

1 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.

1.1 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. After ODBC drivers have been specified, loading data subsets of interest from DEM (.mdb files) can be achieved by using package RODBC [?]. For reading data from AirBase (.csv files) no extra package is required.

1.2 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 1.1 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 1.1 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.

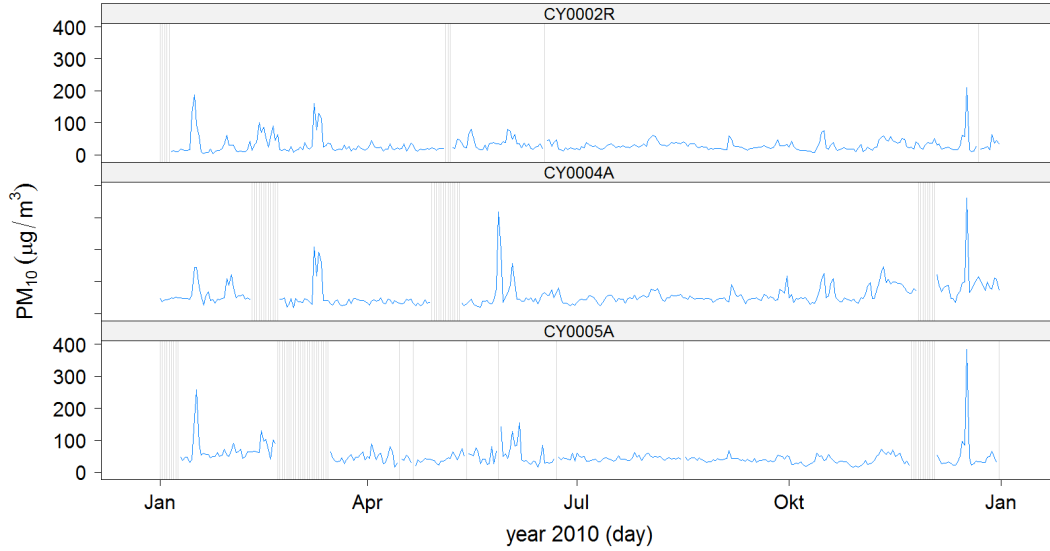


Figure 1.1: Cyprus PM_{10} stations, missing values in grey.

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 1.2), 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 1.2 a boxplot is given for every measurement station considered, with stations ordered by median. Colour coding is used to indicate the country.

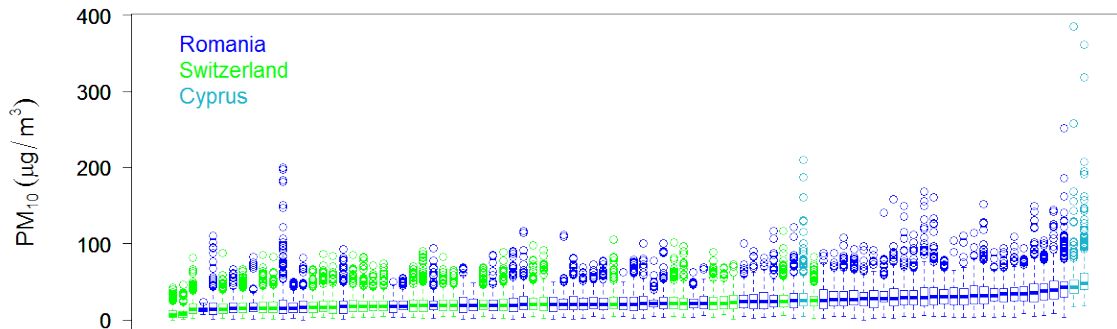


Figure 1.2: Boxplots per station, ordered by median.

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 1.3](#) shows the time series plot where thresholds of potential outliers as identified by the boxplots are indicated by orange lines.

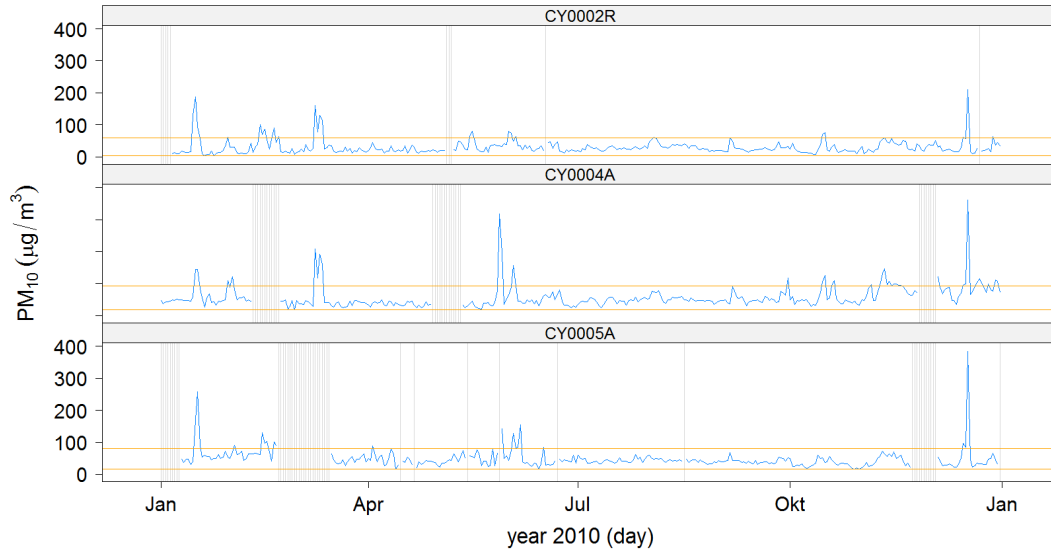


Figure 1.3: Cyprus PM_{10} stations, missing values in grey, boxplot outlier thresholds in orange.

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 1.4](#) 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.

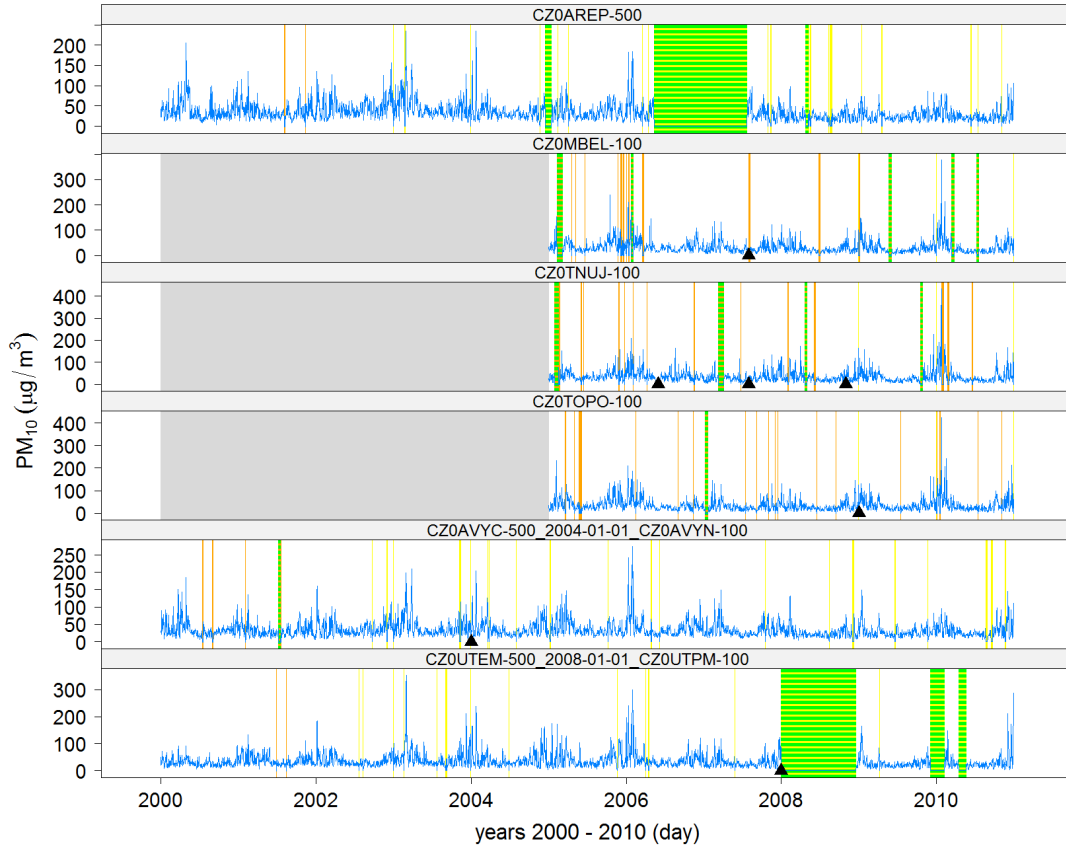


Figure 1.4: Selection of PM_{10} stations from Czech Republic. Missing values are shown in grey, zero values in yellow, negative values in orange, and runs of equal values of length 10 or longer in green. Black triangles mark hypothesized break points.

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 1.4). 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.

1.3 Transformations

Some heuristics in ?? 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 1.5 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.

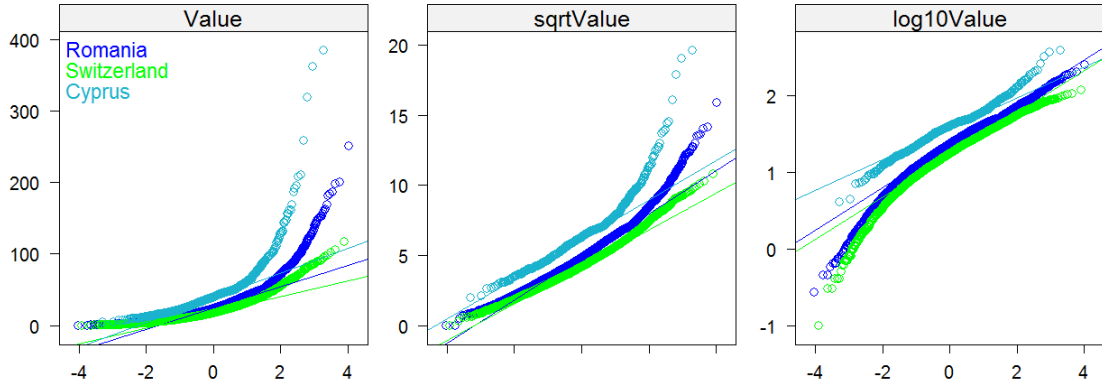


Figure 1.5: Normal-Quantile plots of original and transformed variables for PM_{10} data.

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.

1.4 Time series decomposition

Structural changes as searched for in ?? 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 ?] 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 1.6 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.

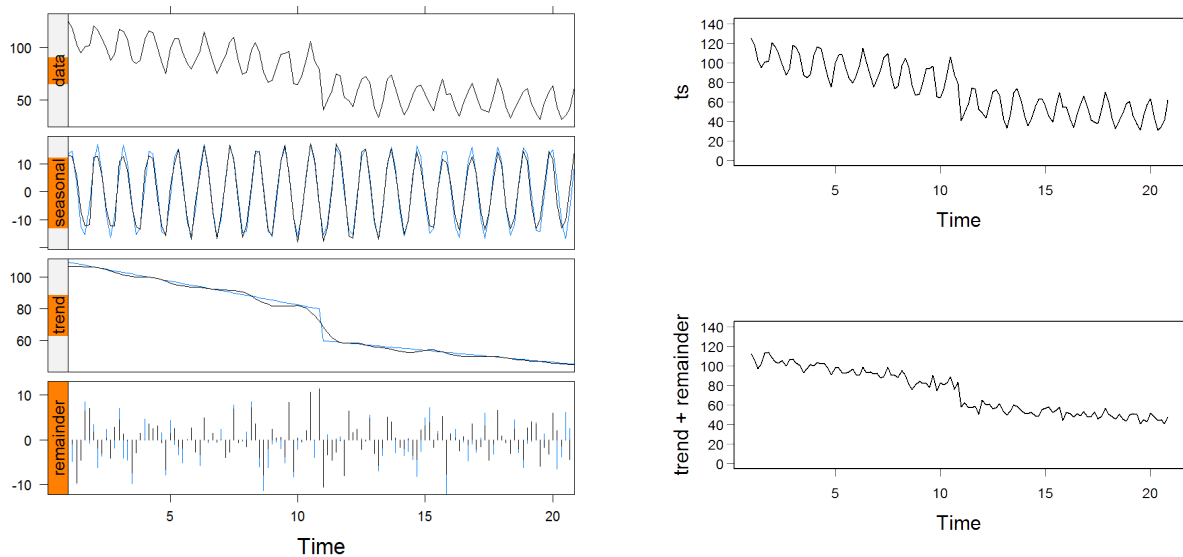


Figure 1.6: Left: Decomposition of simulated time series: true components in blue, STL decomposition in black. Right: Comparison of simulated time series before and after decomposition and removal of seasonal effects.