per time series indicate that the parameters are not very robust and yield only good results for very specific combinations in threshold values and window size. The application of the optimised parameters (Table 9) showed mixed results, with better performance of the minimum comparison. Results on hourly validation data sets were very poor with a strong over-prediction.

Interestingly, the comparison between minimum of the upper and lower window half showed the best results whereas the T-Test was less efficient. This might be due to the relative large number of observations the T-Test needs in each window half. Due to many invalid values (NAs) in the time series it was probably not always possible to reach the required count.

LAG1

The lag-1 difference method failed completely in our tests. For the data sets φ was larger than 0, the optimal threshold and window values covered the whole range. Therefore a sensible application of this method was not possible. A correct detection only occurred for clean data sets as the method under-predicted all types of inhomogeneities. However in pre-tests using the synthetic data, the method worked quite well for type I outliers with a probability c=0.02.

One reason might be the problem of how to deal with NAs. If an NA occurred during a temporal break, no lag-1 differences were available and thus no inhomogeneity could be detected. Using the last value before the break could be a solution, but might be error prone as well. Another reason might be that the probability values which were chosen accordingly to Horálek et al. (2004) are too small.

MA filter

The Moving Average Filter showed reasonable results in the separate time series analysis (Table 8) but weak performance for the validation data. The large differences in results for outliers and clean data sets especially for hourly data (see plots in appendix A.2 Parameter

optimisation plots per time series) indicate an over-detection of outliers. This over-detection shows the sensitivity of the method to very small breaks. Zurbenko et al. (1996) were able to detect breaks much smaller and thus less visible than the ones we assigned as outliers in our AirBase data sets. The method might be able to detect very small changes in the structure which are not assigned as inhomogeneities by us. When looking at the plots of variance it gets clear that the efficiency of the filter could be better if the detection of extremes would be improved. So far, the risk of over-detection in clean data sets is high, as also very small changes in the variance could be assigned as maxima. Iterating the filter more often helps as the results in Table 8 show. However, iterating too often could dilute the results for true structural changes.

6.2 Recommended methods

For **type I** outliers the **Moving Window** approach calculating the deviation of single values from the average of the surrounding values seems to be a fast, effective and robust method. As the results for mean and median were very similar, the use of the mean might be recommended as the calculation is slightly faster. As threshold value a multiplicative factor of the standard deviation for the whole time series seems to be most suitable. This factor is about 3 to 6 depending on the temporal resolution. The window size also needs to be adapted for different temporal resolutions. However, the values are probably stable enough between different pollutant types that they do not need to be adapted. However, these parameter values should be validated by further tests.

The **Moving Average Filter** is the most suitable method for detecting **type II** structural changes. Results were less robust than for the MW detection method as shown in section 5.3. Compared to the MW2 approach the MA filter showed a better potential to be used for data sets with different temporal resolution. The main reason for poor performances seemed to be a clear over-prediction due to the detection method for extremes in the variance. As the method is sensitive to very small breaks, for clean data set an over-prediction occurred. In Figure 8 the MA filter analysis of the type III (clean) validation data set used for the raw hourly data is shown. It is clear that even small breaks which are considered to be negligible get variance peaks even after several iterations.

To make use of the MA filter, more analyses have to take place. First, the definition and assignment of breaks in test data has to be reconsidered thoroughly. The detection of local extremes in the variance could be improved which would prevent false predictions for clean data sets as it occurred here.

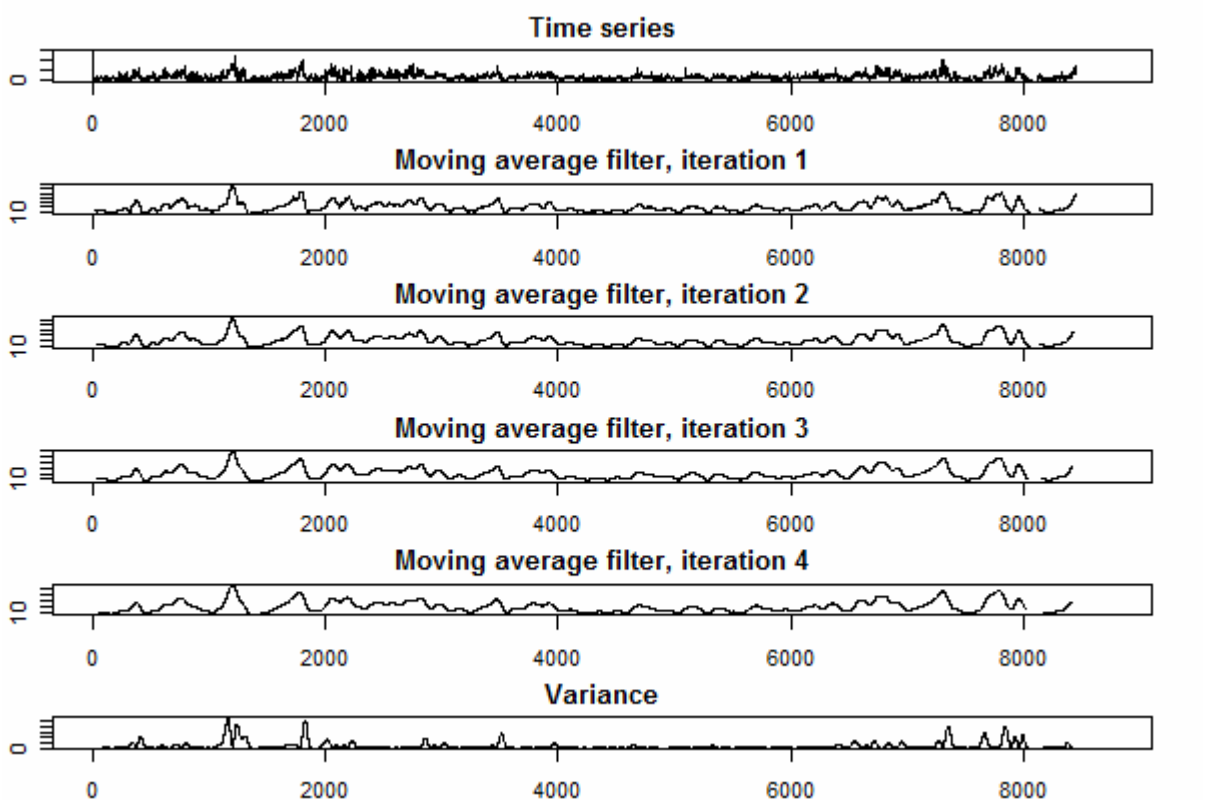


Figure 8: Moving average filter results for four iterations and the variance of the fourth iteration for clean raw data validation data set.

6.3 Conclusion and outlook

The aim of the presented work was the review, testing and evaluation of suitable methods for detecting inhomogeneities in air quality measurement time series. The focus was on simple methods (purely statistical approach) to allow simple, automatic detection with the long term aim of implementing a tool allowing such outlier analysis automatically. For each of the two identified groups of inhomogeneities, i.e. outliers and structural changes, a useful method could be identified. Both methods show reasonable to very good results for the limited number of data sets used for the presented study. For final tuning of the parameters and evaluation of the methods, a larger number of time series than considered in this report should be included.

The interpretation of the Jaccard's coefficient φ is limited. For example, for clean data sets a detection of even one outlier leads to a value of 0. For outliers, the performance depended strongly on the correct assignment of outliers manually which was not always easily possible as some changes could be small and therefore overseen. Looking at the MA filter results, for example, revealed much better results as when looking only at the φ values. Furthermore, the Jaccard's coefficient could be adapted by changing the weights for either over- or under-prediction. In the current form both are weighted equally.

For the future implementation of a free and open source tool, the implemented R functions used for this report could form the basis; they are available upon request from the authors. An open issue is how to integrate such a tool with the existing or future infrastructure of reporting and processing of data that feed into AirBase. One aspect of this work was on computation time to allow relatively fast checks on outliers. It is possible that too complicated methods with long run time would prevent users from applying the tool on their data. On the other hand, when outlier detection is needed for validating data, it is carried out relatively

infrequently and only the best methods should be considered, even when they are computationally expensive.

In this report we only consider single time series at a station, whereas in principle time series of nearby stations in similar conditions or time series of other pollutants at the same station should be informative and could help to establish whether measurements are outliers or not. The further integration of spatio-temporal methods for outlier detection requires additional information about the spatial location of the stations providing the time series. Thus, a detailed requirement analysis on the intended use cases and user groups and the technical infrastructure would be necessary to develop such a tool.

References

AirBase, European air quality database, http://airbase.eionet.europa.eu/

Alexandersson, H. 1986. A homogeneity test applied to precipitation data. Journal of Climate, 6:661–675.

Chatfield, C. (2004). The Analysis of Time Series. An Introduction. 6th edition. Texts in statistical science. Chapman & Hall/CRC.

Basu, S. and M. Meckesheimer (2007). Automatic outlier detection for time series: an application to sensor data. Knowledge and Information Systems 11(2);137-154.

Fox, A. J. (1972). Outliers in Time Series. Journal of the Royal Statistical Society. Series B 3:350-263.

Gräler, B., Gerharz, L., Pebesma, E. (2011). Spatio-temporal (3D) analysis and interpolation. ETC/ACM Technical Paper 2011/10 *(in prep.; to be released at http://acm.eionet.europa.eu/reports/#tp)*

Grubbs, F. E. (1950). Sample Criteria for Testing Outlying Observations. The Annals of Mathematical Statistics. 21(1):27–58.

Horálek, J., J. Fiala, M. Brabec and J. Hrdá (2004). Detection of gross errors in air pollution data sets. In: QA/QC in the field of emission and air quality measurements: harmonisation, standardisation and accreditation (pp. 242-248). European Commission, Luxembourg.

Ihaka, R., Gentleman, R. (1996). A language for data analysis and graphics. Journal of Computational and Graphical Statistics 5:299-314.

Junninen, H., Niska, H., Tuppurainen, K., Ruuskanen, J. and M. Kolehmainen (2004). Methods for imputation of missing values in air quality data sets. Atmospheric Environment 38:2895-2907.

Kohler, M. A. 1949. Double-mass analysis for testing the consistency of records and for making adjustments. Bull. Amer. Meteorol. Soc., 30:188–189.

Peterson, T.C., Easterling, D.R., Karl, T.R., Groisman, P., Nicholls, N., Plummer, N., Torok, S., Auer, I., Boehm, R., Gullett, D., Vincent, L., Heino, R., Tuomenvirta, H., Mestre, O., Szentimrey, T., Salinger, J., Forland, E., Hanssen-Bauer, I., Alexandersson, H., Jones, P. and D. Parker (1998). Homogeneity adjustment of in situ atmospheric climate data: a review. International Journal of Climatology 18:1493-1517.

Potter, K. W. 1981. Illustration of a new test for detecting a shift in mean in precipitation series. Mon. Wea. Re6., 109:2040–2045.

Rao, S. T., and I. Zurbenko (1994). Detecting and tracking changes in ozone air quality. Journal of Air and Waste Management Association 44:1089-1095.

Tsay, R.S. (1988). Outliers, Level Shifts, and Vairance Changes in Time Series. Journal of Forecasting 7:1-20.

Van der Loo, M. P. J. (2010). Distribution based outlier detection in univariate data. Discussion paper (10003). Statistics Netherlands, The Hague/Heerlen.

Vincent, L. 1998. A technique for the identification of inhomogeneities in Canadian temperature series. J. Climate, 11, 1094–1104.

Zurbenko, I., P. S. Porter, S. T. Rao, J. Y. Ku, R. Gui and R. E. Eskridge (1996). Detecting Discontinuities in Time Series of Upper-Air Data: Development and Demonstration of an Adaptive Filter Technique. Journal of Climate 9:3548-3560.

Zhang, Y., N. Hamm, N. Meratnia, A. Stein, P. Havinga (2011). Statistically based outlier detection for wireless sensor networks. Poster presented at the 1st conference on Spatial Statistics – Mapping global change.

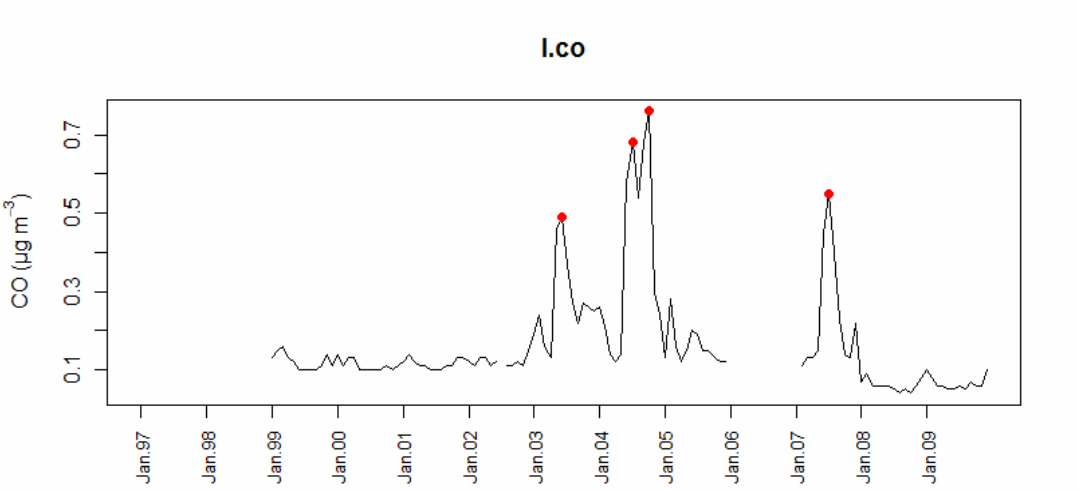
Appendix/Supplementary data

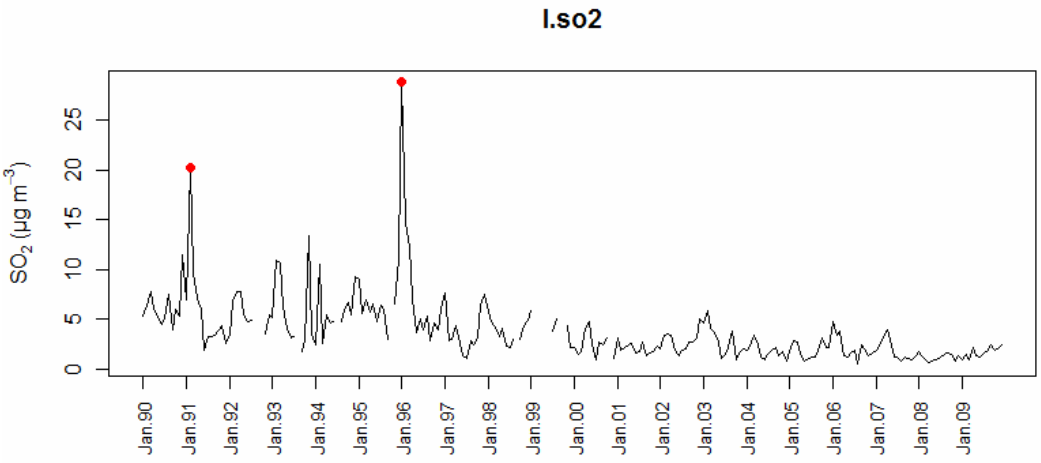
A.1 Test data plots

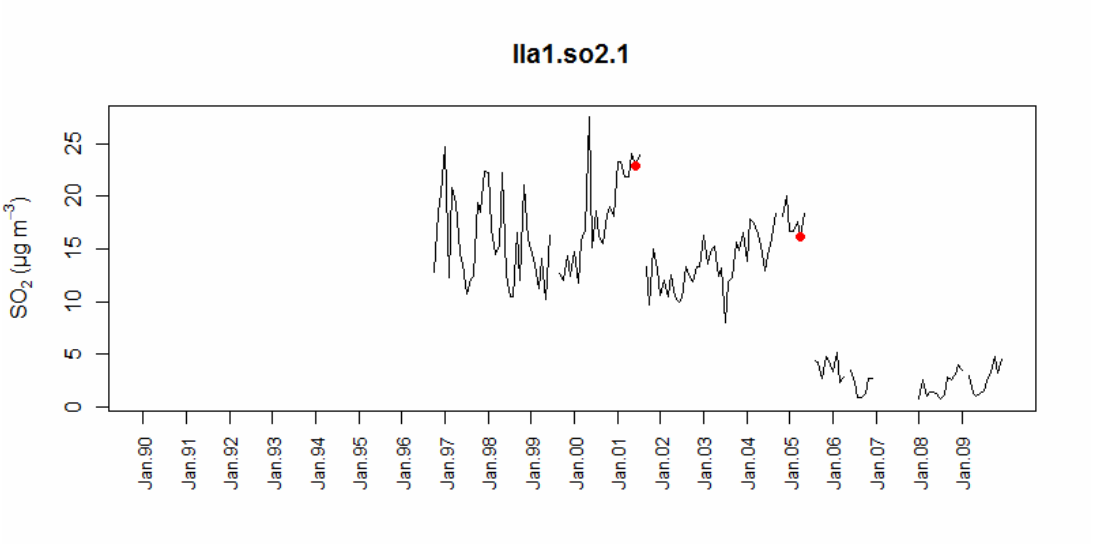
Plots of time series of the different pollutants and temporal resolution selected from AirBase used in the tests: monthly data of CO, NO_2, O_3, SO_2 (1990/97-2009) for Europe, and hourly data of PM_{10} (2007) for Germany.

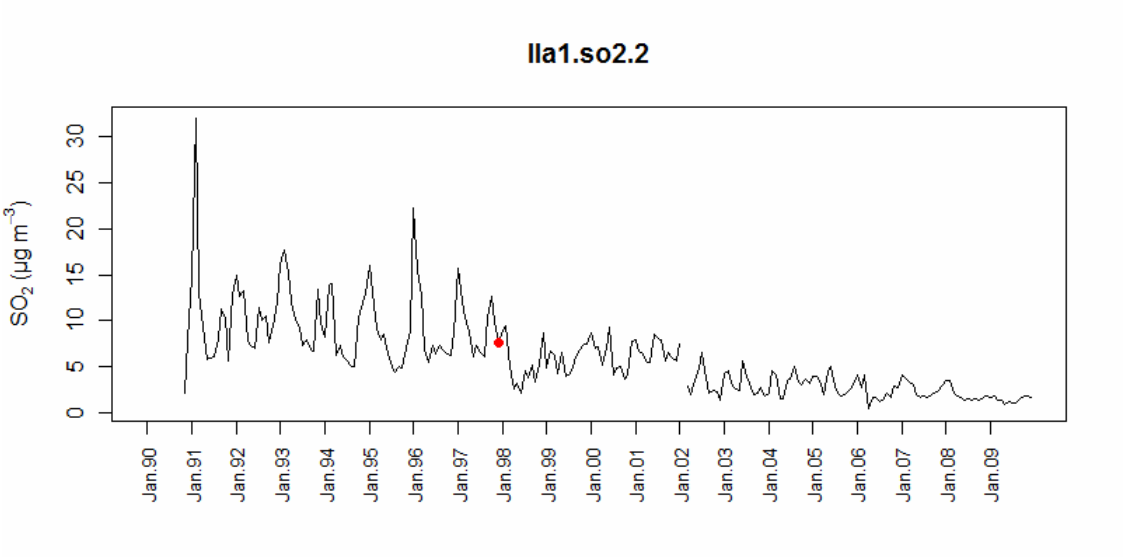
The plots represent the time series with inhomogeneities of type I, IIa1, IIa2, IIb and clean time series as reference, depicted as type III. For each time series the locations of the inhomogeneities was recorded: the red dots in the plots.

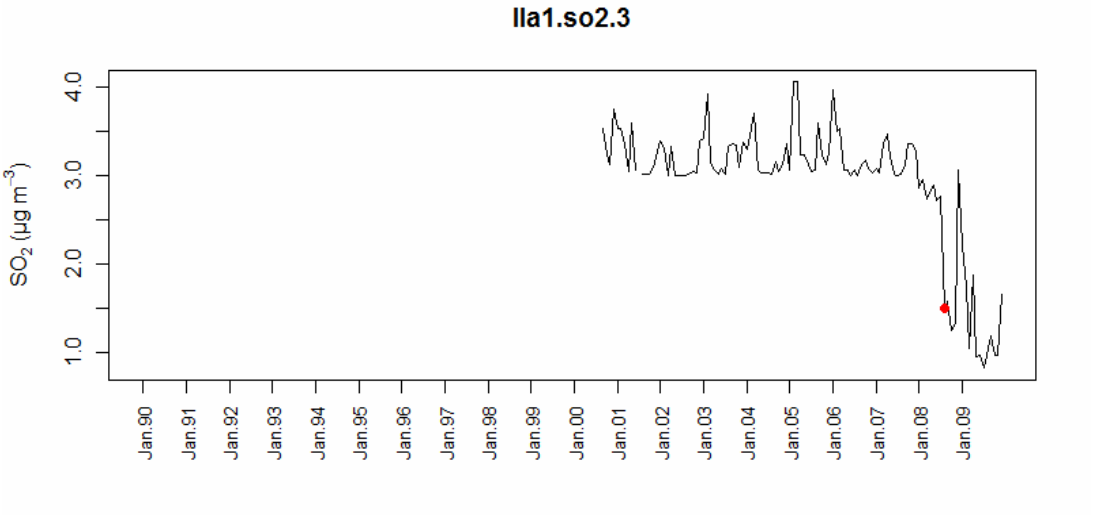
Monthly average data

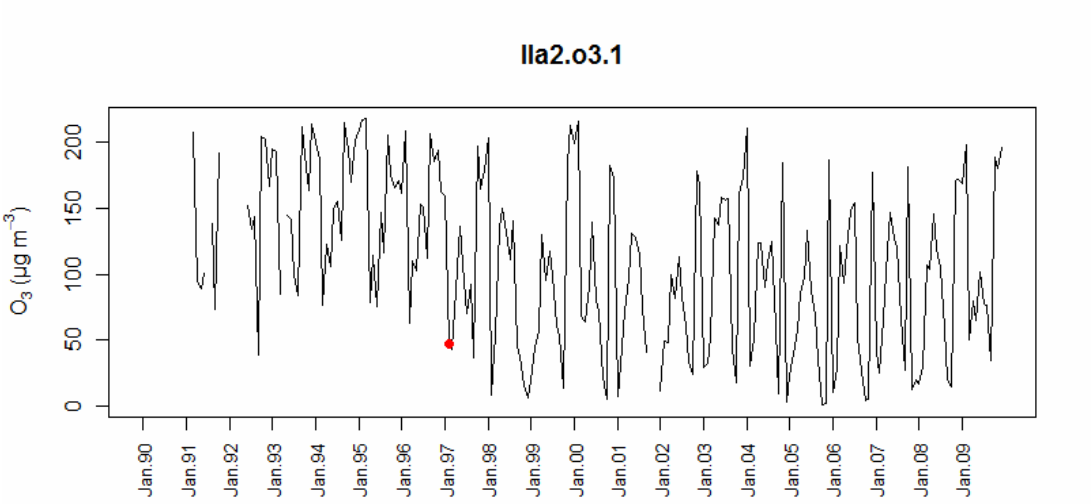


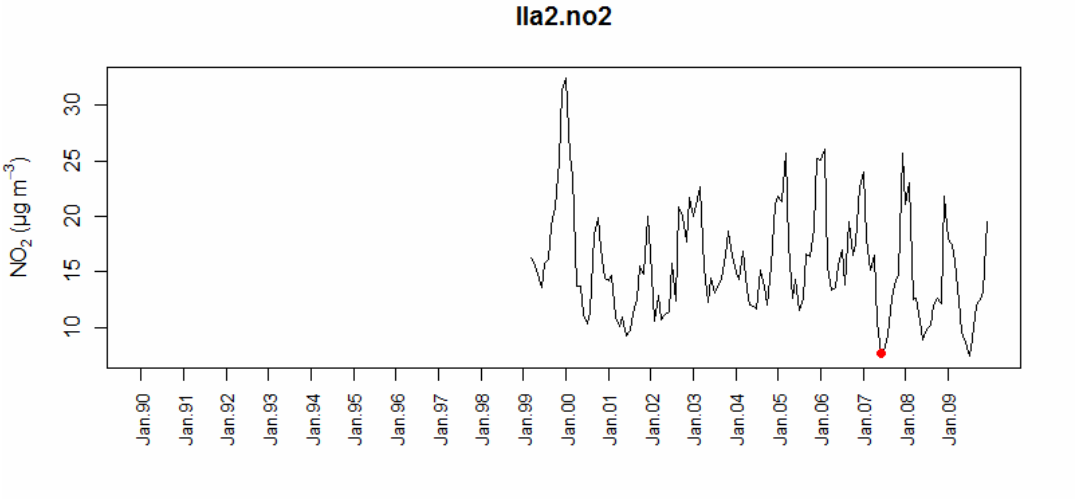


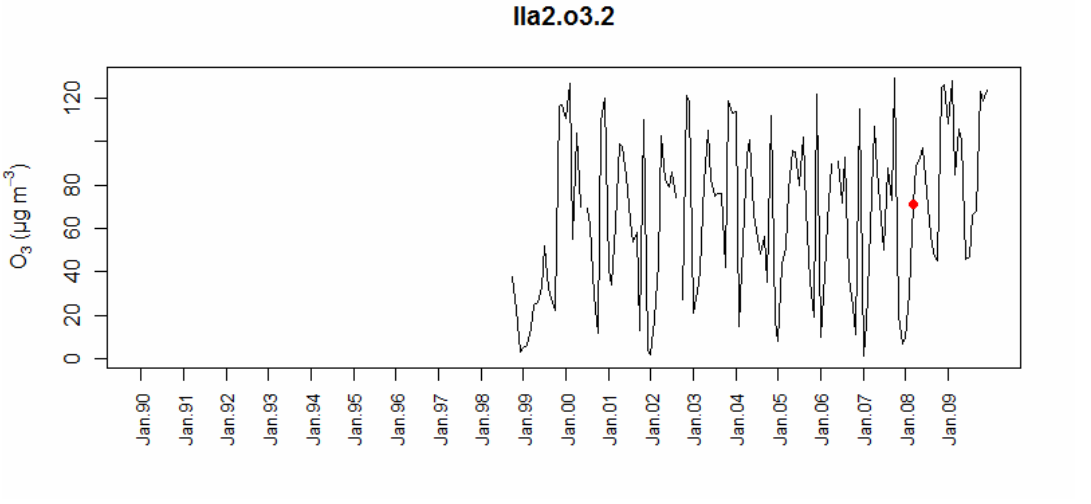

IIa1.so2.1
SO2 (µg m-3)
0
5
10
15
20
25
Jan.90
Jan.91
Jan.92
Jan.93
Jan.94
Jan.95
Jan.96
Jan.97
Jan.98
Jan.99
Jan.00
Jan.01
Jan.02
Jan.03
Jan.04
Jan.05
Jan.06
Jan.07
Jan.08
Jan.09

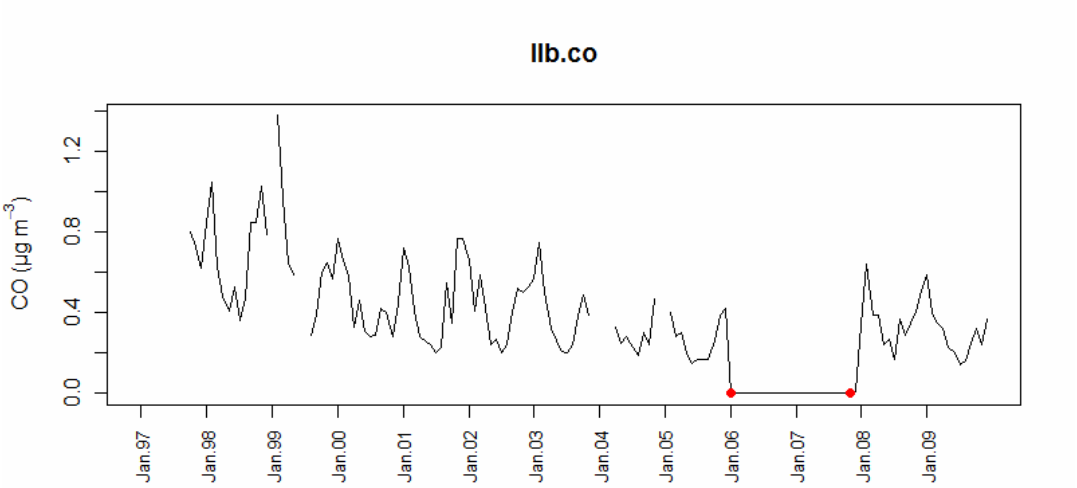

IIa1.so2.2
SO2 (µg m-3)
0
5
10
15
20
25
30
Jan.90
Jan.91
Jan.92
Jan.93
Jan.94
Jan.95
Jan.96
Jan.97
Jan.98
Jan.99
Jan.00
Jan.01
Jan.02
Jan.03
Jan.04
Jan.05
Jan.06
Jan.07
Jan.08
Jan.09

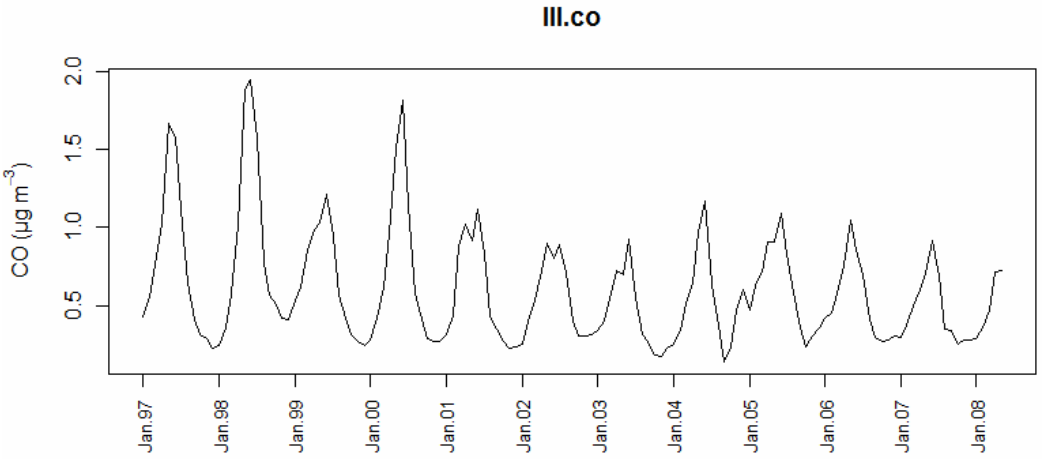

IIa1.so2.3
SO2 (µg m-3)
1.0
2.0
3.0
4.0
Jan.90
Jan.91
Jan.92
Jan.93
Jan.94
Jan.95
Jan.96
Jan.97
Jan.98
Jan.99
Jan.00
Jan.01
Jan.02
Jan.03
Jan.04
Jan.05
Jan.06
Jan.07
Jan.08
Jan.09

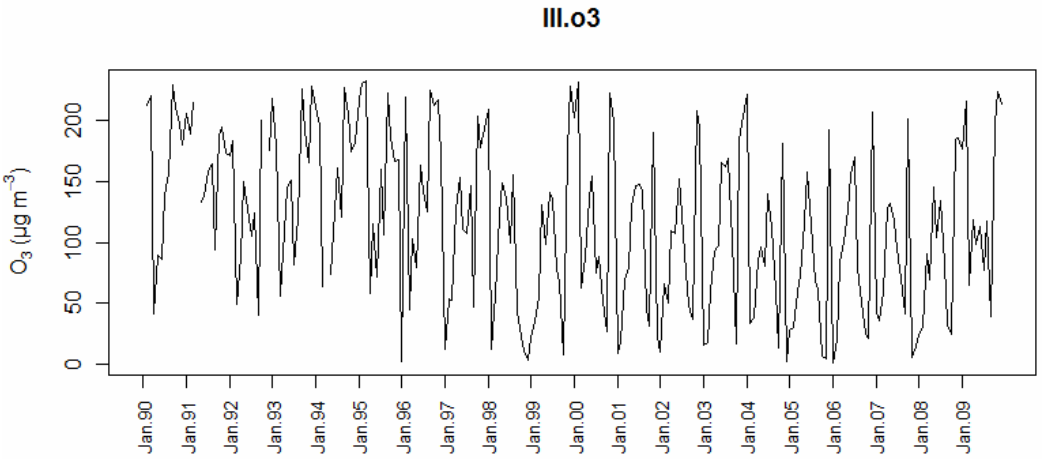

IIa2.o3.1
O3 (µg m-3)
0
50
100
150
200
Jan.90
Jan.91
Jan.92
Jan.93
Jan.94
Jan.95
Jan.96
Jan.97
Jan.98
Jan.99
Jan.00
Jan.01
Jan.02
Jan.03
Jan.04
Jan.05
Jan.06
Jan.07
Jan.08
Jan.09

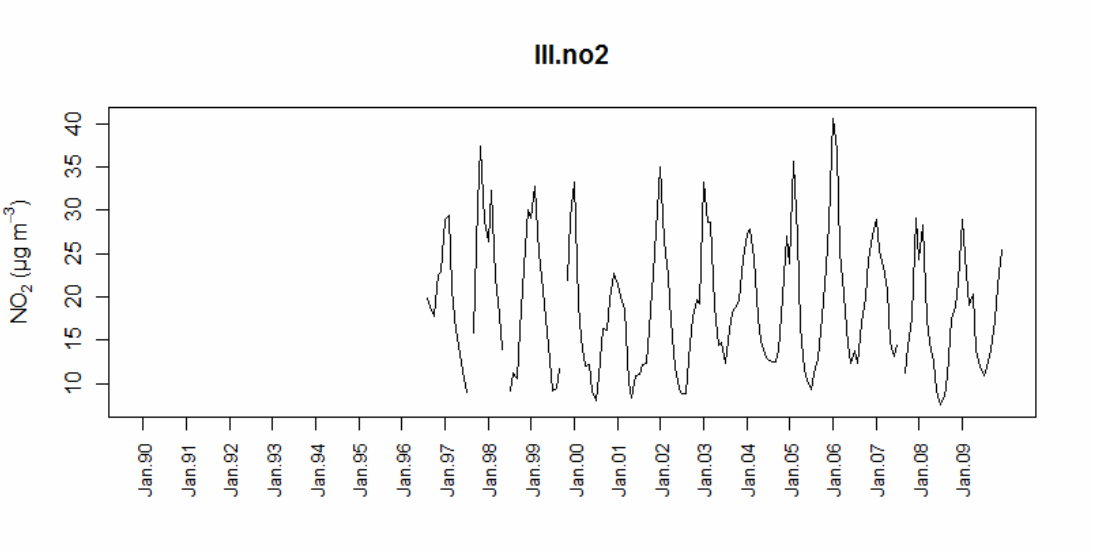

IIa2.no2
NO2 (µg m-3)
10
15
20
25
30
Jan.90
Jan.91
Jan.92
Jan.93
Jan.94
Jan.95
Jan.96
Jan.97
Jan.98
Jan.99
Jan.00
Jan.01
Jan.02
Jan.03
Jan.04
Jan.05
Jan.06
Jan.07
Jan.08
Jan.09


IIa2.o3.2
O3 (µg m-3)
0
20
40
60
80
120
Jan.90
Jan.91
Jan.92
Jan.93
Jan.94
Jan.95
Jan.96
Jan.97
Jan.98
Jan.99
Jan.00
Jan.01
Jan.02
Jan.03
Jan.04
Jan.05
Jan.06
Jan.07
Jan.08
Jan.09

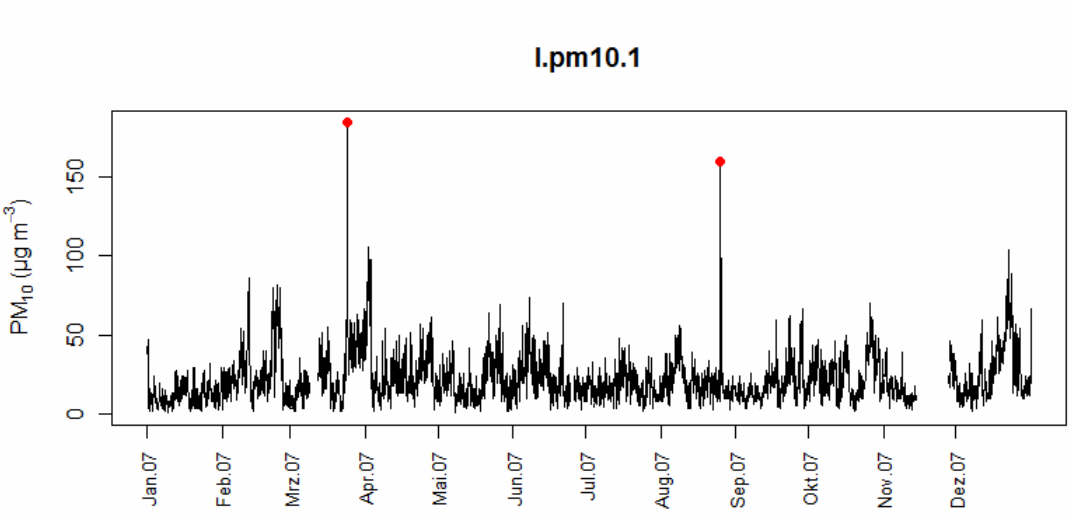

IIb.co
CO (µg m-3)
0.0
0.4
0.8
1.2
Jan.97
Jan.98
Jan.99
Jan.00
Jan.01
Jan.02
Jan.03
Jan.04
Jan.05
Jan.06
Jan.07
Jan.08
Jan.09

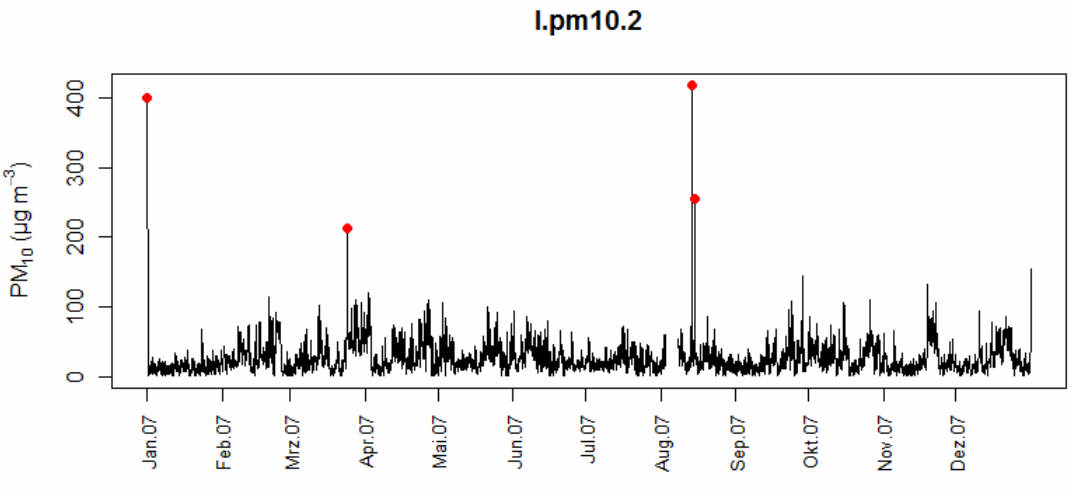

III.co
CO (µg m-3)
0.5
1.0
1.5
2.0
Jan.97
Jan.98
Jan.99
Jan.00
Jan.01
Jan.02
Jan.03
Jan.04
Jan.05
Jan.06
Jan.07
Jan.08

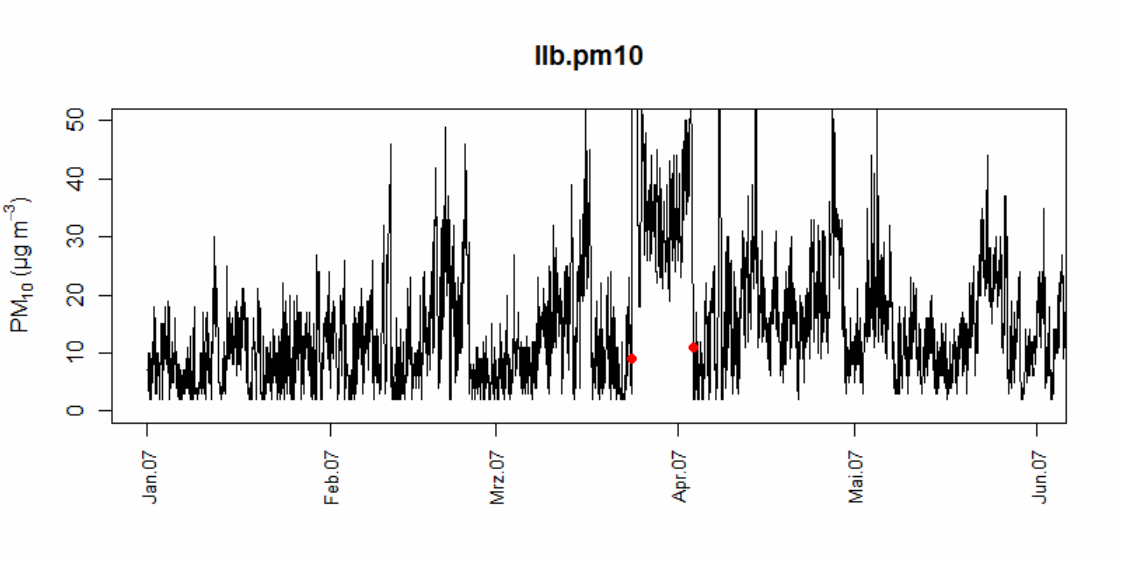

III.o3
O3 (µg m-3)
0
50
100
150
200
Jan.90
Jan.91
Jan.92
Jan.93
Jan.94
Jan.95
Jan.96
Jan.97
Jan.98
Jan.99
Jan.00
Jan.01
Jan.02
Jan.03
Jan.04
Jan.05
Jan.06
Jan.07
Jan.08
Jan.09

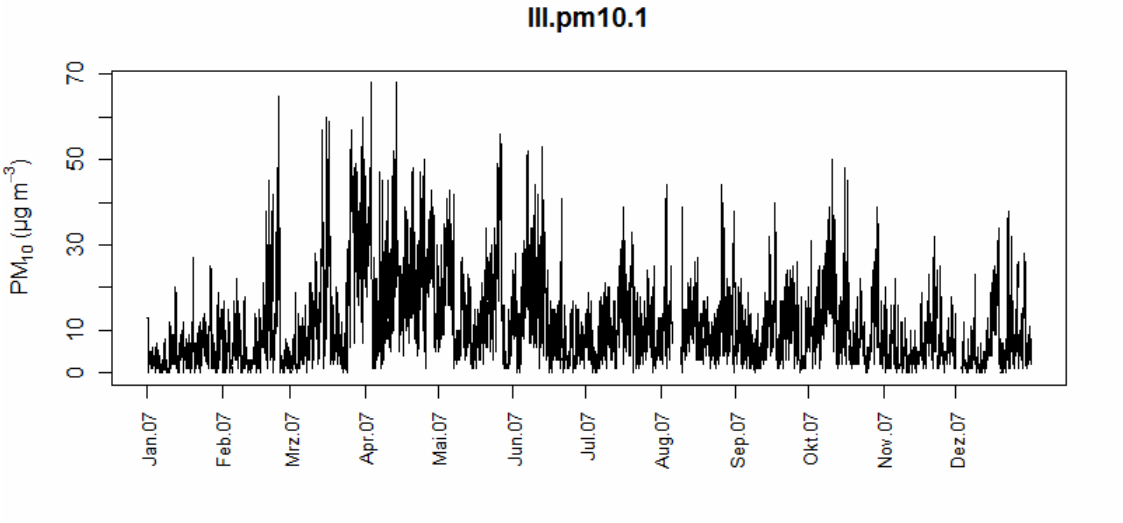

III.no2
NO2 (µg m-3)
10
15
20
25
30
35
40
Jan.90
Jan.91
Jan.92
Jan.93
Jan.94
Jan.95
Jan.96
Jan.97
Jan.98
Jan.99
Jan.00
Jan.01
Jan.02
Jan.03
Jan.04
Jan.05
Jan.06
Jan.07
Jan.08
Jan.09

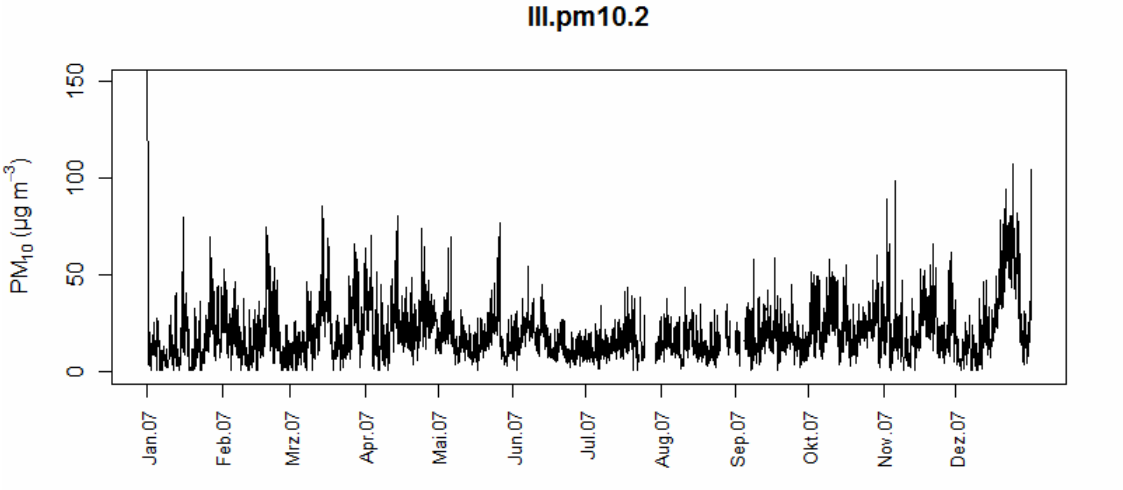
Hourly data


I.pm10.1
PM10 (µg m-3)
0
50
100
150
Jan.07
Feb.07
Mrz.07
Apr.07
Mai.07
Jun.07
Jul.07
Aug.07
Sep.07
Okt.07
Nov.07
Dez.07


I.pm10.2
PM10 (µg m-3)
0
100
200
300
400
Jan.07
Feb.07
Mrz.07
Apr.07
Mai.07
Jun.07
Jul.07
Aug.07
Sep.07
Okt.07
Nov.07
Dez.07


IIb.pm10
PM10 (µg m-3)
0
10
20
30
40
50
Jan.07
Feb.07
Mrz.07
Apr.07
Mai.07
Jun.07


III.pm10.1
PM10 (µg m-3)
0
10
30
50
70
Jan.07
Feb.07
Mrz.07
Apr.07
Mai.07
Jun.07
Jul.07
Aug.07
Sep.07
Okt.07
Nov.07
Dez.07


III.pm10.2
PM10 (µg m-3)
0
50
100
150
Jan.07
Feb.07
Mrz.07
Apr.07
Mai.07
Jun.07
Jul.07
Aug.07
Sep.07
Okt.07
Nov.07
Dez.07

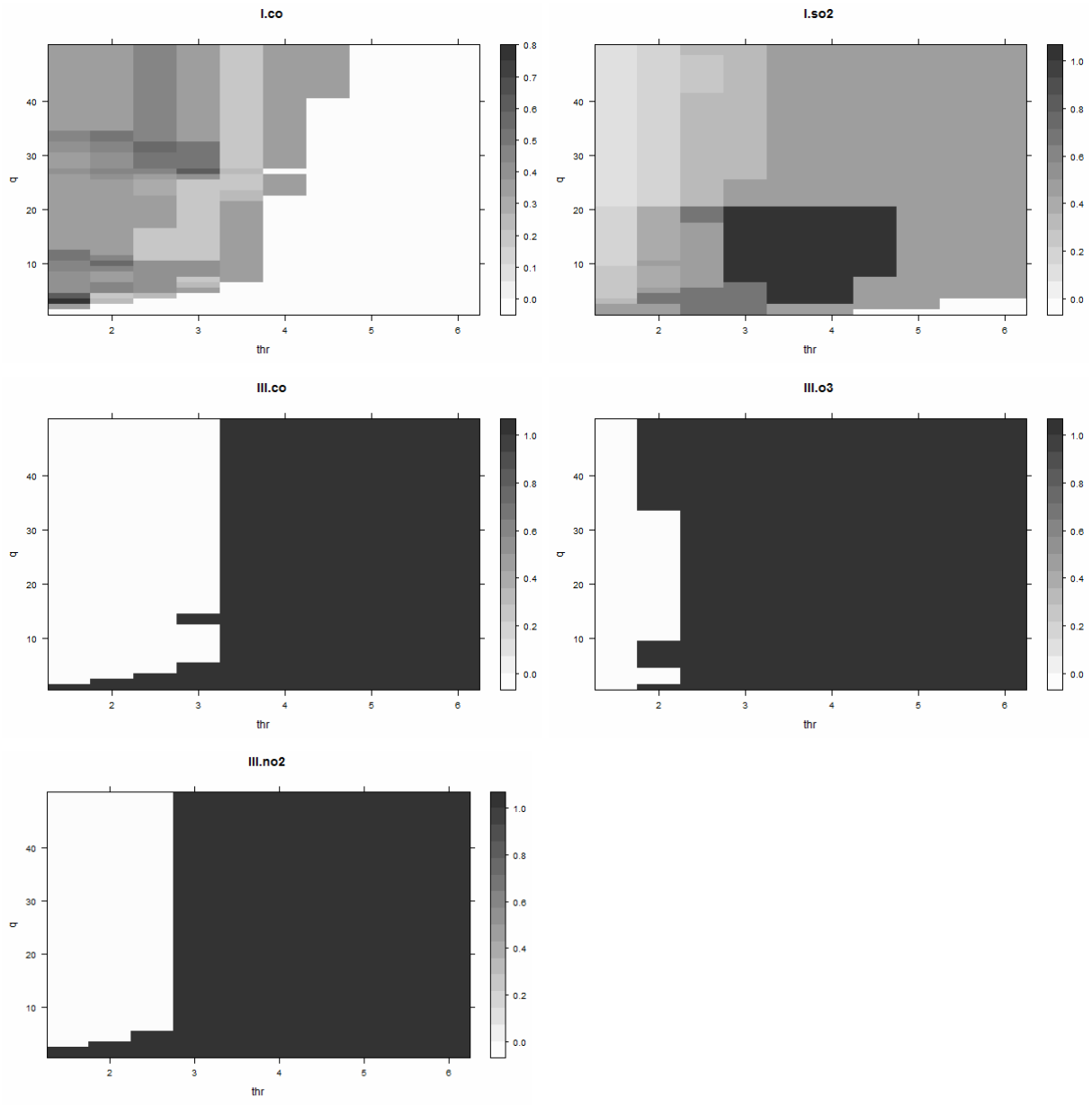
A.2 Parameter optimisation plots per time series

For well performing methods level plots of φ per window size and threshold value are shown for the combined analysis. In these plots the threshold values (thr) are given on the x-axis and the window sizes (q) are given on the y-axis. For each combination of threshold and window size the resulting value for φ is given as grey shade. Areas with high values for φ are depicted darker than areas with low φ indicating bad performance.

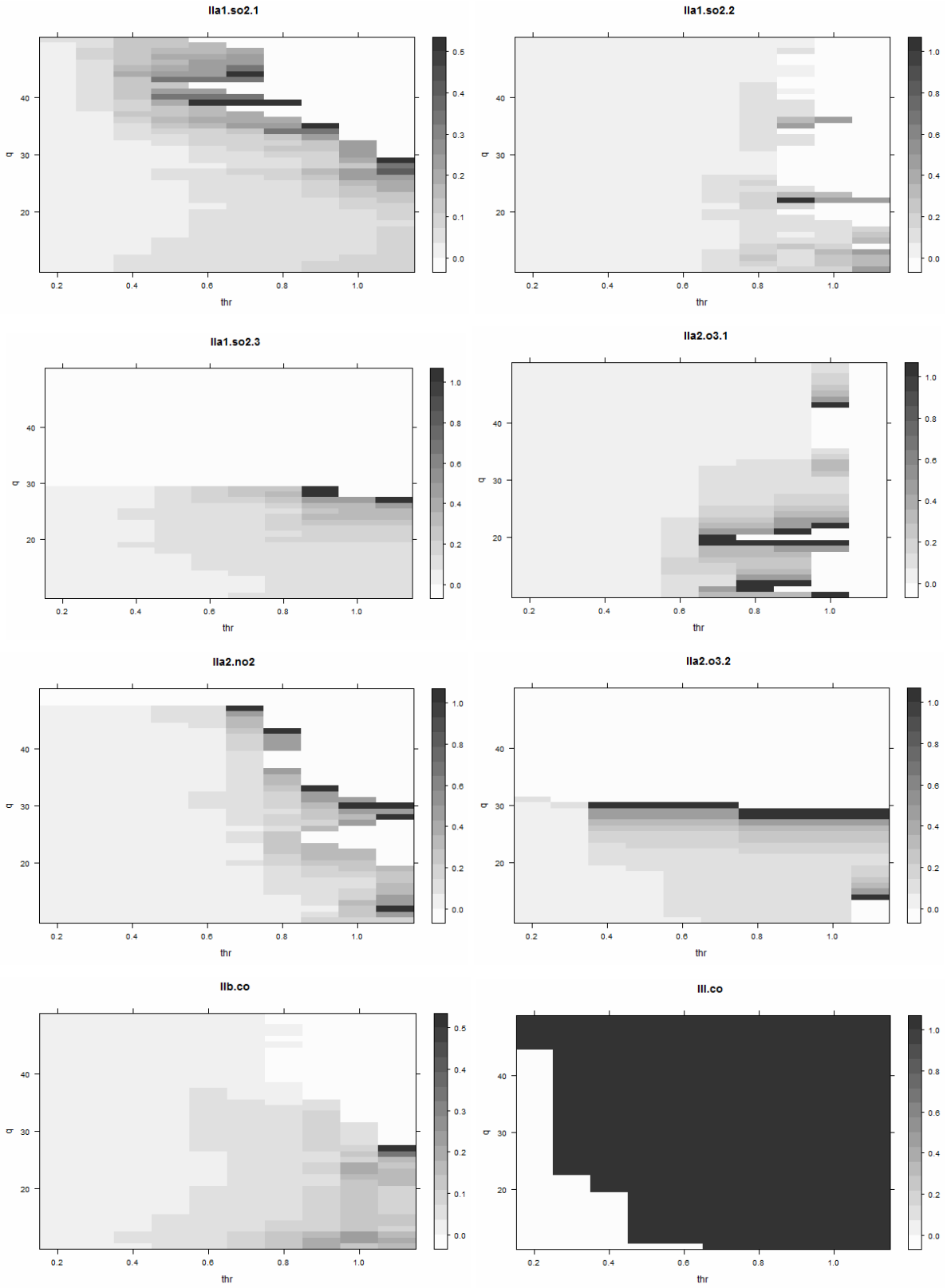
Plots for AR2 and LAG1 are omitted because of the poor results over the whole parameter range.

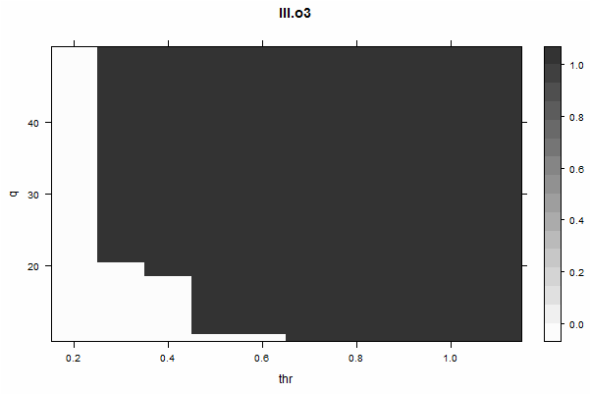
Monthly data

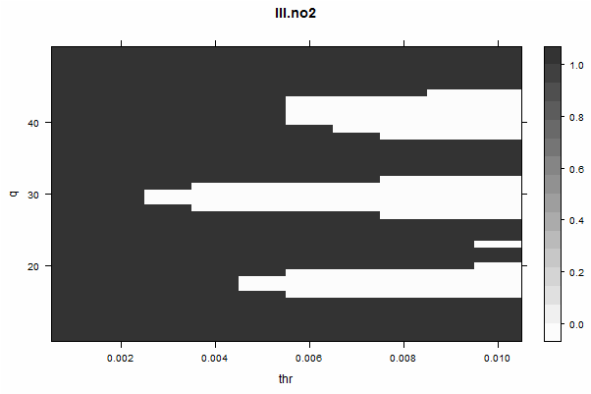
Moving Window – Whole window (MW)



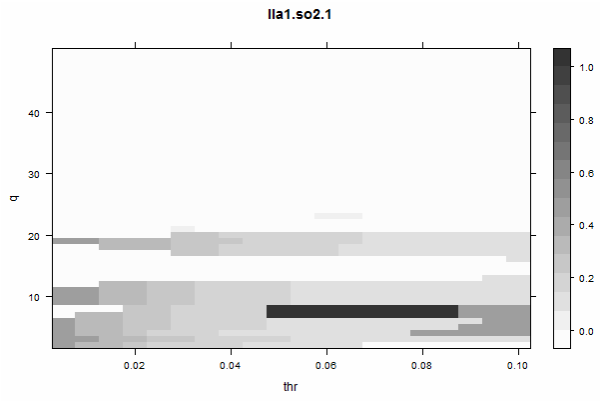
Moving Window – Two-sided window (MW2)

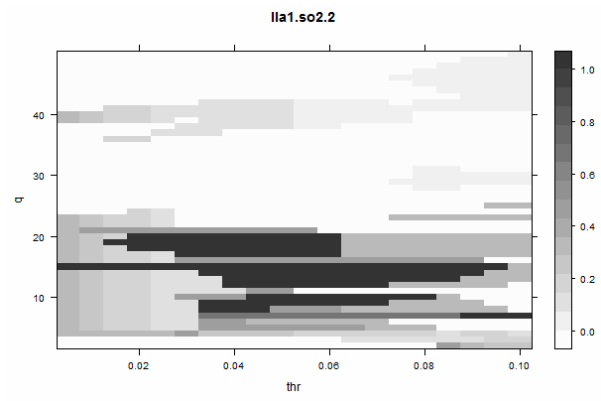


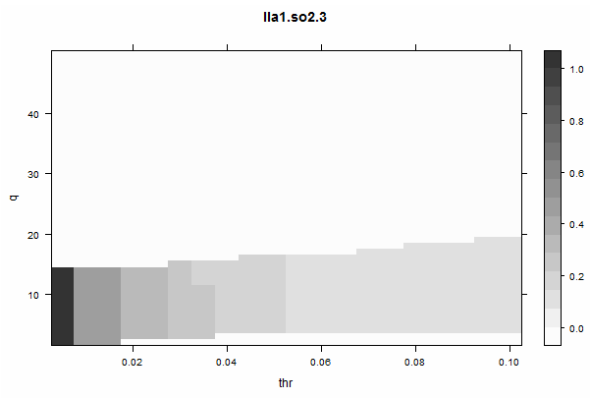


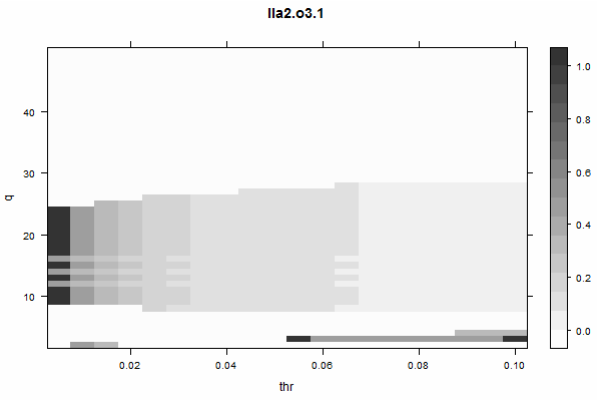


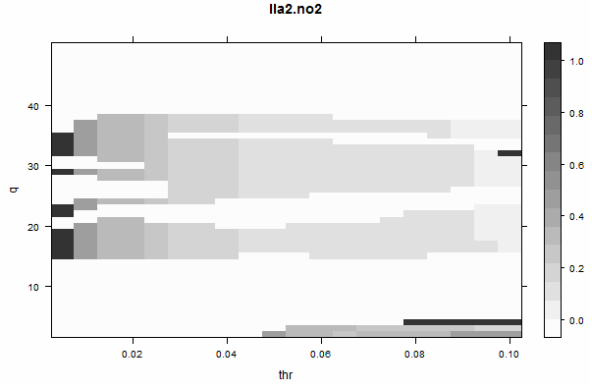
Moving Average filter (MA filter)

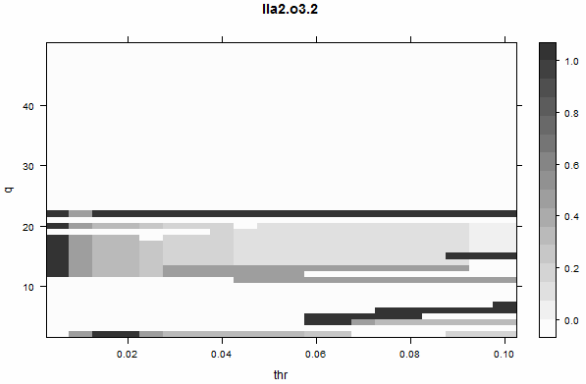


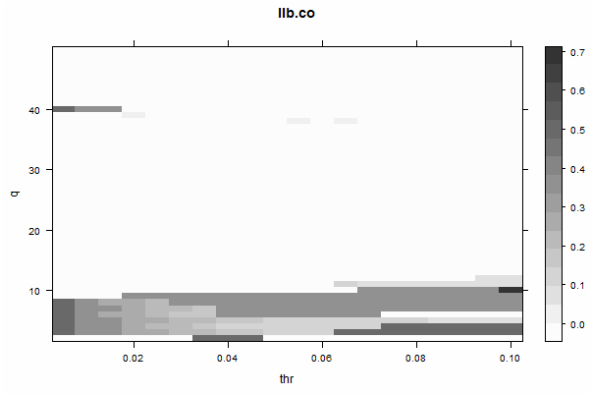


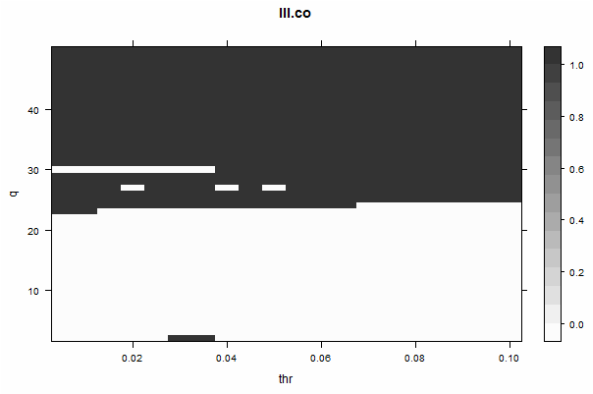


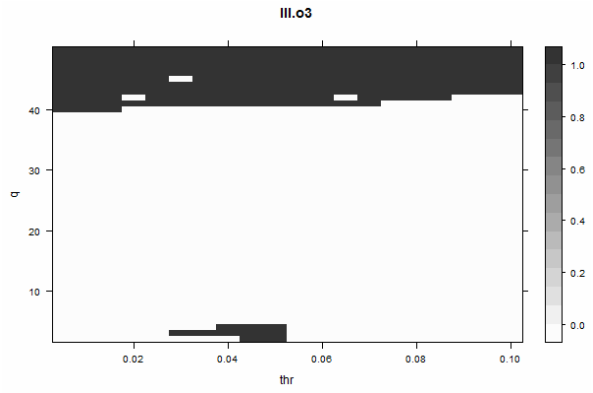


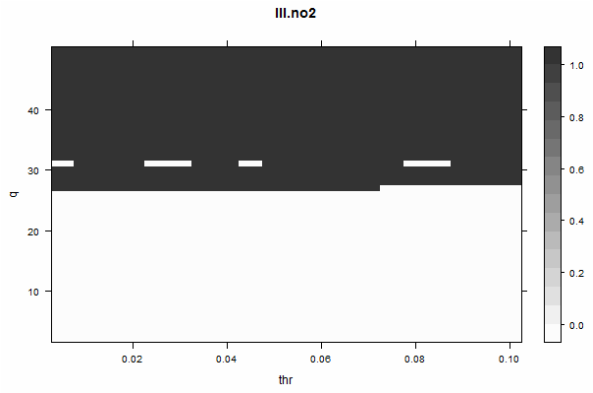






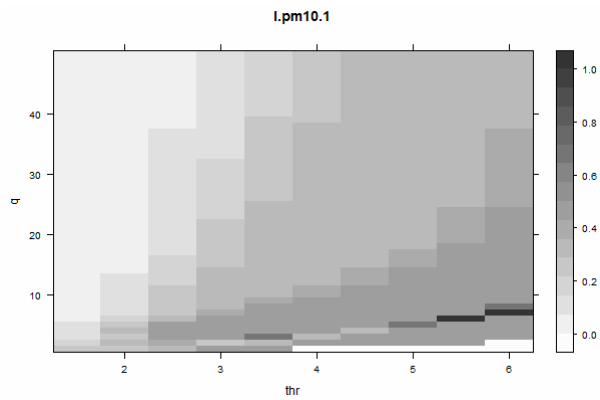


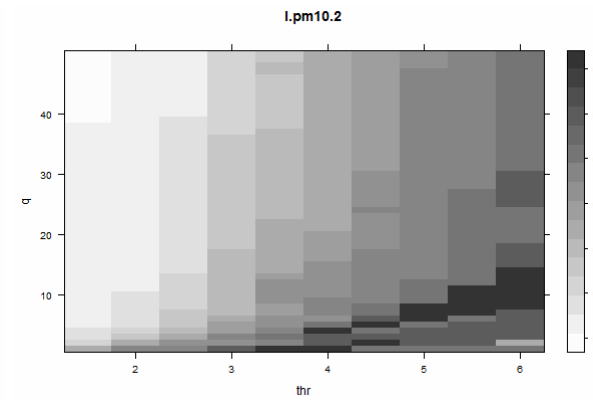


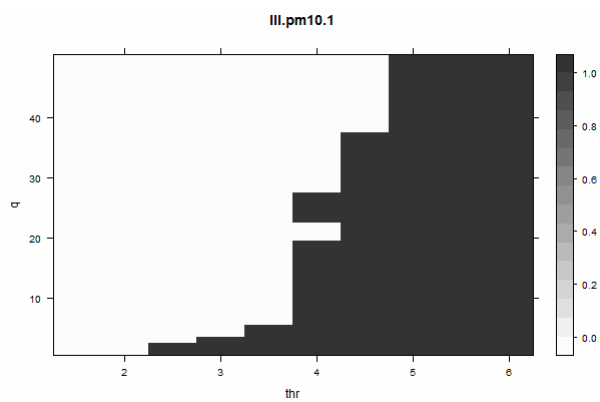


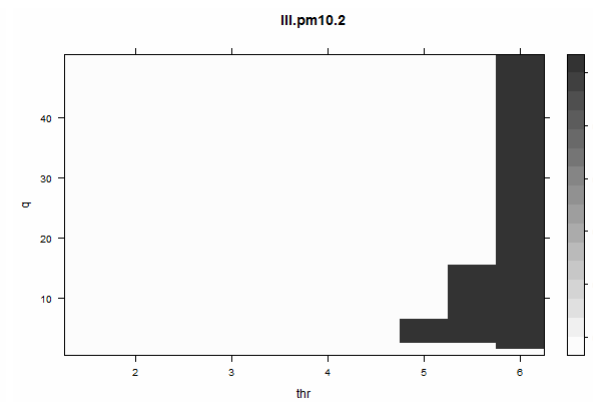
Hourly data

Moving Window – Whole window (MW)

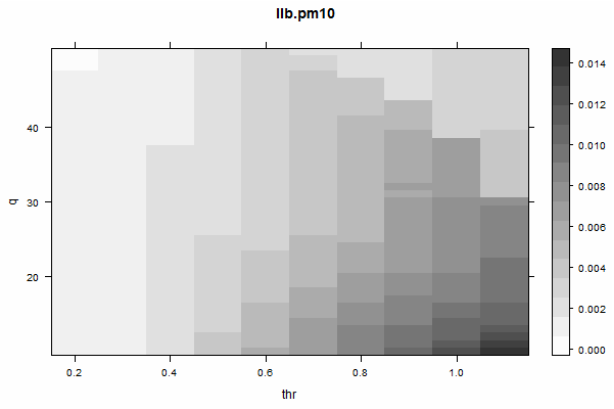








Moving Window – Two-sided window (MW2)



Moving Average filter (MA filter)