DI Markus Zahrnhofer

Further development of prediction models for PM10 in Graz

MASTER'S THESIS

to achieve the university degree of

Diplom-Ingenieur

Master's degree programme: Technical Mathematics, Operations Research and Statistics

submitted to

Graz University of Technology

Supervisor

Univ.-Prof.i.R. Dipl.-Ing. Dr.techn. Ernst Stadlober

Institut für Statistik

Graz, April 2018

AFFIDAVIT

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly indicated all material which has been quoted either literally or by content from the sources used. The text document uploaded to TUGRAZonline is identical to the present master‘s thesis.

Date

Signature

Abstract

The scope of this thesis is a further development of the prediction model for 24h average values of particulate matter PM10 in Graz. A first step is the extension of the current one-day ahead prediction model to a two-days ahead model using linear regression models. This model should be in line with the current model and should support the Styrian provincial government for actions which can be planned in advanced. In the following we apply *Classification and Regression Trees (CART)* to the data-set of Graz-Mitte and Graz-Süd. Moreover, we apply the method of bagging and random forests which combine the results of many tree models to one final output. Finally neural networks which are a very popular method of machine learning were applied to the data. Especially the CART models could help to give a fast and suitable indication of the exceedance of the critical threshold of 50 $\mu g/m^3$.

Zusammenfassung

Diese Arbeit befasst sich mit der Weiterentwicklung der statistischen Modelle für die Vorhersage der 24h Mittelwerte der PM10 Konzentration in Graz. Das aktuell verwendete Regressionsmodell wird in einem ersten Schritt erweitert, um längerfriste Planungen und Maßnahmen zur Reduktion der Feinstaubkonzentration zu ermöglichen. Darüber hinaus werden *Classification and Regression Trees (CART)* auf die Datensätze der Messstationen Graz-Mitte und Graz-Süd angewendet. Diese ermöglichen eine schnelle erste Indikation ob am darauffolgenden Tag eine Überschreitung der kritischen Schwellen von 50 $\mu g/m^3$ beobachtet wird. Zusätzlich werden noch Random Forests und Neuronale Netze auf die Daten angewendet. Die Machine Learning Algorithmen sind sehr mächtig, bringen jedoch im Vergleich zum derzeitigen Modell nur eine geringe Verbesserung.

Acknowledgment

I would like to thank my supervisor, Prof. Dr. Stadlober for giving me the possibility of writing the thesis about this very interesting and complex topic. Further I thank him for his faithful supervision and his patience. Special thanks go to all people, who helped me with the work on the thesis.

To Cornelia and Samuel Leonard

All errors are my own.

Graz, April 2018 Markus Zahrnhofer

Contents

Chapter 1

Introduction

The scope of this thesis is a further development of the prediction model for particulate matter PM10 in Graz. Air pollution in Graz attracts remarkable attention in the last years, because during the winter season there is a permanent exceedance of the threshold value for PM10. A detailed descriptive analysis, the origin of PM10 and the adverse influence on health can be found in Hörmann et al. [11]. The current prediction model has been introduced by Hörmann et al. [11] in the year 2004 and has been yearly updated and reviewed. This work includes a detailed analysis of the data and a one day prediction of particulate matter PM10 (24h average value) for different locations in Graz.
Based on the existing analysis we will implement possible, further developments of the model within this thesis. The further developments should be used for prediction and not only for statistical description of possible dependencies. The current model is a multiple regression model for a one day ahead prediction of particulate matter PM10. Parameters for predicting PM10 are meteorological values (like temperature and wind) and lagged particulate matters values, as well as categorical parameters like weekdays and weekends.
The goal of the first model is to expand the current model to a 2-day prediction model. This enlargement of the model might help planning long-term actions in case of exceeding the limit of $50\mu g/m^3$. To be in line with the current model, this model is based on a multiple regression with additional parameters.
In the second model we will introduce tree based models for predicting one day ahead PM10 concentration or the exceedance of the critical threshold $50\mu g/m^3$. This models can be seen as supplementary models for the current prediction models described in Stadlober et al. [23] and [24]. The big advantage of the tree models are their simple overseeing structure and representation. This models can help to get in a simple and very fast way a first indication if there will be a breach of the critical limit on the next day or not.
Based on the tree models we use the bagging algorithm and random forest models to get models with a good predictive power, which smooth out the weakness of single tree models out. The advantage of these models is that in each run an implicit test of the forecasting error is performed by the out-of-bag data. This helps to stabilize the results of the tree model.
In the final chapter we apply neural networks to PM10 data of Graz-Mitte and Graz-Süd.

Neural networks are widely used for predicting PM10 concentrations (compare Perez [19]). The disadvantage of the last 3 model types is the black box characteristic. The formal representation of the relationship (e.g. regression model) of the response and predictor variables is not anymore available. But their predictive power is undisputed.

Chapter 2

2-Day Prediction Model of PM 10

This chapter is about the first possible extension of the current model described in Stadlober et al. [23]. There is a continuous further development of the model described in Stadlober et al. [23] and the current model for the prediction differs slightly. In general the model uses similar input variables but they can sightly differ after the seasonal recalibration of the model. The current model is based on a multiple regression predicting the 24-hour average PM10 value of the following day $(t+1)$. The prediction is based on the available information at midday at day t. This condition is also essential for the extension to a 2-day prediction model. The prediction 2 days ahead will be calculated at the same available information as the current model. The extended model should be in line with the current model. Therefore the modeling and calibration of models in the following sections is done for the winter season (November-March) only. The idea of the extension to a 2-day ahead prediction model is shown in Figure 2.1. The forecast of the 2 day prediction model is based on the similar information available at midday at time t.

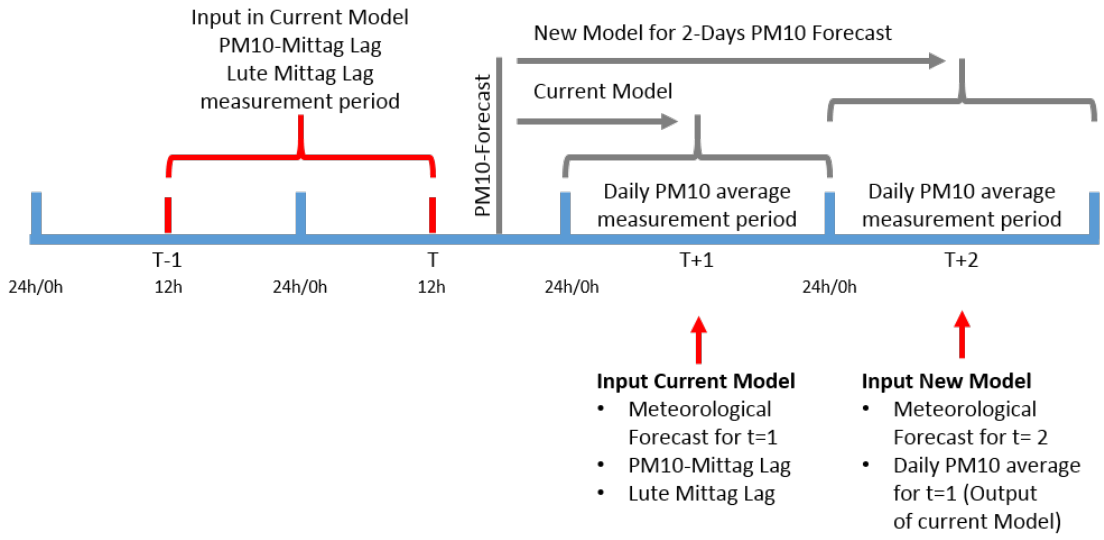

Figure 2.1: Extension to a 2-day prediction model

2.1 Graz-Mitte

The first model is developed for the site Graz-Mitte. This site represents the PM10 concentration for the center of Graz. For the model calibrating the available data sets for the winter seasons (November 1^{st}-March 31^{st}) from 2003/2004-2013/2014 are used. The seasons 2014/2015 and 2015/2016 are used for back-testing of the calibrated model to measure the goodness of the prediction.

2.1.1 Data

The data set for Graz-Mitte contains 4841 observations (including summer season), starting at January 1^{st} 2003. In addition to particulate matter variables, the set contains meteorological and date information for each observation. Hörmann et al. [10] mention that a calibration of a reliable prediction model for PM10 without meteorological data is not possible. A detailed analysis of the correlation of weather data and particulate matter concentration can be found in Stadlober et al. [23]. Within this thesis the main focus lies on the extension of the model and not on the descriptive statistical data analysis.
The data set includes 2 different PM10 variables. The first variable indicates the daily average PM10 concentration from 0:30h to 24h. This size is the average of half hour values observed at a certain day. The daily average is the essential size for possible actions in case of exceeding the limit value of 50 μ/m^3 and is the response variable in the following models. The second PM10 variable in the data set represents the average of the PM10 concentration from 12:30h to 12h (**PM_mittag**). This variable is essential for the current prediction model. The prediction of the average daily PM10 concentration (0:30h-24h) at $t + 1$ is based on the **PM_mittag** at t see Stadlober [23]. The prediction will be done in the afternoon of the previous day and uses **PM_mittag_lag1** instead of the the lagged value of PM10. This improves the accuracy of the model and reduces the lack of information from 24 hours to 12 hours.
In addition to the PM10 variables and other air pollutants like NOx, which are not considered for the model, the data set includes meteorological variables like average temperature, wind speed, precipitation and humidity. Some of these variables are transformed to dummy variables with values 0 and 1. In the following we list the number of missing PM10 observations for the winter seasons.

Winter Season	Missing Values
2002/2003	no values before January 1^{st} 2003
2003/2004	March 11^{th}-March 22^{nd} 2004
2004/2005-2008/2009	no missing values
2009/2010	December 11^{th} 2009-February 17^{th} 2010
2013/2014	January 26^{th}-January 27^{th} 2014
2014/2015-2015/2016	no missing values

Table 2.1: Missing PM10 Values Graz-Mitte

Considering the time series of daily PM10 concentration (Figure 2.2) from 2003 on, one can see a slightly decrease of the PM10 concentration over time. This trend is emphasized by considering the mean of the PM10 concentrations of each winter season in Table 2.2. The decrease of the PM10 concentration originates from two effects. There is a obvious correlation between meteorological data and the PM10 concentration. Higher temperatures and fewer frost days might lead to a lower particulate matter concentration. In addition there is a positive correlation between number of days with temperature inversion and the PM10 concentration. The second reason for the reduction of PM10 in the last years is linked to other actions of the Styrian provincial government (like reduction of Speed Limits on highways next to Graz) based on the current forecasting model.

Season	Observations	Average PM10	Average Temp in C°	Inversion Days	Frost Days	Strong Wind
2002/2003	90	71.60	2.10	42	33	63
2003/2004	140	56.85	2.68	48	39	71
2004/2005	151	54.17	2.53	47	48	91
2005/2006	151	61.64	1.60	27	52	80
2006/2007	151	47.00	5.78	47	10	98
2007/2008	152	46.24	3.86	33	30	105
2008/2009	151	39.30	3.35	17	33	103
2009/2010	83	35.99	6.17	22	6	30
2010/2011	151	44.22	3.13	43	51	52
2011/2012	152	39.00	3.69	35	32	59
2012/2013	151	35.70	3.01	31	32	50
2013/2014	149	30.20	5.47	31	15	56
2014/2015	151	30.23	5.26	21	13	64
2015/2016	152	33.29	5.15	54	17	56

Table 2.2: Summary Graz-Mitte

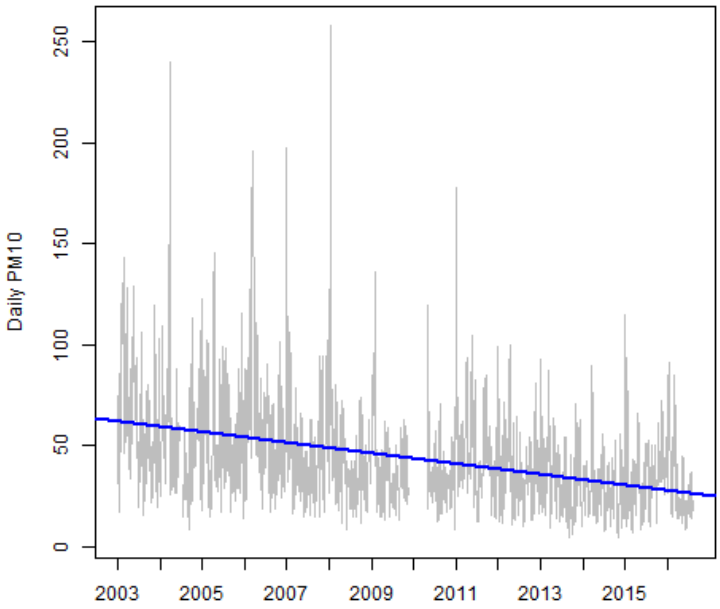

Figure 2.2: Daily PM10 average values for the winter seasons, Graz-Mitte

2.1.2 Model

To get a good and reliable model for the 2-days ahead prediction of PM10 at $t + 2$, we choose all variables which may have positive impact on the calibration of the model. The goal is to find the best model in the sense of described variance adjusted R^2 and the number of explanatory variables. Therefore we choose the following variables.

- **PM10_lag1** This variable is the measurement of PM10 to the previous day $t + 1$. This variable should model the strong autocorrelation effect within the PM10 concentration. The autocorrelation function for 3 seasons can be seen in Figure 2.3. Simply spoken this high autocorrelation means that a day with high PM10 concentration is followed by another day with high PM10 concentration.

- **überschreitung_lag1** This variable is the binary version of **PM10_lag1** with respect to the critical limit of 50 $\mu g/m^3$.

- **lute** The variable represents the average temperature for the time period 0:30h-24h at $t + 2$ and is a weather forecast. The variable is the average of half hour values temperature measurements. As already seen in Table 2.2, there exists a correlation between temperature and PM10 concentration.

- **lute_mittag** This variable gives the average temperature of half hour values from mid-day 12:30 day $t + 1$ till 12:00 at day $t + 2$. In contrast to the variable **lute**, which is the

average temperature at day $t+2$, **lute_mittag_lag1** is the variable, whose measurement period ends already at midday of $t+1$.

- **ltusg_k** The variable representing the average temperature difference of a measurement point $360m$ above (Kalkleiten) the site Graz Mitte (350 meters above sea level) and is based on weather forecasts for $t+2$. Higher temperature at Kalkleiten (710 meters above sea level) during the winter months means temperature inversion. The following variable is the binary version of **ltusg_k**.

- **invers** is 0 if no temperature inversion and 1 if temperature inversion for $t+2$. Temperature inversion has a significant impact on PM10 concentration. see: Hörmann et al. [11] (weather forecast for $t+2$)

- **frost** is a dummy variable, which is 1 if the average temperature is below 0 degree and 1 otherwise for time $t+2$ (weather forecast).

- **wind** is a dummy variable which is 0 if the wind velocity is greater than $0.6m/s$ and 1 otherwise. Strong wind reduces the PM10 concentration significantly.

- **nied01** is a dummy variable which is 0 if no precipitation and 1 otherwise. Precipitation, especially rain reduces the PM10 concentration significantly (weather forecast for $t+2$).

- **suho** is a categorical variable which represents Sundays and public holidays. Volume of traffic is an essential factor for higher PM10 values, therefore a distinction between holidays with lower traffic volume and week days might be reasonable.

- **sat, wd** similar to **suho**, representing saturdays and weekdays.

- **nov, dec, jan, feb, mar** dummy variables of every single month during the winter season.

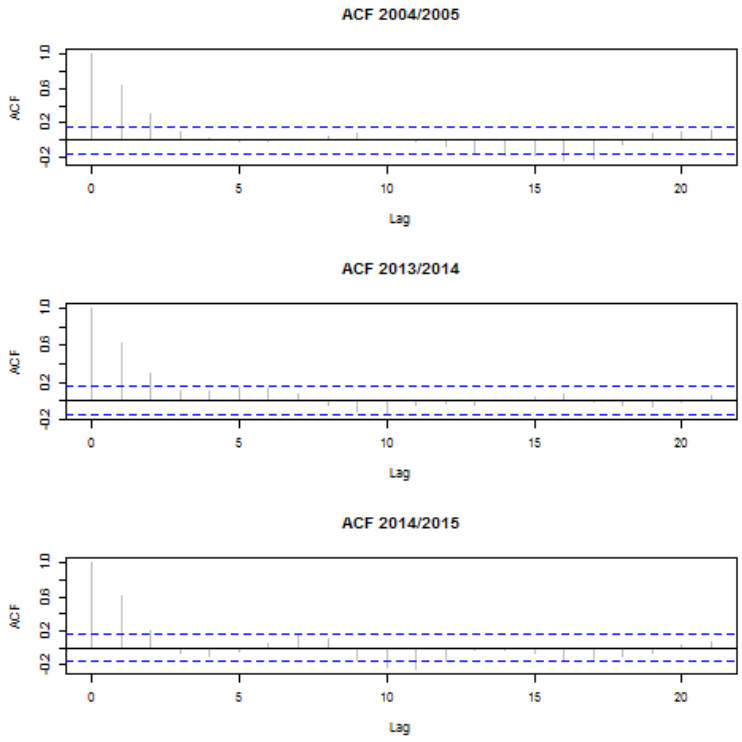

Figure 2.3: Autocorrelation of PM10 Values for Graz-Mitte

To find the best model for predicting the dependent variable $\sqrt{PM10}$, linear models with all possible combinations of the variables described above are constructed. The linear model consists of at least 4 explanatory variables. These calculations results 32192 different linear models. A huge number of them are not very useful, because of too low adjusted R^2 and the combination of the prediction variables. But there are a lot of models which have very similar size of variance described with different number of variables. The model with the highest adjusted R^2 of 0.677 is composed of 12 variables (see: Table 2.3).

Model Parameters	Adjusted R^2
PM10_lag1,lute_mittag, ltusg_k, lute, suho, jan, mar, dec, sat, wind, nied01	0.6772
PM10_lag1,lute_mittag, ltusg_k, lute, feb, mar, dec, sat, wind, nied01	0.6766
PM10_lag1,lute_mittag, suho, ltusg_k, lute, frost, feb, mar, sat, wind	0.6693
PM10_lag1,lute_mittag, suho, ltusg_k, lute, frost, mar, sat, wind	0.6664
PM10_lag1,lute_mittag, suho, ltusg_k, feb, mar, sat, wind	06618
PM10_lag1,lute_mittag, suho, ltusg_k, mar, sat, wind	0.6588

Table 2.3: Summary Graz-Mitte

For descriptive purposes this model might be satisfying, but this model should be used for the prediction of PM10 concentration. This model includes explanatory variables which are

transformations of other variables included in the model (e.g *frost* and *lute*). Such a model is not desirable, therefore a reduction of the number of variables and a more rational selection of them is necessary.
Reducing the number of variables to a maximum of 7 variables leads to a model without meteorological dummy variables and an adjusted R^2 of 0.6588. The $\sqrt{PM10}$ at $t+2$ is described by the PM10 concentration at $t+1$ (**PM10_lag1**) and **lute_mittag, suho, ltusg_k, mar, sat, wind** for $t+2$. The selection of the variables is reliable and in line with the current one day prediction model. For the 2-days ahead forecast the information of the PM10 concentration at $t+1$ is given by the current one day prediction model. Nevertheless the prediction of the PM10 concentration depends essentially on the quality of the numerical meteorological forecast values for **lute_mittag** and **ltusg_k** at $t+2$. Therefore a possible improvement of this model might be the replacement of all metric meteorological variables by there transformed dummy variables (e.g **lute** by **frost**).
The assumption behind this step is that a prediction of the binary variable might be easier than an exact numerical forecast of meteorological values. Especially forecasting daily average values of temperature and temperature differences could be a source of inaccuracy in the latter 2 days ahead prediction of the PM10 concentration. Therefore we take into account the trade-off of the goodness of fit of the model and the vulnerability of the model against predictive errors of meteorological data. The lower described adjusted R^2 of 62.3% for the model excluding metric meteorological values will be compensated by its stability for the prediction.
Models with the additional variable **nied01** yields to slightly higher adjusted R^2, but they do not lead to better results in backtesting nor improving the quality of forecast.(In the sense of the quality function see: Section 2.1.4). Therefore we choose for Graz-Mitte a model without the precipitation variable.
In the following sections we are analyzing the models in detail and perform a backtesting.

2.1.3 Model Validation

The aim of this section is to present the estimation of the model parameters and to verify the underlying normality assumptions. The estimation of parameters for the model including metric meteorological data is based on the seasons 2002/2003-2013/2014.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.348008   0.082280  64.998  < 2e-16
pm10_lag1    0.036283   0.001083  33.507  < 2e-16
lute_mittag -0.096278   0.006049 -15.917  < 2e-16
suho        -0.698200   0.069432 -10.056  < 2e-16
ltusg_k     -0.230781   0.013915 -16.585  < 2e-16
wind         0.379790   0.058162   6.530 8.75e-11
```

```
mar          0.591254   0.072641   8.139 7.76e-16
sat         -0.539718   0.076203  -7.083 2.09e-12
```

The model fits quite good and all variables have a significant impact on the response variable $\sqrt{PM10}$. For the verification of the assumptions of the regression model we have to analyze the residuals. Considering Figure 2.4 and Figure 2.5 there is no indication of a violation of the assumptions. The distribution looks symmetric and reasonable and the standard error of the residuals is 1.051.

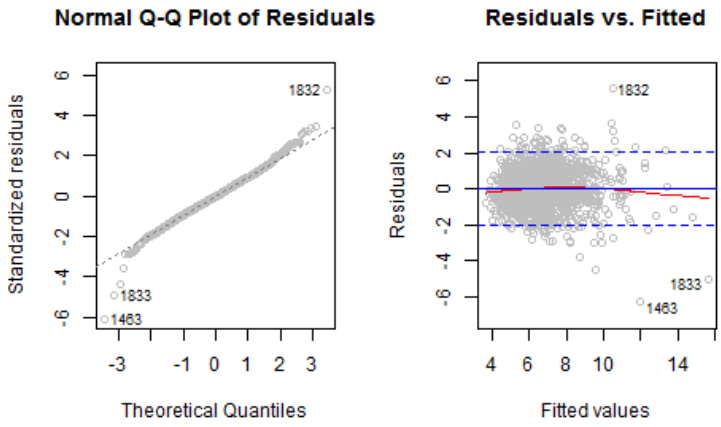

Figure 2.4: Q-Q-Plot of residuals vs. normal quantiles and Scatter-Plot of fitted values vs. residuals

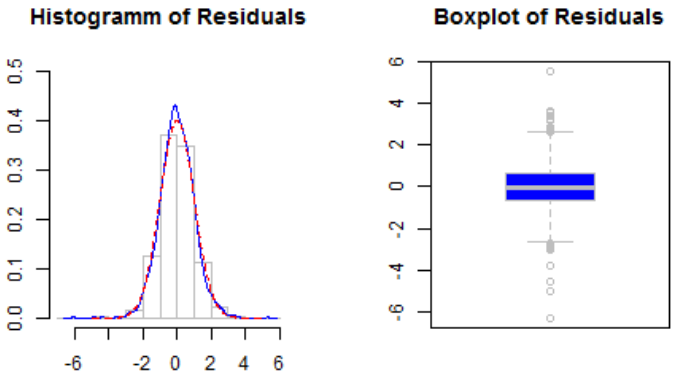

Figure 2.5: Histogram and Box-Plot of residuals

There are some outliers in the tails which leads to a rejection of the Shapiro-Wilk normality test.

```
Shapiro-Wilk normality test

W = 0.98791, p-value = 1.626e-10
```

The model will underestimate low as well extreme high concentrations of PM10. The consequence of this fact is a worse fit of the prediction values and a systematic under-prediction of high concentrations. The predictions are more conservative, because the prediction values of the model have a lower variance then the observed data points.

The fit of the model using only binary meteorological variables is reasonable and all variables within the model have a significant impact on the response variable.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.201956   0.069841  60.165  < 2e-16
pm10_lag1    0.039855   0.001098  36.297  < 2e-16
invers       0.892807   0.070246  12.710  < 2e-16
frost        0.824005   0.066970  12.304  < 2e-16
suho        -0.715761   0.072924  -9.815  < 2e-16
mar          0.313964   0.072953   4.304 1.78e-05
sat         -0.593133   0.079990  -7.415 1.94e-13
wind         0.493994   0.060407   8.178 5.72e-16
```

Similar to the first model the Figures 2.6 and 2.7 show a symmetric and reasonable fit of the model, but in comparison to the model with metric variables the standard error of the residuals is 1.10 and therefore sightly higher. The Shapiro-Wilk test reject the normality assumption caused by some outliers.

```
Shapiro-Wilk normality test

W = 0.9906, p-value = 7.935e-09
```

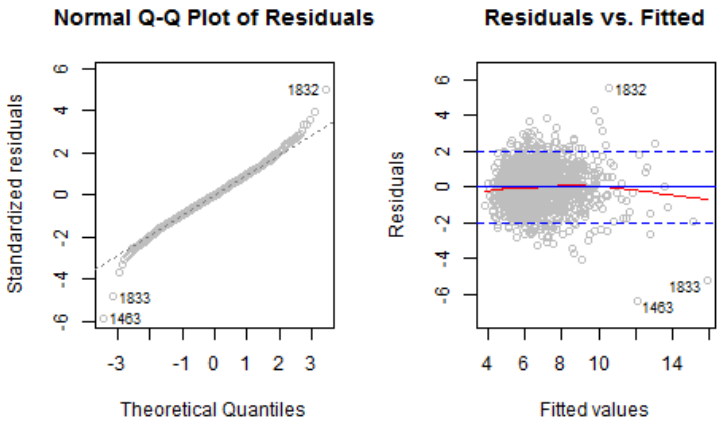

Figure 2.6: Q-Q-Plot of residuals vs. normal quantiles and Scatter-Plot of fitted values vs. residuals

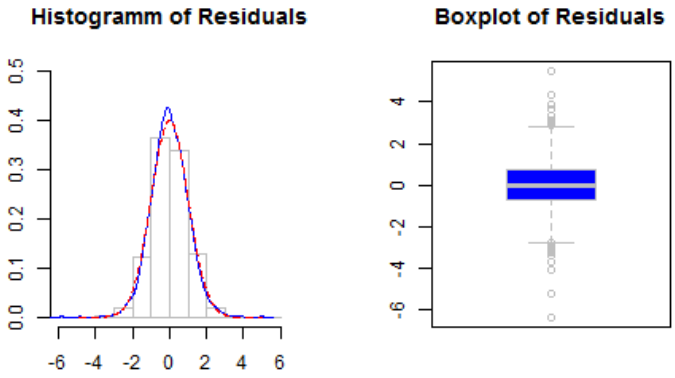

Figure 2.7: Histogram and Box-Plot of residuals

2.1.4 Quality of Forecasting

In the previous section we described the calibration and the accuracy of two different linear models for Graz-Mitte. The calculation of the coefficients β_i and the evaluation of reasonable models were performed. In this section we measure the quality of the model in forecasting PM10 values. To fulfill the needed assumptions the linear models yield prediction values of $\sqrt{PM10}$. By taking the square of the predicted values we receive the prediction of PM10. For our purpose the resulting bias is negligible.

Measuring the quality of forecasting is performed for both models and 2 winter seasons (2014/2015 and 2015/2016) separately. The first quality check of the model in forecasting is performed with measured values of the prediction variables to check the theoretical performance of the models. This assumption is very unrealistic in practice, because 2-days

meteorological forecasts and one day PM10 forecast values might produce considerable additional errors within the prediction.
The forecast results for the winter season 2014/2015 based on observed meteorological values and observed lagged PM10 values deliver satisfactory results for both models. For the PM10 forecast the EU limit of 50 $\mu g/m^3$ is essential for possible actions. Therefore our main focus lies on that limit. In the season 2014/2015 at 13 days an exceedance of the limit was observed. At 138 days the daily mean PM10 concentration was below the critical limit of 50 $\mu g/m^3$. Considering the forecast results of the model using metric meteorological regression variables in Table 2.4 we observe that the model fits the observation reasonably. Only at 17 days the 2-days ahead prediction model would forecast the wrong trend, which may imply wrong actions linked to the limit of 50 $\mu g/m^3$. The quality of the results of the model using only binary meteorological values (binary model) is similar to the first model. The model returns only in 16 cases a wrong indication of the PM10 concentration (Table 2.5).
But at one day a PM10 concentration higher than 100 $\mu g/m^3$ was observed. Both models cannot forecast this extreme value of PM10 concentration. Both models forecast a value below 100 $\mu g/m^3$, which gives a wrong indication. This high PM10 concentration was observed at New Years day 2015 and this high concentration might stem from the fireworks. Despite similar weather conditions at New years eve 2014 and at 2^{nd} January 2015 the daily average concentration is much lower. Such high concentrations of PM10 caused by external events can not be covered by the models. Similar to the forecast of the previous season 19 forecasts for the winter season 2015/2016 of the metric model indicates a wrong tendency and 23 in the model which uses binary meteorological variables (Table 2.6 and Table 2.7).

	Forecast 0-50 $\mu g/m^3$	Forecast 50-100 $\mu g/m^3$	Forecast $>$ 100 $\mu g/m^3$	Total
Observation 0-50 $\mu g/m^3$	131	7	0	138
Observation 50-100 $\mu g/m^3$	8	3	1	12
Observation $>$ 100 $\mu g/m^3$	0	1	0	1
Total	139	11	1	151

Table 2.4: Forecasting Results of the metric Model, Winter Season 2014/2015

	Forecast 0-50 $\mu g/m^3$	Forecast 50-100 $\mu g/m^3$	Forecast $>$ 100 $\mu g/m^3$	Total
Observation 0-50 $\mu g/m^3$	134	4	0	138
Observation 50-100 $\mu g/m^3$	9	2	1	12
Observation $>$ 100 $\mu g/m^3$	0	1	0	1
Total	143	7	1	151

Table 2.5: Forecasting Results of the binary Model, Winter Season 2014/2015

	Forecast 0-50 $\mu g/m^3$	Forecast 50-100 $\mu g/m^3$	Total
Observation 0-50 $\mu g/m^3$	114	15	129
Observation 50-100 $\mu g/m^3$	4	19	23
Total	118	34	152

Table 2.6: Forecasting Results of the metric Model, Winter Season 2015/2016

	Forecast 0-50 $\mu g/m^3$	Forecast 50-100 $\mu g/m^3$	Forecast > 100 $\mu g/m^3$	Total
Observation 0-50 $\mu g/m^3$	114	15	0	129
Observation 50-100 $\mu g/m^3$	7	15	1	23
Total	121	30	1	152

Table 2.7: Forecasting Results of the binary Model, Winter Season 2015/2016

In more than 84 percent of observations for the winter seasons 2014/2015 and 2015/2016 the 2-days ahead prediction models give the right indication in sense of the EU limit of 50 $\mu g/m^3$ for the average PM10 concentration. It's a first good indication, but to get a better feeling for the quality of fit Stadlober et al. [23] introduced a quality function which is tailored to the particular needs of this limit. This function transforms absolute errors of forecast-values and observed values to a reasonable rating system. In addition, this function takes into account that a large deviation of the forecast from the observation, but indicating the right decision have a better quality than small errors indicating possibly wrong actions. In detail a forecast of 53 $\mu g/m^3$ and an observation of 40 $\mu g/m^3$ has a lower quality than a forecast of 105 $\mu g/m^3$ and an observation of 140 $\mu g/m^3$. The first error implies wrong actions, therefore a higher penalization of this forecasting error is necessary. The main focus of this quality function lies on forecast and observation values above the limit of 50 $\mu g/m^3$ and below the limit of 100 $\mu g/m^3$. Forecasting errors where both values are below the critical limit of 50 $\mu g/m^3$ have a lower penalization then errors causing a wrong action.

Definition 2.1 (Quality function I) *The quality function $Q(O, F)$ assigns to each pair consisting of an observation O and a forecast value F in $[0, 1]$. The quality function $Q : \mathbb{R}^2_+ \to [0, 1]$ is given by:*

$$Q(O, F) = 1 - \min\left\{\frac{a \times |O - F|}{D}, 1\right\} \textit{ with}$$

$$D = 1 + \frac{1}{2}\sqrt{|O - 50| + |F - 50| + b \times \boldsymbol{I}_B + c \times \boldsymbol{I}_C}$$

where $\boldsymbol{I}_B$ is the indicator function, if $\{O \leq 50 \quad and \quad F \leq 50\}$ is fulfilled. $\boldsymbol{I}_C$ is equal to 1 if $\{O \geq 100 \quad and \quad F \geq 100\}$ is fulfilled and zero otherwise. According Stadlober et al. [23] the specific choice of the parameters is $a = 0.1$, $b = 100$ and $c = 1000$.

This quality function is limited to our specific problem in forecasting PM10 vales. The choice of this functions guarantees the strongest penalties for the following events:

$$\{O \leq 50 \quad \text{and} \quad F \geq 50\} \qquad \{O \geq 50 \quad \text{and} \quad F \leq 50\}$$

A contour plot of the quality function can be found in Stadlober et al. [23]. The results of the quality function are thereafter assigned to the following 5 grades (categorical values):

- $Q(O, F) \geq 0.8 \Longrightarrow$ "excellent"
- $0.6 \leq Q(O, F) < 0.8 \Longrightarrow$ "good"
- $0.4 \leq Q(O, F) < 0.6 \Longrightarrow$ "satisfying"
- $0.2 \leq Q(O, F) < 0.4 \Longrightarrow$ "bad"
- $Q(O, F) < 0.2 \Longrightarrow$ "very bad"

The results of the quality function for the winter season 2014/2015 for both models can be found in Table 2.8. More than 90 percent of forecasted values of the metric model and around 90 percent of the binary model are rated with "good" or "excellent", which is a remarkable result for the theoretical prediction accuracy. The results for the season 2015/2016 (Table 2.9) look similar, the quality of the fit is still considerable, more than 90 percent are in the category excellent or good.

			Model with binary variables					Total
			"excellent"	"good"	"satisfying"	"bad"	"very bad"	
Model with	metric variables	"excellent"	107	9	0	0	0	116
		"good"	9	10	1	0	0	20
		"satisfying"	1	0	5	1	1	8
		"bad"	0	1	0	2	0	3
		"very bad"	1	1	1	0	1	4
		Total	118	21	7	3	2	151

Table 2.8: Results of Quality function I for Winter Season 2014/2015

			Model with binary variables					Total
			”excellent”	”good”	”satisfying”	”bad”	”very bad”	
Model with	metric variables	”excellent”	95	10	2	2	0	109
		”good”	10	10	2	2	0	24
		”satisfying”	4	4	4	2	0	14
		”bad”	0	0	1	1	0	2
		”very bad”	0	0	0	2	1	3
		Total	109	24	9	9	1	152

Table 2.9: Results of Quality function I for Winter Season 2015/2016

In the sense of the first quality function both models deliver remarkable results for the forecasting values. The first quality function is a symmetric function, where the absolute difference error between measured and estimated value has the same impact on the quality of the estimation. There is no distinction, if the estimation is higher than the observed value or vice versa. In addition the first quality function does not take into consideration if the observation is below the limit and estimated value is above it or other way round. Especially a forecast below the limit of 50 $\mu g/m^3$ and a measured value above the limit should have a separate penalization. This case has adverse consequences, because of the lower forecast, actions needed to reduce the PM10 concentration are not performed, while the concentration will exceed the limit. An estimation higher then the limit and higher than the observation does not have such adverse consequences. A higher estimated concentration can be seen as a more conservative approach. The higher forecast motivates short time actions for reducing the PM10 concentration and its adverse health effect. Both situations with different penalization are considered in the following quality function. The quality function is not symmetric anymore.

Definition 2.2 (Quality Function II) *The quality function $Q(O,F)$ assigns to each pair consisting of an observation O and a forecast value F in $[0,1]$. The quality function $Q : \mathbb{R}^2_+ \to [0,1]$ is given by:*

$$Q(O,F) = 1 - \min\left\{ \boldsymbol{I}_A * \min\left\{ \frac{d \times |O-F|}{D_2}, 1 \right\} - \min\left\{ \frac{a \times |O-F|}{D}, 1 \right\}, 1 \right\} \qquad \textit{with}$$

$$D_1 = 1 + \frac{1}{2}\sqrt{|O-50| + |F-50| + b \times \boldsymbol{I}_B + c \times \boldsymbol{I}_C}$$

$$D_2 = 1 + \sqrt{\min\{|O-50|, |O-100|\} + \min\{|F-50|, |F-100|\}}$$

where $\boldsymbol{I}_A$ is equal to 1, if $\{F \leq 50 \text{ and } O \geq 50\}$ or $\{F \leq 100 \text{ and } O \geq 100\}$ is fulfilled .$\boldsymbol{I}_B$ is the indicator function, if $\{O \leq 50 \text{ and } F \leq 50\}$ is fulfilled. $\boldsymbol{I}_C$ is equal to 1 if $\{O \geq 100 \text{ and } F \geq 100\}$ is fulfilled and zero otherwise. The specific choice of the parameters is similar to the first quality function: $a = 0.1, b = 100, c = 1000$ and $d = 1$.

The assignment of the quality function values to the different classes is performed similarly to the first quality function. If both values are below 50 $\mu g/m^3$ or above 100 $\mu g/m^3$ the quality function doesn't change. The asymmetric property of the quality function can be seen in Figure 2.8, especially for measured values and estimated values next to the limits of 50 $\mu g/m^3$ and 100 $\mu g/m^3$.
The results in the higher ratings of Quality Function II are very similar to the results of the first Quality Function, but the number of "very bad" forecasts increased sharply (see Tables 2.10 and 2.11). The reason for this increase is the higher penalization of lower forecasted values linked to higher measured values.

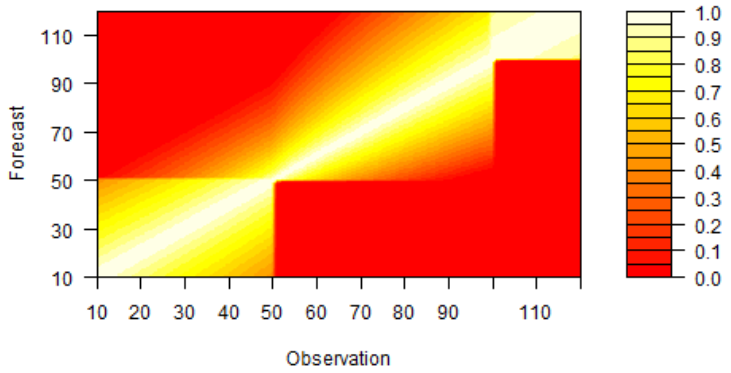

Figure 2.8: Contour plot of the quality function II

			Model with binary variables					Total
			"excellent"	"good"	"satisfying"	"bad"	"very bad"	
Model with metric variables		"excellent"	107	9	0	0	0	116
		"good"	9	8	0	0	0	17
		"satisfying"	1	0	2	0	2	5
		"bad"	0	1	0	0	0	1
		"very bad"	1	1	1	0	9	12
		Total	118	19	3	0	11	151

Table 2.10: Results of Quality function II for Winter Season 2014/2015

			Model with binary variables					Total
			”excellent”	”good”	”satisfying”	”bad”	”very bad”	
Model with	metric variables	”excellent”	94	9	2	1	2	108
		”good”	10	8	2	2	2	24
		”satisfying”	4	4	2	2	0	12
		”bad”	0	0	1	1	0	2
		”very bad”	1	0	0	2	3	6
		Total	109	21	7	8	7	152

Table 2.11: Results of Quality function II for Winter Season 2015/2016

2.1.5 Possible further Improvements of the Model

Within this section we try to improve the chosen models for Graz-Mitte. In the first approach we are reducing the seasons for the calibration of the model. In the data we see a slightly decreasing trend in the PM10 concentration, therefore another approach could be a detrending of the PM10 concentration for the calibration of the models.

2.1.5.1 Reducing the amount of seasons

In the 2-days ahead forecasting of PM10 for the winter seasons 2014/2015 and 2015/2016 we see that the forecast is in general higher than the observation. Considering Figure 2.2 this effect is caused by the decreasing PM10 concentration over time. Within the data set (see Table 2.1) we see in the season 2009/2010 a long period without any PM10 values. The reason for that is the relocation of the measurement station for Graz-Mitte (February 18^{th} 2010) see [23]. In addition to the decreasing PM10 concentration we have a structural change in the data, due the relocation of the station. In the following the calibration of the models is based on 4 winter seasons 2010/2011-2013/2014. Within this analysis the choice of the parameters stays unchanged. Reducing the number of seasons lead to a reduction of the variance described to 61.97% for the model using metric meteorological regression variables. For the model using only binary meteorological variables the adjusted R^2 is 0.5922.
The residuals are symmetrical distributed, except some outliers in the upper tail. Similar to the model, with the whole data history and the structural break of measurements, the Shapiro-Wilk Test is rejected. The standard error of the residuals with 0.912 for the metric model and 0.944 for the binary model is lower than in the models in the former section.

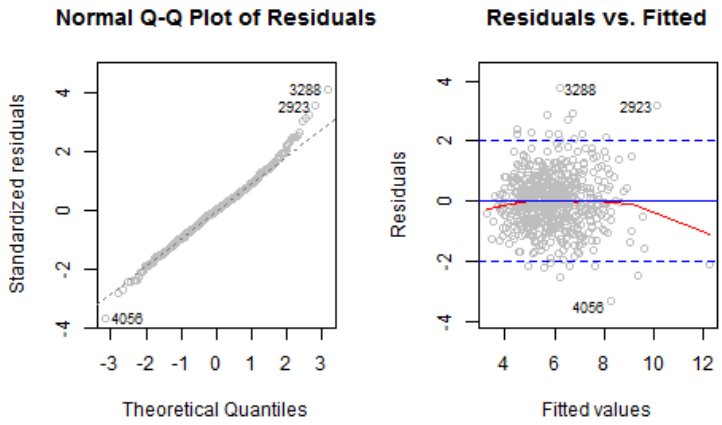

Figure 2.9: Q-Q-Plot of residuals vs. normal quantiles and Scatter-Plot of fitted values vs. residuals

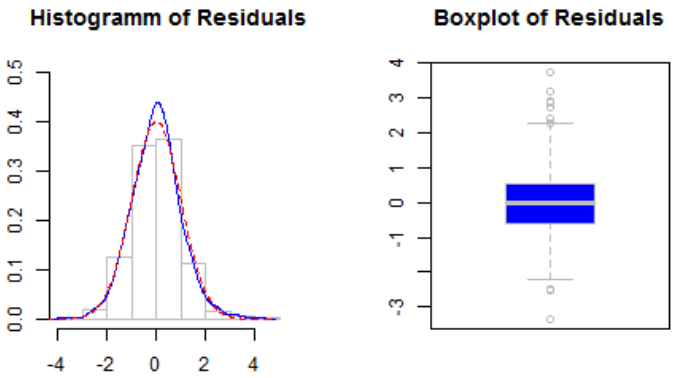

Figure 2.10: Histogram and Box-Plot of residuals

```
Shapiro-Wilk normality test

W = 0.99223, p-value = 0.003118
```

For the model using only binary regression variables the residuals looks very similar and the model could be a good choice for the 2-days ahead prediction.

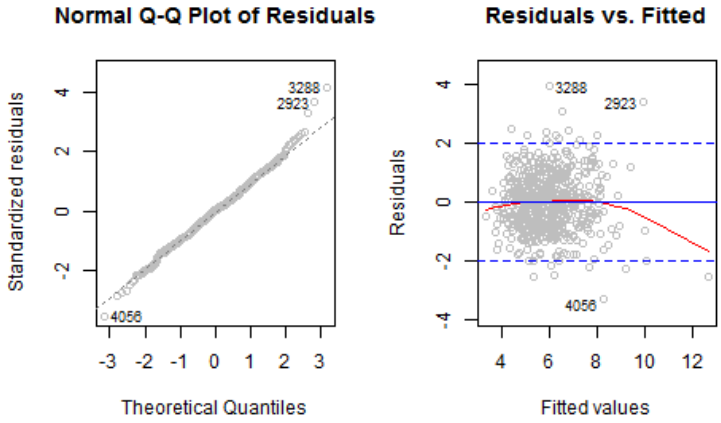

Figure 2.11: Q-Q-Plot of residuals vs. normal quantiles and Scatter-Plot of fitted values vs. residuals

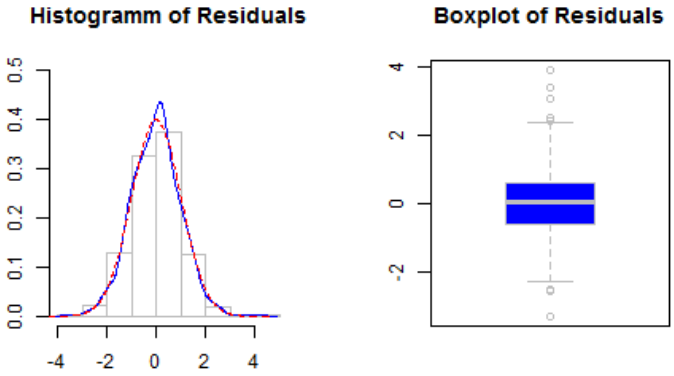

Figure 2.12: Histogram and Box-Plot of residuals

```
Shapiro-Wilk normality test

W = 0.99466, p-value = 0.03395
```

Essential for the choice of the model is the quality of the forecast for the seasons 2014/2015 and 2015/2016 (compare: Tables 2.12 - 2.15). The forecast results for the critical level of the 50 and 100 $\mu g/m^3$ daily average concentration in the season 2014/2015 are very similar to the results of the forecasted values produced by the model using all available seasons. The increase of the rate predicting the right indication for the season 2014/2015 is only 1 percent. But for the season 2015/2016 the model calibrated to a data set of 4 winter seasons gives a much higher ratio of true classification. The ratio of the right indication increases from 84% to 89% for the binary model and from 87.5% to 91.4% for the metric model. Especially the

increase of the ratio for the model using only binary weather variables is very pleasant. If higher predictions are not rated as a wrong prediction, the hit ratio would increase to 92.7%.

	Forecast 0-50 $\mu g/m^3$	Forecast 50-100 $\mu g/m^3$	Forecast $>$ 100 $\mu g/m^3$	Total
Observation 0-50 $\mu g/m^3$	134	4	0	138
Observation 50-100 $\mu g/m^3$	10	1	1	12
Observation $>$ 100 $\mu g/m^3$	1	0	0	1
Total	145	5	1	151

Table 2.12: Forecasting Results of the metric Model using 4 Seasons, Winter Season 2014-2015

	Forecast 0-50 $\mu g/m^3$	Forecast 50-100 $\mu g/m^3$	Total
Observation 0-50 $\mu g/m^3$	136	2	138
Observation 50-100 $\mu g/m^3$	10	2	12
Observation $>$ 100 $\mu g/m^3$	1	0	1
Total	147	4	151

Table 2.13: Forecasting Results of the binary Model using 4 Seasons, Winter Season 2014-2015

	Forecast 0-50 $\mu g/m^3$	Forecast 50-100 $\mu g/m^3$	Total
Observation 0-50 $\mu g/m^3$	124	5	129
Observation 50-100 $\mu g/m^3$	8	15	23
Total	133	19	152

Table 2.14: Forecasting Results of the metric Model using 4 Seasons, Winter Season 2015-2016

	Forecast 0-50 $\mu g/m^3$	Forecast 50-100 $\mu g/m^3$	Total
Observation 0-50 $\mu g/m^3$	123	6	129
Observation 50-100 $\mu g/m^3$	10	13	23
Total	133	19	152

Table 2.15: Forecasting Results of the binary Model using 4 Seasons, Winter Season 2015-2016

Considering quality function II for both seasons in Table 2.16 and Table 2.17 we see that most of the predicted values are categorized with ”excellent” or ”good”. In contrast to the

results of the model calibrated with the whole data history more predicted values are classified with ”very bad”. The model calibrated by using 4 winter seasons is reasonable, because the bias from the relocation of the measurement station is removed.

			Model with binary variables					Total
			”excellent”	”good”	”satisfying”	”bad”	”very bad”	
Model with	metric variables	”excellent”	113	8	0	0	0	121
		”good”	8	6	1	0	0	15
		”satisfying”	1	1	0	0	0	2
		”bad”	1	0	1	0	0	2
		”very bad”	0	0	0	0	11	11
		Total	123	15	2	0	11	151

Table 2.16: Results of Quality function II for Winter Season 2014/2015

			Model with binary variables					Total
			”excellent”	”good”	”satisfying”	”bad”	”very bad”	
Model with	metric variables	”excellent”	108	10	1	0	3	122
		”good”	4	7	0	1	0	12
		”satisfying”	0	1	5	0	1	7
		”bad”	0	1	1	1	0	3
		”very bad”	1	0	1	0	6	8
		Total	113	19	8	2	10	152

Table 2.17: Results of Quality function II for Winter Season 2015/2016

By comparing the observed PM10 concentration with the forecasted values for the winter season 2014-2015 in Figure 2.13 we see that the models underestimate a longer period of higher PM10 values in February 2015 and in March 2015. Otherwise the model is very close to the observations. For the winter season 2015/2016 except the overestimation of the PM10 concentration in march the comparison looks very good (Figure 2.14).

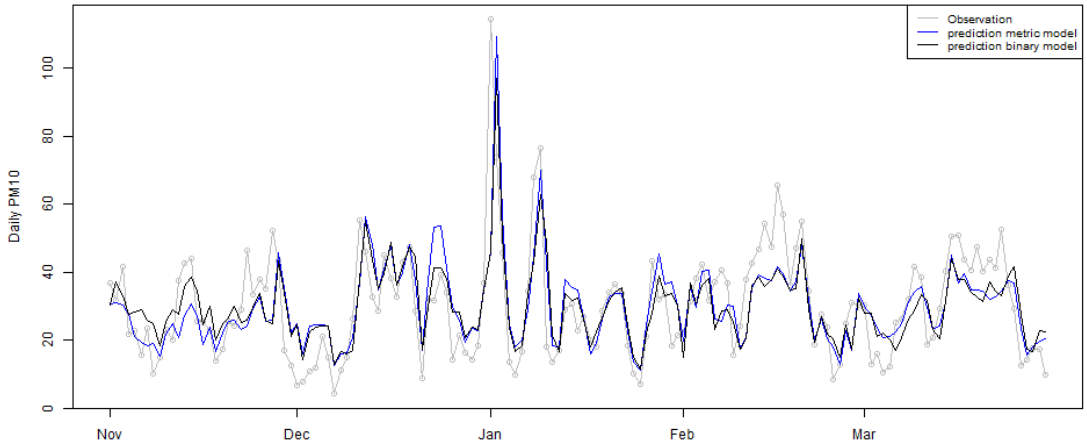

Figure 2.13: Comparison of observed and predicted values for the winter season 2014/2015

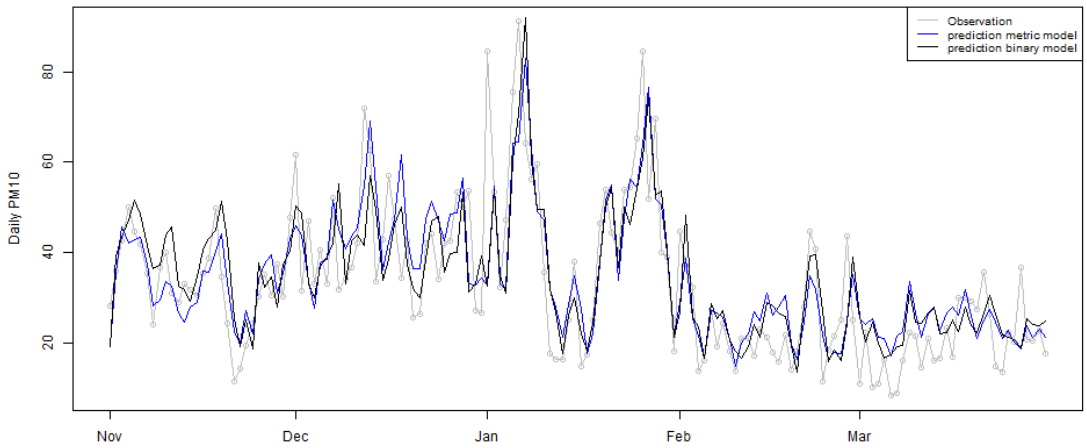

Figure 2.14: Comparison of observed and predicted values for the winter season 2015/2016

In the Figures 2.15 and 2.16 see in the left upper corner and in the right lower corner a Scatter-Plot of the observations and the predicted values of the according model combined with the quality function. We see that especially for extreme observations the quality of the fit is bad. In the right upper plot on can see that both models predict similar values. The mass of the dots is close to the diagonal line.

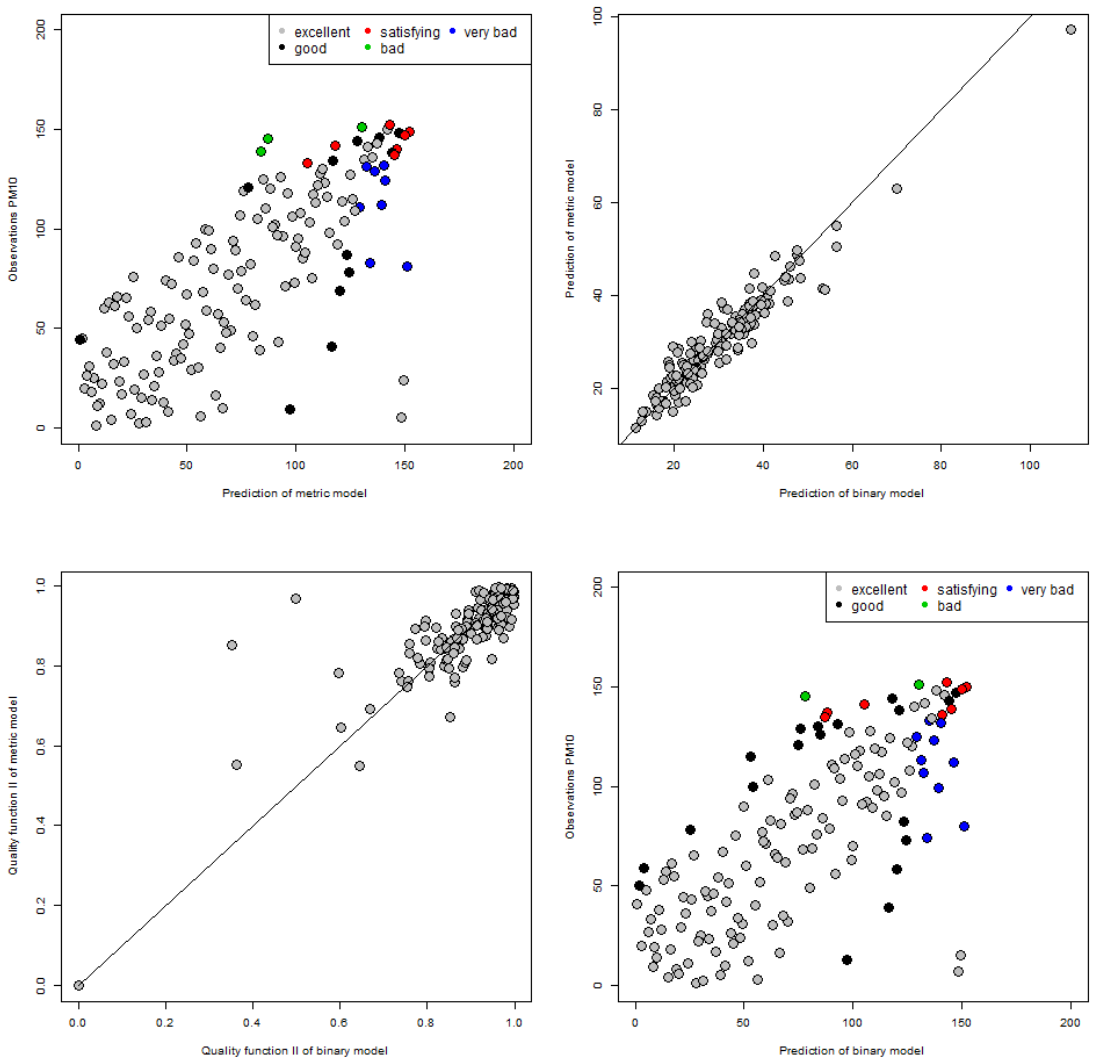

Figure 2.15: Analysis of the Forecasts for the winter season 2014/2015

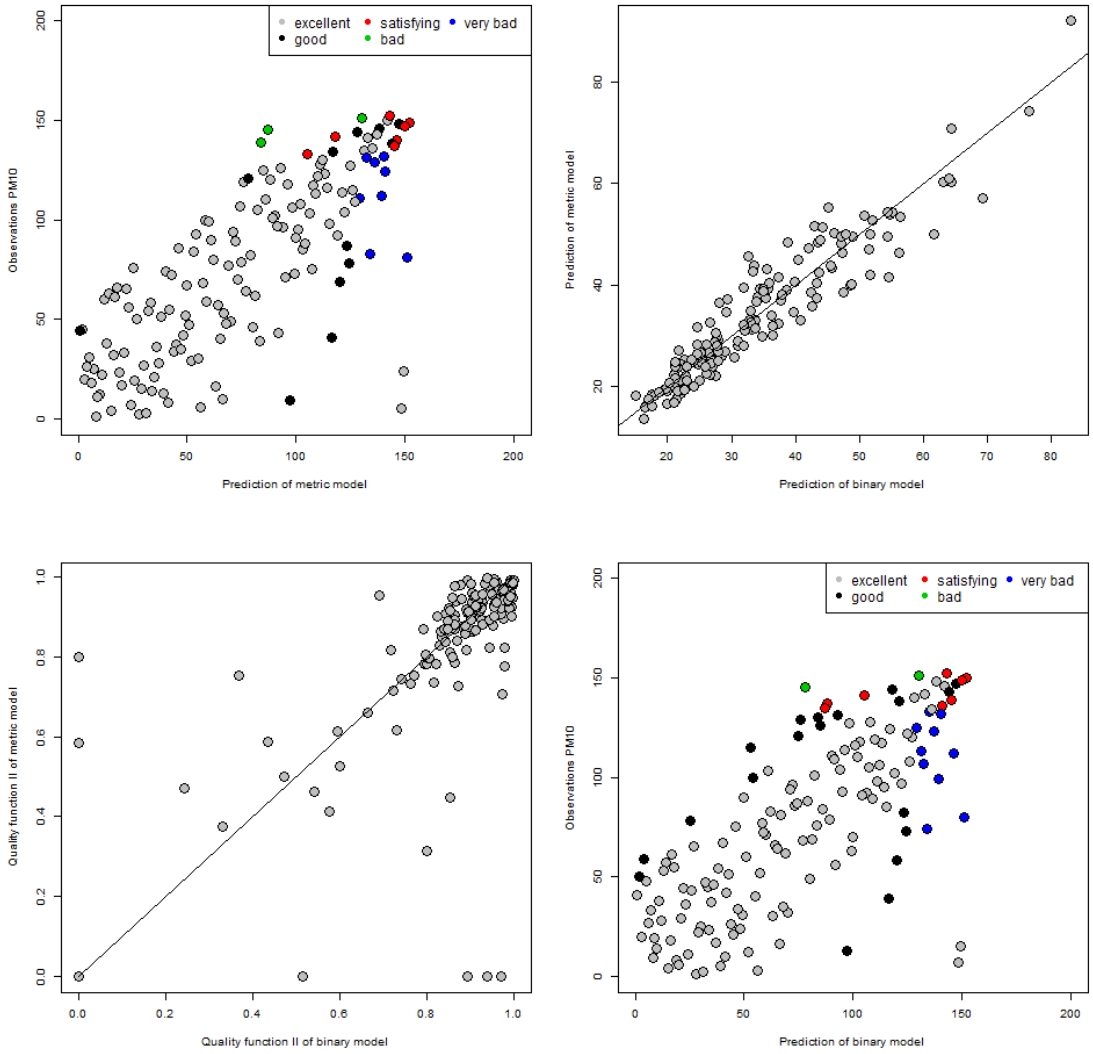

Figure 2.16: Analysis of the Forecasts for the winter season 2015/2016

In the next section we are trying to improve the model by detrending the particulate matter concentration.

2.1.5.2 Detrending of PM10 concentration

We observe a continuously decrease of the PM10 concentration over the last years (see Figure 2.2). An alternative approach to model the particulate matter concentration is to detrend the given observations using a linear trend. Detrending the daily average of the PM10 concentration yields to negative values of concentrations. Therefore modeling the square root of the concentration is not possible anymore. For a linear model to the detrended PM10 concentrations the requirements for the residuals are not fulfilled. The distribution of the residuals is not symmetric nor the variance of the residuals seems to be constant. In addition the fit of the model is not an improvement of the models above.

2.1.6 Backtesting under more realistic assumptions

In the previous section we carried out a continuous improvement of the models for Graz-Mitte. The final model delivers reasonable results in the backtesting for the seasons 2014/2015 and 2015/2016. The backtesting in the previous section was based on observed weather and particulate matter values. In this section we perform a backtesting of the models under more realistic conditions. Therefore we have to adjust the data set used for the prediction.
For the lagged PM10 concentration in the model the observed values are replaced by the predicted values of the current one-day ahead model for PM10. This model is described in detail in Stadlober et al. [23]. In addition we are adding some random noice to the metric variables in the model. The metric variables **lute_mittag** and **ltusg_k** are adjusted by adding a random noise. The random noise is normally distributed with mean 0 and standard deviation, which is two times the empirical standard deviation of the measured data set. For the binary variables **frost**, **invers** and **wind** we assume that 35% of the meteorological 2-days ahead predictions are wrong. Every third observation, i.e. around 50 days in the winter season are replaced by a wrong value. These modifications of the data used for the prediction should give us an insight in the quality of the predictions under more realistic circumstances and the robustness of the model. Considering the results for the season 2014/2015 in the Tables 2.18 - 2.19 we still see a remarkable high hit ratio of the predictions. The ratio is still above 85% for the metric model and almost unchanged for the binary model. Looking at the quality function II of forecasts for this season (see Table 2.22) we see a slightly loss of quality. The quality of some forecasts moves from 'excellent' to 'good'.
For the winter season 2015/2016 the hit ratio for the metric model decreases from 89% to 78% and for the binary model from 91.4% to 80%. Similar to the theoretical prediction the quality of the predictions under realistic assumptions decreases for season 2015/2016.
Within this backtesting we take into account the variability of the meteorological two-days ahead forecast. On the one hand this analysis shows that the model delivers good forecasts if the meteorological forecast is not exact, but on the other hand these results raises a question if a model without any meteorological input could deliver reasonable results. In the following section we calibrate a model which includes only lagged PM10 concentration and calendarian variables.

	Forecast 0-50 $\mu g/m^3$	Forecast 50-100 $\mu g/m^3$	Forecast > 100 $\mu g/m^3$	Total
Observation 0-50 $\mu g/m^3$	126	12	0	138
Observation 50-100 $\mu g/m^3$	9	3	0	12
Observation > 100 $\mu g/m^3$	1	0	0	1
Total	136	15	0	151

Table 2.18: Forecasting Results of the metric Model using 4 Seasons under realistic assumptions, Winter Season 2014-2015

	Forecast 0-50 $\mu g/m^3$	Forecast 50-100 $\mu g/m^3$	Total
Observation 0-50 $\mu g/m^3$	136	2	138
Observation 50-100 $\mu g/m^3$	11	1	12
Observation $>$ 100 $\mu g/m^3$	1	0	1
Total 148	3	151	

Table 2.19: Forecasting Results of the binary Model using 4 Seasons under realistic assumptions, Winter Season 2014-2015

	Forecast 0-50 $\mu g/m^3$	Forecast 50-100 $\mu g/m^3$	Total
Observation 0-50 $\mu g/m^3$	110	19	129
Observation 50-100 $\mu g/m^3$	14	9	23
Total	124	28	152

Table 2.20: Forecasting Results of the metric Model using 4 Seasons under realistic assumptions, Winter Season 2015-2016

	Forecast 0-50 $\mu g/m^3$	Forecast 50-100 $\mu g/m^3$	Total
Observation 0-50 $\mu g/m^3$	117	12	129
Observation 50-100 $\mu g/m^3$	18	5	23
Total	135	17	152

Table 2.21: Forecasting Results of the binary Model using 4 Seasons under realistic assumptions, Winter Season 2015-2016

			Model with binary variables					Total
			”excellent”	”good”	”satisfying”	”bad”	”very bad”	
Model with	metric variables	”excellent”	78	12	0	0	1	91
		”good”	20	10	0	0	0	30
		”satisfying”	3	5	1	0	1	10
		”bad”	6	0	0	0	0	6
		”very bad”	2	2	0	0	10	14
		Total	109	29	1	0	12	151

Table 2.22: Results of Quality function II for Winter Season 2014/2015

			Model with binary variables					Total
			”excellent”	”good”	”satisfying”	”bad”	”very bad”	
Model with	metric variables	”excellent”	51	23	4	2	4	84
		”good”	14	11	1	0	2	28
		”satisfying”	6	2	0	0	1	9
		”bad”	6	1	1	1	0	9
		”very bad”	2	3	2	0	15	22
		Total	79	40	8	3	22	152

Table 2.23: Results of Quality function II for Winter Season 2015/2016

2.1.6.1 Model without meteorological variables

In the previous section we have demonstrated that the models give a good 2 days ahead prediction of particulate matter for Graz Mitte if we simulate the meteorological prediction errors and use the one day prediction values from the current model. The ratio of the right indication according the EU limit of 50 $\mu g/m^3$ for the model is still above 79% using simulated values.

Hörmann et al. [10] mention that the PM10 concentration for Graz is highly correlated with meteorological values like wind, temperature, inversion temperature. Therefore all considered regression models in the previous sections include meteorological values. But for the considered 2-days ahead prediction models, 2-days meteorological predictions are needed as input for a reasonable forecast of PM10. If a good indication of the PM10 concentration is sufficient, a much simpler model without any meteorological components will deliver the desired results. Using the model for forecasting in reality we deal with the compromise between the complexity of the model and the accuracy and quality of the forecast. Therefore within this section we are analysing a model without any meteorological parameters.

The model will be calibrated to the data set, after the relocation of the measurement station (4 winter seasons). Similar to the procedure for the model including meteorological we try to find a model with a high adjusted R^2 and a low number of variables. The following variables are available for a possible model: **suho, sat, wd, nov, dec, jan, feb, mar** and **PM10_lag1**. The model which explains the biggest part of the variance in the data is given by the following variables:

PM10_lag1, **suho**, **sat**, **mar**

The adjusted R^2 is 45.98% which is considerable lower than the R^2 of the models including meteorological variables. (see Section 2.1.5.1). The summary of the model shows that **PM10_lag1**, **suho**, **sat** are very significant, while the significance of **mar** is much lower.

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.239499   0.109388  38.757  < 2e-16
pm10_lag1    0.050127   0.002352  21.308  < 2e-16
suho        -0.483630   0.119642  -4.042 5.99e-05
```

```
sat          -0.469583    0.130106  -3.609 0.000333
mar          -0.214212    0.110455  -1.939 0.052929
```

By considering the residuals in Figure 2.17 and 2.18 more outliers in the tails can be observed. Although the Shapiro-Wilk Test rejects the normal distribution of the residuals, the histogram and the boxplot represent a sightly right skewed distribution. This fact will lead to an underestimation of high PM10 concentrations. The standard error of the residuals (1.087) is also higher than in the former models which emphasizes the underestimation. The variance of the residuals is indicated in Figure 2.17 and looks constant. One of the outliers marked in Figure 2.17 are New years day with a very high PM10 concentration caused by fireworks.

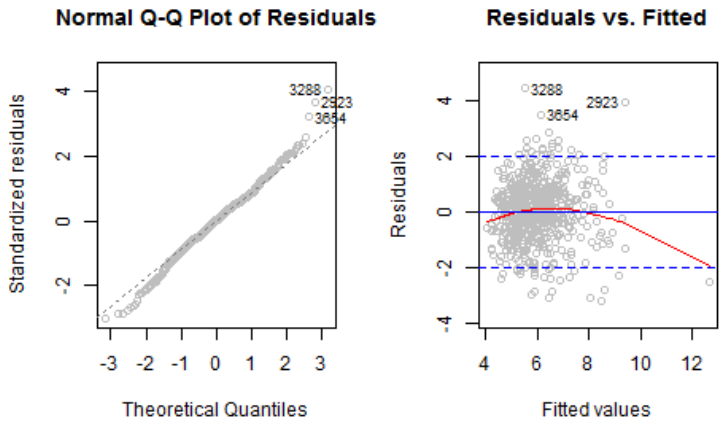

Figure 2.17: Q-Q-Plot of residuals vs. normal quantiles and Scatter-Plot of fitted values vs. residuals

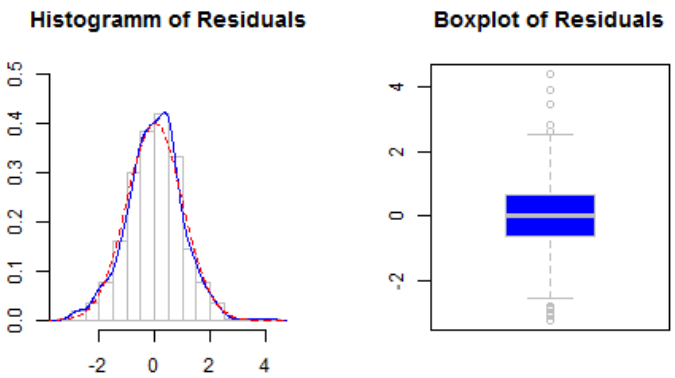

Figure 2.18: Histogram and Box-Plot of residuals

```
Shapiro-Wilk normality test
```

```
W = 0.9931, p-value = 0.006869
```

The fit of this model is worse than the model including meteorological variables, but it is essential how this model performs in forecasting for the winter seasons 2014/2015 and 2015/2016. The performance test is already based on the predicted values of the current one day PM10 model described in Stadlober et al. [23].
According the 50 $\mu g/m^3$ limit the ratio of right indication is remarkable high for both seasons. For the winter season 2014/15 the ratio is 91% (see Table: 2.24) and for the season 2015/16 the ratio is 86% (see Table: 2.25). The Table for the season 2015/2016 indicates an underestimation of PM10-concentrations above 50 $\mu g/m^3$ and emphasizes the assumption assumed by considering the residual plots.

	Forecast 0-50 $\mu g/m^3$	Forecast 50-100 $\mu g/m^3$	Total
Observation 0-50 $\mu g/m^3$	134	4	138
Observation 50-100 $\mu g/m^3$	9	3	12
Observation $>$ 100 $\mu g/m^3$	1	0	1
Total	144	7	151

Table 2.24: Forecasting Results of the Model without meteorological variables under realistic assumptions, Winter Season 2014-2015

	Forecast 0-50 $\mu g/m^3$	Forecast 50-100 $\mu g/m^3$	Total
Observation 0-50 $\mu g/m^3$	127	2	129
Observation 50-100 $\mu g/m^3$	18	5	23
Total	145	7	152

Table 2.25: Forecasting Results of the Model without meteorological variables under realistic assumptions, Winter Season 2015-2016

In addition to the analysis of the right indication of the PM10 concentration we are analysing the quality (in sense of quality function II) of the forecasts. On can see in Table 2.26 and 2.27 that the quality of the model predictions is considerable. Most of the forecast are classified as 'excellent' or 'good'. The underestimated values are classified as 'very bad', which is in line with the results shown in Table 2.24 and 2.25.

Model with no meteorologic variables					Total
"excellent"	"good"	"satisfying"	"bad"	"very bad"	
85	44	9	2	11	151

Table 2.26: Results of Quality function II for Winter Season 2014/2015

Model with no meteorologic variables					Total
”excellent”	”good”	”satisfying”	”bad”	”very bad”	
96	34	3	1	18	152

Table 2.27: Results of Quality function II for Winter Season 2015/2016

In this section we analyzed the data for Graz-Mitte, considered different models for a two-days prediction and improved them continuously. In the following section we calibrate different models for the location Graz Süd.

2.2 Graz-Süd

Within this section we develop different 2-days prediction models for the site Graz-Süd. This measurement station is located next the industrial zone and in an area with high traffic frequency. The first PM10 concentration at this site was measured in April 2003.

2.2.1 Data

The measurement of the daily PM10 concentration at the site Graz Süd has been started at April 25^{th} 2003, therefore the data set contains 4725 entries (including summer season) of daily average PM10 concentration data points. To set up a valid model for Graz-Süd the first 4 seasons (2002/03-2005/06) can not be considered, because recording of meteorological values for Graz-Süd has been started at April 5^{th} 2006. (see [23]) The variables contained in the data set of Graz-Süd are the same as in the data set of Graz-Mitte. The number of missing data for the individual winter seasons is represented in Table 2.28.

Winter Season	**Missing Values**
2002/2003	no data available
2003/2004-2004/2005	no missing values
2005/2006	November 20^{th} 2005-November 21^{st} 2005
2006/2007	November 7^{th} 2006-November 13^{th} 2006
2007/2008-2015/2016	no missing values

Table 2.28: Missing PM10 Values Graz-Süd

Considering Figure 2.19 on can see a downward sloping trend of the average PM10 concentration over the last winter seasons. The reasons for this evolutions are similar to Graz-Mitte, better action of the Styrian provincial government and the climate change. By comparing the daily average PM10 concentration of Graz-Mitte with the average PM10 concentration at Graz-Süd on can see, that the PM10 concentration at Graz-Süd is higher (see: Table: 2.29). The reason for this effect is that the measurement station is located next to the industrial zone and in an area with higher traffic frequency. Stadlober et al. [23] show

that correlation of PM10 concentration and meteorological factors at Graz Süd exists and is similar to Graz-Mitte.

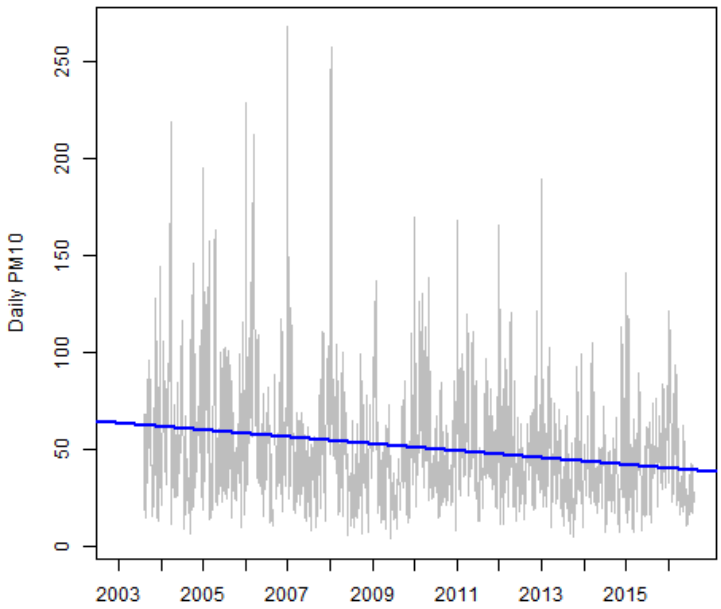

Figure 2.19: Daily PM10 average values for the winter seasons, Graz Süed

Season	Observations	Average PM10	Average Temp in C°	Inversion Days	Frost Days	Wind Days
2006/2007	144	51.88	4.37	70	17	44
2007/2008	152	55.15	2.69	58	37	64
2008/2009	151	42.60	2.38	45	42	52
2009/2010	151	53.86	1.92	56	58	43
2010/2011	151	53.13	1.95	69	62	39
2011/2012	152	49.37	2.44	69	44	62
2012/2013	151	46.43	2.13	51	43	57
2013/2014	151	39.84	4.39	61	27	51
2014/2015	151	43.10	4.34	40	21	58
2015/2016	152	44.76	3.87	77	22	46

Table 2.29: Summary Graz-Süd

2.2.2 Model

To calibrate a reasonable model for Graz-Süd the same methodology as for Graz-Mitte is applied. First we generate the regression models with all possible combinations of regression variables (at least 4) described in section 2.1.2. Next we try to find a model which fulfills the criteria of a low number of variables and a high ratio of described variance. As for Graz-Mitte the model should be useful for 2 days ahead predication. In the following we consider 2 models, with a manageable number of variables and a high adjusted R^2. Similar to Graz-Mitte we choose a model (metric model) containing metric meteorological variables (**lute**) and one model (binary model) containing only binary meteorological variables (**frost**). The use of the model forecasting the PM10 concentration 2-days ahead could be more stable if the 2-days meteorological forecast isn't a metric but rather a indication by binary variables of the weather situation.
For the metric version the following model, which describes 65.5% of the variance the following variables are chosen:

pm10_lag1, lute, ltusg_k, suho, mar, dec, sat, wind, nied01

In contrast to Graz-Mitte for Graz-Süd the variable **nied01** is added to the model, because its significance is much higher than in the models of Graz-Mitte. The adjusted R^2 of the binary model for Graz-Süd is 0.6162 and the regression variables are.

pm10_lag1, invers, frost, suho, dec, nov, sat, wind, nied01

The reason for increasing the number for the models to 9 is the higher ratio of true indications in backtesting of the seasons 2014/2015 and 2015/2016.

2.2.3 Model Validation

The summary of the model calibrated on seasons 2006/2007-2013/14 shows that all variables used in the model have a significant impact. To fulfill the assumptions of a constant variance of the residuals the chosen response variable is $\sqrt{PM10}$. (similar to Graz-Mitte)

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.586907   0.105509  52.952  < 2e-16
pm10_lag1    0.024157   0.001289  18.742  < 2e-16
lute        -0.129713   0.007948 -16.320  < 2e-16
ltusg_k     -0.233297   0.015433 -15.117  < 2e-16
suho        -0.413396   0.082802  -4.993 6.84e-07
mar          0.426490   0.091538   4.659 3.53e-06
dec         -0.456805   0.082456  -5.540 3.72e-08
sat         -0.514484   0.091435  -5.627 2.29e-08
wind         0.959632   0.075629  12.689  < 2e-16
nied01      -0.298811   0.068057  -4.391 1.23e-05
```

The summary of the binary model looks similar:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.504113	0.086141	52.288	< 2e-16
pm10_lag1	0.027177	0.001315	20.668	< 2e-16
invers	0.830944	0.077865	10.672	< 2e-16
frost	0.962314	0.079097	12.166	< 2e-16
suho	-0.415890	0.087316	-4.763	2.14e-06
dec	-0.390942	0.087228	-4.482	8.11e-06
nov	-0.362952	0.088958	-4.080	4.80e-05
sat	-0.546232	0.096404	-5.666	1.83e-08
wind	1.142708	0.077809	14.686	< 2e-16
nied01	-0.383627	0.071198	-5.388	8.57e-08

In Figure 2.20 on can see some extreme outliers of the residuals in the lower as well as in the upper tail. The marked data point in the right upper area of the QQ-Plot is New years day 2007 with a PM10 concentration above 200 $\mu g/m^3$ caused by fireworks. The outlier in the lower tail is January 2^{nd} 2007, where a very high PM10 concentration is modeled. This is caused by the extremely high observation at $t-1$ (New Years day). Within the models we will not remove such high concentrations caused by external events.

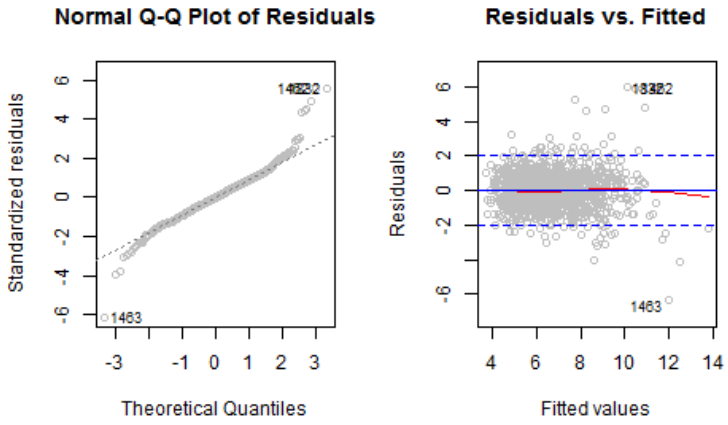

Figure 2.20: Q-Q-Plot of residuals vs. normal quantiles and Scatter-Plot of fitted values vs. residuals

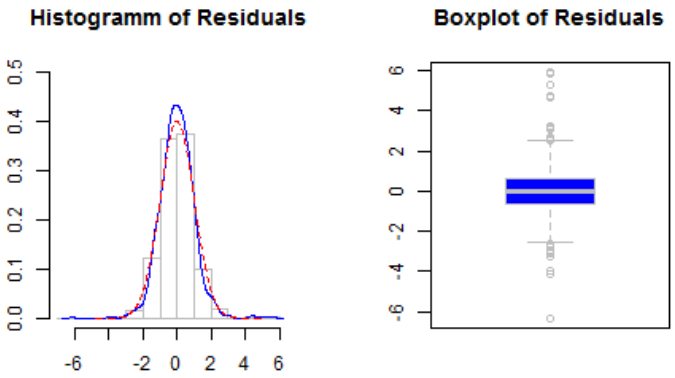

Figure 2.21: Histogram and Box-Plot of residuals

There are some outliers in the tails which leads to a rejection of the Shapiro-Wilk normality test.

```
Shapiro-Wilk normality test

W = 0.96581, p-value = 3.303e-16
```

The Shapiro Test rejects the normality assumption but the distribution of the residuals is symmetric (Figure 2.21). Using this model will lead to an underestimation of extremely high values and it overestimates low PM10 concentrations. Overestimation of low values is a more conservative approach, because they have not any adverse consequences. The standard error of the residuals for the metric model 1.071 as well the error for the binary model 1.13 is higher than for Graz-Mitte caused by the higher variability in the observations of PM10 at Graz-Süd.
The residuals (Figure 2.22 and 2.23) of the model using only binary meteorological regression variables, look very similar to the residuals of the model calibrated with metric variables. The residuals are symmetric distributed, but some outliers yield a rejection of the Shapiro-Wilk Test.

```
Shapiro-Wilk normality test

W = 0.9642, p-value < 2.2e-16
```

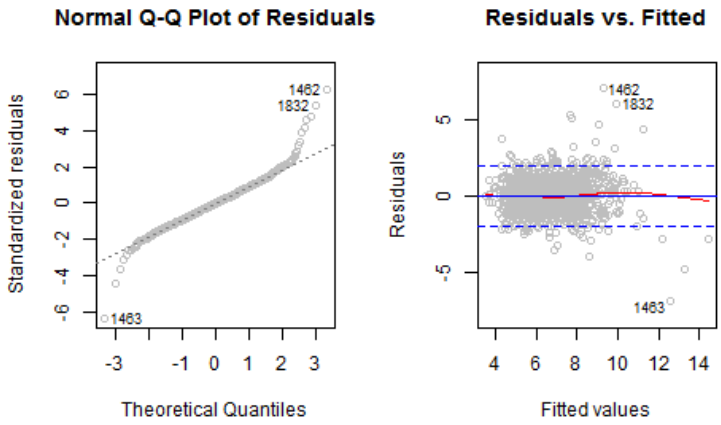

Figure 2.22: Q-Q-Plot of residuals vs. normal quantiles and Scatter-Plot of fitted values vs. residuals

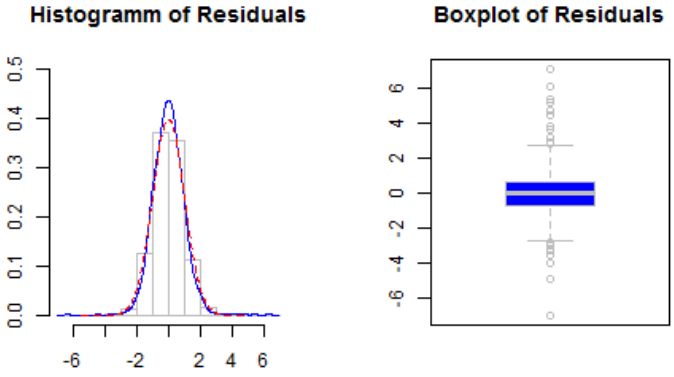

Figure 2.23: Histogram and Box-Plot of residuals

2.2.4 Quality of Forecasting

After calibrating the models and the calculation of the regression coefficients, we evaluate the quality of the models by backtesting them with the winter seasons 2014/2015 and 2015/2016.
Similar to Graz-Mitte we first evaluate the right indication according the limit of 50 $\mu g/m^3$ and thereafter we link the predicted values to a quality function. The focus is on the asymmetric quality function II defined in Section 2.1.4.

	Forecast 0-50 $\mu g/m^3$	Forecast 50-100 $\mu g/m^3$	Forecast $>$ 100 $\mu g/m^3$	Total
Observation 0-50 $\mu g/m^3$	90	9	0	99
Observation 50-100 $\mu g/m^3$	21	26	2	49
Observation $>$ 100 $\mu g/m^3$	0	3	0	3
Total	111	38	2	151

Table 2.30: Forecasting Results of the metric Model Graz-Süd, Winter Season 2014/2015

	Forecast 0-50 $\mu g/m^3$	Forecast 50-100 $\mu g/m^3$	Forecast $>$ 100 $\mu g/m^3$	Total
Observation 0-50 $\mu g/m^3$	87	12	0	99
Observation 50-100 $\mu g/m^3$	19	29	1	49
Observation $>$ 100 $\mu g/m^3$	0	3	0	3
Total	106	44	1	151

Table 2.31: Forecasting Results of the binary Model Graz-Süd, Winter Season 2014/2015

For the Season 2014/2015 both models indicate in 3 of 4 cases (around 76.8%) the right range (Table 2.30 and 2.31). But one can see that there are 21 (respectively 19) predicted values which give a too low indication of the PM10 concentration. The models indicate an observation of 0-50 $\mu g/m^3$ while the observed values are in the range of 50-100 $\mu g/m^3$. These prediction errors are critical, because they cause wrong actions for reducing the particulate matter concentration. Replacing the regression variables or increasing the number of variables in the model will not reduce this prediction error. By considering the results for the season 2015/2016 we see that the chosen models are still reasonable (Table 2.32 and 2.33). The ratio for the right indication of the range increases for the season 2015/2016 to 84.8% in the metric model resp. 81.5% in the binary model. The number of predicted values, which are higher than the observed values decreases to 7 in the metric model and to 10 in the binary model.

	Forecast 0-50 $\mu g/m^3$	Forecast 50-100 $\mu g/m^3$	Forecast $>$ 100 $\mu g/m^3$	Total
Observation 0-50 $\mu g/m^3$	82	14	0	96
Observation 50-100 $\mu g/m^3$	5	47	2	54
Observation $>$ 100 $\mu g/m^3$	0	2	0	2
Total	87	63	2	152

Table 2.32: Forecasting Results of the metric Model Graz-Süd, Winter Season 2015/16

	Forecast 0-50 $\mu g/m^3$	Forecast 50-100 $\mu g/m^3$	Forecast $>$ 100 $\mu g/m^3$	Total
Observation 0-50 $\mu g/m^3$	79	17	0	96
Observation 50-100 $\mu g/m^3$	8	45	1	54
Observation $>$ 100 $\mu g/m^3$	0	2	0	2
Total	87	64	1	152

Table 2.33: Forecasting Results of the binary Model Graz-Süd, Winter Season 2015/2016

In the Tables 2.34 and 2.35 the categorization according quality function II of the backtesting results is represented. The results reflect the results shown in Tables 2.30 - 2.33.

			Model with binary variables					Total
			”excellent”	”good”	”satisfying”	”bad”	”very bad”	
Model with	metric variables	”excellent”	74	18	1	1	1	95
		”good”	8	7	1	0	1	17
		”satisfying”	2	2	5	4	0	13
		”bad”	0	0	0	1	1	2
		”very bad”	2	1	1	0	20	24
		Total	86	28	8	6	23	151

Table 2.34: Results of Quality function II for Winter Season 2014/2015 Graz-Süd

			Model with binary variables					Total
			”excellent”	”good”	”satisfying”	”bad”	”very bad”	
Model with	metric variables	”excellent”	77	13	3	0	2	95
		”good”	17	6	4	2	2	31
		”satisfying”	3	3	4	2	1	13
		”bad”	1	0	0	3	1	5
		”very bad”	2	0	1	0	5	8
		Total	100	22	12	7	11	152

Table 2.35: Results of Quality function II for Winter Season 2015/2016 Graz-Süd

2.2.5 Possible further Improvements of the Model

Similar to the improvement of the model for Graz-Mitte, we try to get a better calibration of the model by reducing the number of seasons. The regression factors of the current model are calibrated using all available seasons (2006/2007-2013/2014). By reducing the number of seasons, the first seasons with high PM10 concentrations are not considered. This step might improve the fit to the data and could yield to a higher indication ratio in backtesting. In contrast to Graz-Mitte there is no relocation of the measurement station at Graz-Süd. The

iterative reduction of the number of seasons and the calculation of the linear model yields the following result (Table 2.36 and 2.36).

Seasons	Number of Seasons	Adjusted R^2
2006/2007-2013/2014	8	0.6550
2007/2008-2013/2014	7	0.6637
2008/2009-2013/2014	6	0.6620
2010/2011-2013/2014	5	0.6276
2011/2012-2013/2014	4	0.6248

Table 2.36: Adjusted R^2 of Metric Model for Graz-Süd

Seasons	Number of Seasons	Adjusted R^2
2006/2007-2013/2014	8	0.6161
2007/2008-2013/2014	7	0.6319
2008/2009-2013/2014	6	0.6326
2010/2011-2013/2014	5	0.5979
2011/2012-2013/2014	4	0.5847

Table 2.37: Adjusted R^2 of Binary Model for Graz-Süd

The reduction of the number of seasons from 8 to 6 yields to a slighty higher adjusted R^2 for both models. The estimated regression parameters of the metric model are as follows:

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.438474   0.118080  46.058  < 2e-16
pm10_lag1    0.026285   0.001543  17.033  < 2e-16
lute        -0.124257   0.008340 -14.899  < 2e-16
ltusg_k     -0.199084   0.016740 -11.892  < 2e-16
suho        -0.442630   0.089397  -4.951 8.81e-07
mar          0.457142   0.099113   4.612 4.56e-06
dec         -0.519810   0.088656  -5.863 6.38e-09
sat         -0.570695   0.098352  -5.803 9.06e-09
wind         0.959116   0.081311  11.796  < 2e-16
nied01      -0.314765   0.071923  -4.376 1.35e-05
```

The binary model is given by:

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.451579   0.095515  46.606  < 2e-16
pm10_lag1    0.029129   0.001557  18.708  < 2e-16
invers       0.754579   0.082232   9.176  < 2e-16
frost        0.920970   0.081316  11.326  < 2e-16
```

```
suho         -0.447454   0.093136  -4.804 1.82e-06
dec          -0.479646   0.093054  -5.155 3.13e-07
nov          -0.332061   0.094698  -3.507 0.000477
sat          -0.611417   0.102402  -5.971 3.40e-09
wind          1.104226   0.082548  13.377  < 2e-16
nied01       -0.391067   0.074034  -5.282 1.60e-07
```

All variables are significant within the models. Similar to the model calibrated to 8 seasons the QQ-Plot (Fig: 2.24) indicates some outliers in the lower and upper tail. Similar to the previous model the outliers are New Years days with high PM10 concentrations caused by fireworks. The Shapiro normality test for the metric model is rejected, but considering the histogram and the boxplot (Figure 2.25) a symmetric distribution of the residuals is given.

```
Shapiro-Wilk normality test

W = 0.98137, p-value = 2.393e-09
```

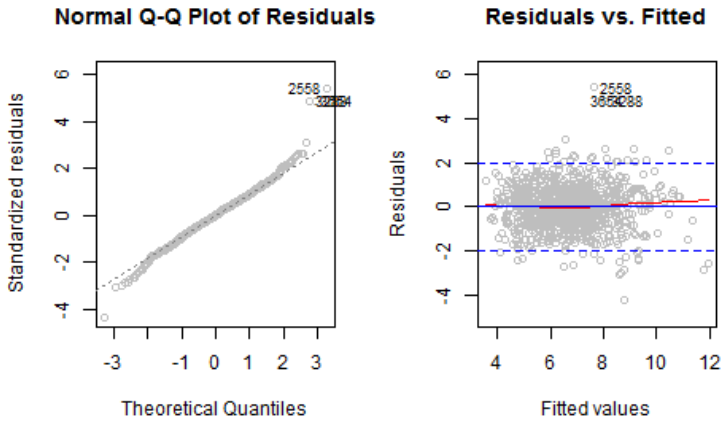

Figure 2.24: Q-Q-Plot of residuals vs. normal quantiles and Scatter-Plot of fitted values vs. residuals

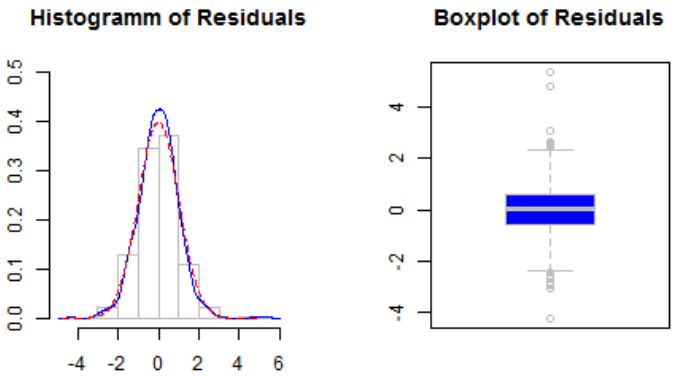

Figure 2.25: Histogram and Box-Plot of residuals

The residuals of the binary model (Figures: 2.26 and 2.27) look similar to the residuals of the metric model. The Shapiro Wilk test is rejected, but the distribution of the residuals is symmetric and similar to a standard normal distribution (Histogram 2.26) and the standard error is simliar to the model for Graz-Süd seen above.

```
Shapiro-Wilk normality test

W = 0.98477, p-value = 4.219e-08
```

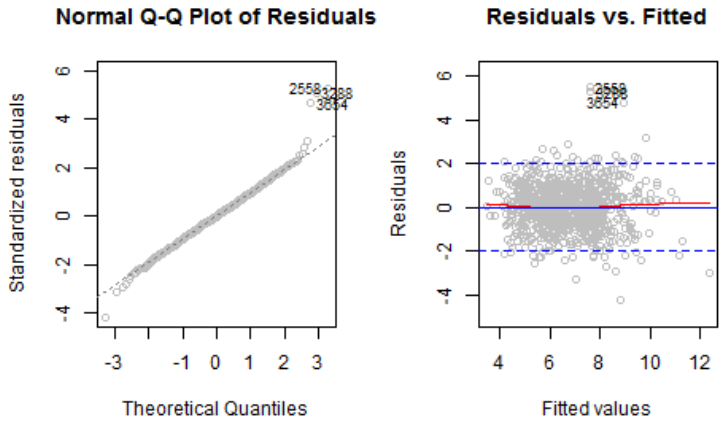

Figure 2.26: Q-Q-Plot of residuals vs. normal quantiles and Scatter-Plot of fitted values vs. residuals

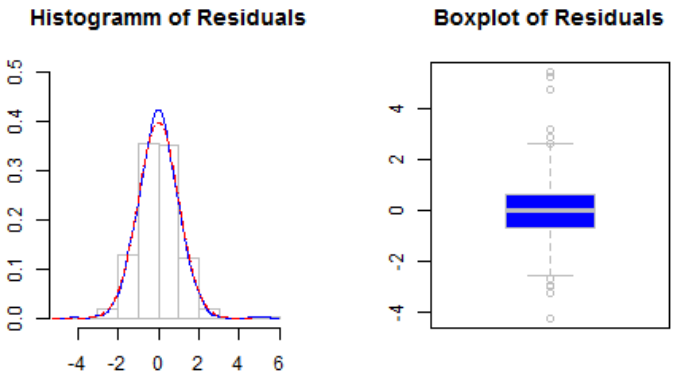

Figure 2.27: Histogram and Box-Plot of residuals

The increase of the described variance of the model by decreasing the number of seasons, does not yield to better results in the backtesting of the seasons 2014/2015 and 2015/2016. The results of the backtesting, illustrated in the Tables 2.38-2.40 show an insignificant increase of the right indication according the 50 $\mu g/m^3$ limit for the metric model and a slight decrease for the binary model.

	Forecast 0-50 $\mu g/m^3$	Forecast 50-100 $\mu g/m^3$	Forecast $>$ 100 $\mu g/m^3$	Total
Observation 0-50 $\mu g/m^3$	92	7	0	99
Observation 50-100 $\mu g/m^3$	20	28	1	49
Observation $>$ 100 $\mu g/m^3$	0	3	0	3
Total	112	38	1	151

Table 2.38: Forecasting Results of the metric Model Graz-Süd based on 6 winter seasons, Winter Season 2014/2015

	Forecast 0-50 $\mu g/m^3$	Forecast 50-100 $\mu g/m^3$	Forecast $>$ 100 $\mu g/m^3$	Total
Observation 0-50 $\mu g/m^3$	87	12	0	99
Observation 50-100 $\mu g/m^3$	21	27	1	49
Observation $>$ 100 $\mu g/m^3$	0	3	0	3
Total	108	42	1	151

Table 2.39: Forecasting Results of the binary Model Graz-Süd based on 6 winter seasons, Winter Season 2014/2015

	Forecast 0-50 $\mu g/m^3$	Forecast 50-100 $\mu g/m^3$	Forecast $>$ 100 $\mu g/m^3$	Total
Observation 0-50 $\mu g/m^3$	83	13	0	96
Observation 50-100 $\mu g/m^3$	11	42	1	54
Observation $>$ 100 $\mu g/m^3$	0	2	0	2
Total	94	57	1	152

Table 2.40: Forecasting Results of the binary Model Graz-Süd based on 6 winter seasons, Winter Season 2015/2016

	Forecast 0-50 $\mu g/m^3$	Forecast 50-100 $\mu g/m^3$	Forecast $>$ 100 $\mu g/m^3$	Total
Observation 0-50 $\mu g/m^3$	84	12	0	96
Observation 50-100 $\mu g/m^3$	8	45	1	54
Observation $>$ 100 $\mu g/m^3$	0	2	0	2
Total	92	59	1	152

Table 2.41: Forecasting Results of the metric Model Graz-Süd based on 6 winter seasons, Winter Season 2015/16

2.2.6 Backtesting under more realistic assumptions

Reducing the number of seasons for the calibration of the models yields an immaterial increase of the adjusted R^2, but no essential enhancement of the right indication can be observed. In this section we are interested in the stability of the models under more realistic circumstances. Similar to the procedure described in Section 2.1.6 the predictions are not based on observed values but on predicted values. For the prediction the observed **pm10_lag1** is replaced by the one day prediction value of the current model specified in Stadlober et al. [23]. In addition we add a random noise to the meteorological values in the model to simulate the meteorological prediction error. For the binary meteorological variables in the model we assume that 35% are wrongly predicted.
The results for the indication according the limit of 50 $\mu g/m^3$ are given in the Tables 2.42 - 2.45. We see that the number of predictions in the range 0-50 $\mu g/m^3$ increases and the number the underestimated observations too. Especially the underestimation of observations in the range 50-100 $\mu g/m^3$ rises by approximately 50% in the season 2014/2015. The ratio of right predictions decreases from 76.8% to 69.5% for the metric model in the season 2014/2015 and from 76.8% to 67.5% for the binary model. For the season 2015/2016 the number of underestimated observations bigger than 50 $\mu g/m^3$ and smaller than 100 $\mu g/m^3$ doubled. The ratio of right predictions decreased by 15% for both models to 69.7% for the metric model and 67.1% for the binary model.

	Forecast 0-50 $\mu g/m^3$	Forecast 50-100 $\mu g/m^3$	Forecast >100 $\mu g/m^3$	Total
Observation 0-50 $\mu g/m^3$	92	7	0	99
Observation 50-100 $\mu g/m^3$	34	13	2	49
Observation >100 $\mu g/m^3$	0	3	0	3
Total	126	23	2	151

Table 2.42: Forecasting Results of the metric Model Graz-Süd based on 6 winter seasons under more realistic assumptions, Winter Season 2014/2015

	Forecast 0-50 $\mu g/m^3$	Forecast 50-100 $\mu g/m^3$	Forecast >100 $\mu g/m^3$	Total
Observation 0-50 $\mu g/m^3$	84	15	0	99
Observation 50-100 $\mu g/m^3$	31	18	0	49
Observation >100 $\mu g/m^3$	2	1	0	3
Total	127	34	0	151

Table 2.43: Forecasting Results of the binary Model Graz-Süd based on 6 winter seasons under more realistic assumptions, Winter Season 2014/2015

	Forecast 0-50 $\mu g/m^3$	Forecast 50-100 $\mu g/m^3$	Forecast >100 $\mu g/m^3$	Total
Observation 0-50 $\mu g/m^3$	69	26	1	96
Observation 50-100 $\mu g/m^3$	18	33	3	54
Observation >100 $\mu g/m^3$	0	2	0	2
Total	87	61	4	152

Table 2.44: Forecasting Results of the metric Model Graz-Süd based on 6 winter seasons under more realistic assumptions, Winter Season 2015/16

	Forecast 0-50 $\mu g/m^3$	Forecast 50-100 $\mu g/m^3$	Forecast >100 $\mu g/m^3$	Total
Observation 0-50 $\mu g/m^3$	61	35	0	96
Observation 50-100 $\mu g/m^3$	19	35	0	54
Observation >100 $\mu g/m^3$	0	2	0	2
Total	80	72	0	152

Table 2.45: Forecasting Results of the binary Model Graz-Süd based on 6 winter seasons under more realistic assumptions, Winter Season 2015/2016

Applying the asymmetric quality function to the predictions we see in Table 2.46 and Table 2.47 a sharply increase of ”very bad” forecasts. More than one fifth of the forecasted

values doesn't give any considerable fit. In addition only two third of the forecasted values are marked as "excellent" or "good" by using the quality function. Despite the decrease of the prediction quality of the model, it may help to get a right indication of the PM10 concentration two days ahead.
By comparing the observed PM10 concentration with the forecasted values for the winter season 2014-2015 in Figure 2.28 we see that the models underestimate much more observations than for Graz-Mitte. For the winter season 2015/2016 the results looks very simliar. (Figure 2.29).

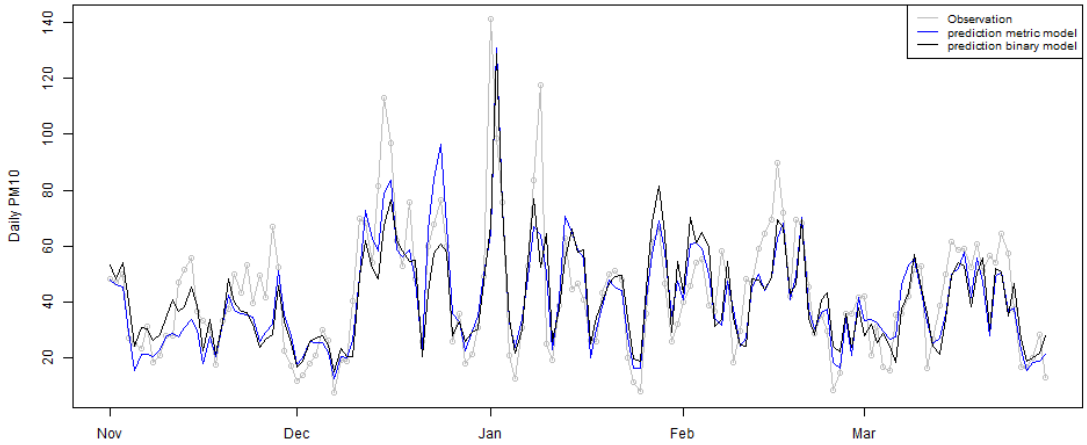

Figure 2.28: Comparison of observed and predicted values for the winter season 2014/2015

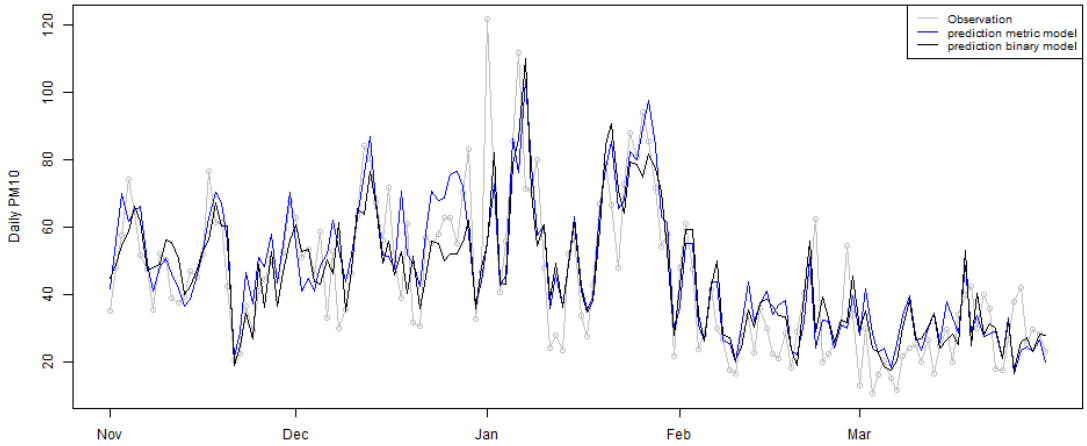

Figure 2.29: Comparison of observed and predicted values for the winter season 2015/2016

In the Figures 2.30 and 2.31 one can see in the left upper corner and in the right lower corner a Scatter-Plot of the observations and the predicted value of the according model combined with the quality function. We see that especially for extreme observations the quality of the fit is bad. In the right upper plot on can see that both models predict similar values. The mass of the dots is close to the diagonal line.

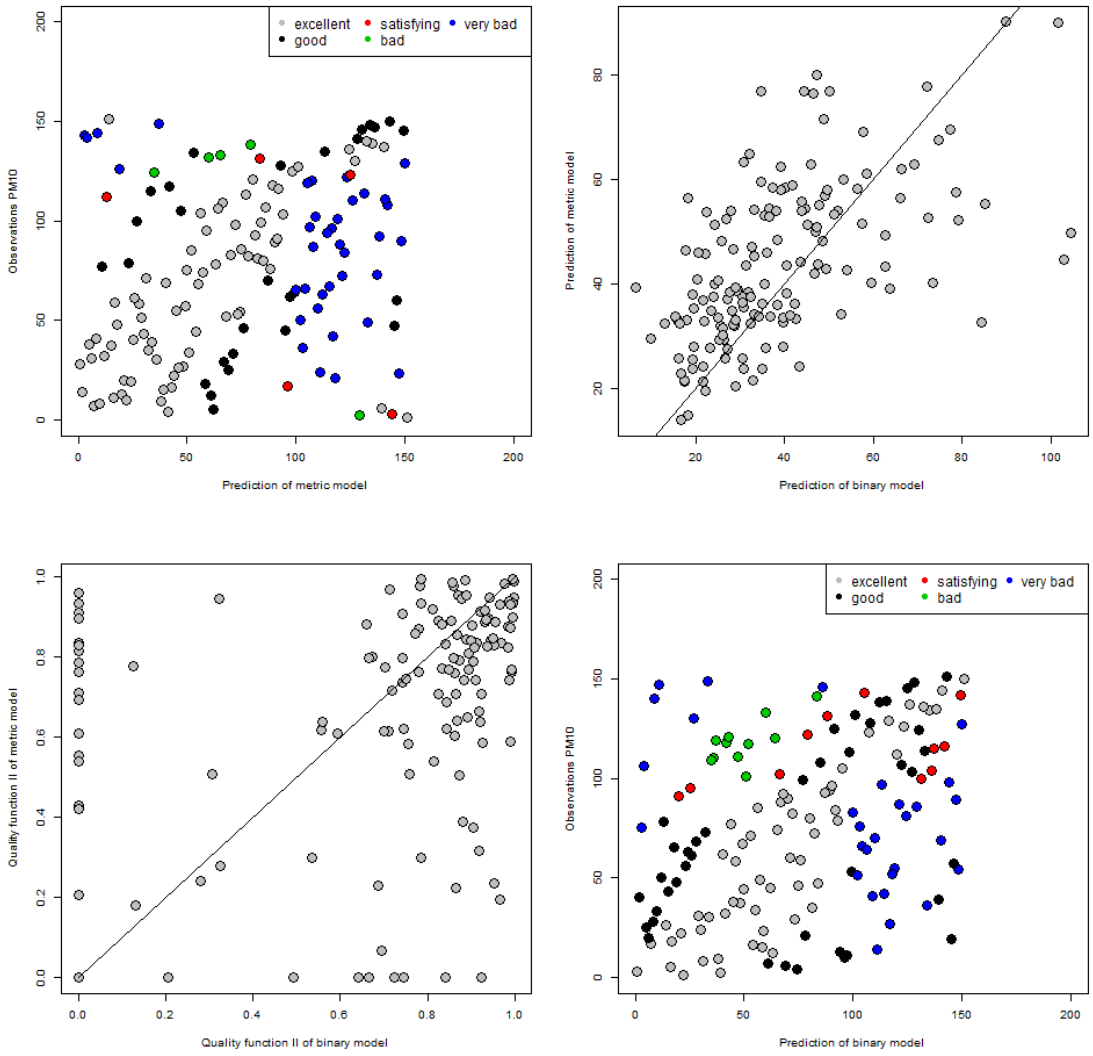

Figure 2.30: Analysis of the Forecasts for the winter season 2014/2015

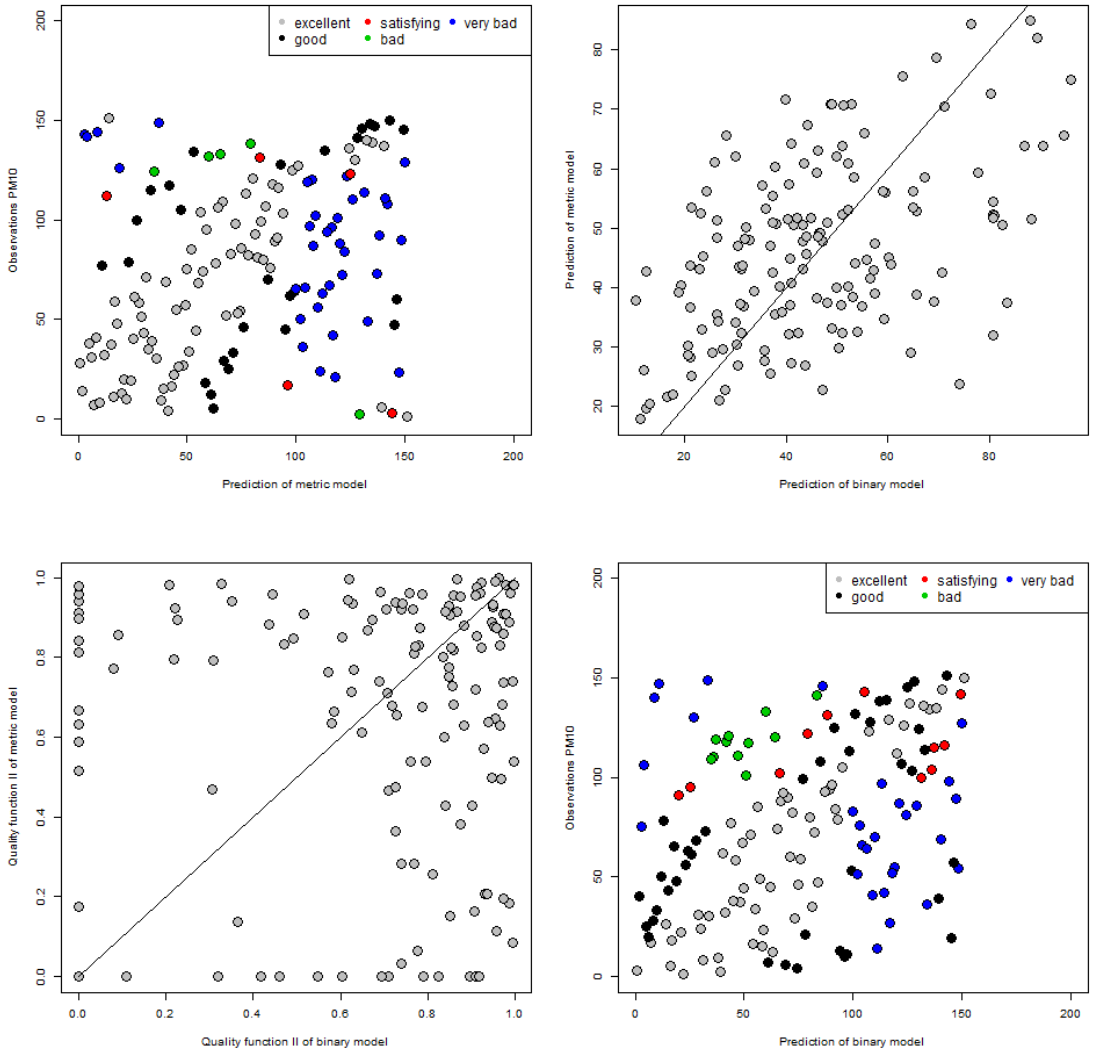

Figure 2.31: Analysis of the Forecasts for the winter season 2015/2016

			Model with binary variables					Total
			”excellent”	”good”	”satisfying”	”bad”	”very bad”	
Model with	metric variables	”excellent”	34	26	1	6	4	71
		”good”	12	10	5	1	4	32
		”satisfying”	2	2	0	1	1	6
		”bad”	0	1	1	1	0	3
		”very bad”	3	4	5	0	27	39
		Total	51	43	12	9	36	151

Table 2.46: Results of Quality function II for Winter Season 2014/2015 Graz-Süd

			Model with binary variables					Total
			”excellent”	”good”	”satisfying”	”bad”	”very bad”	
Model with	metric variables	”excellent”	26	10	8	4	5	53
		”good”	11	10	5	3	10	39
		”satisfying”	5	3	2	0	6	16
		”bad”	3	3	0	0	3	9
		”very bad”	10	8	5	0	12	35
		Total	55	34	20	7	36	152

Table 2.47: Results of Quality function II for Winter Season 2015/2016 Graz-Süd

A model for Graz-Süd which does not take any meteorological variables into consideration, does not yield valuable results. The quality of the forecast and the stability of the model are not valuable.

Chapter 3

Prediction Model of PM 10 based on classification trees

Whitin this chapter we introduce models for the PM10 prediction based on classification and regression trees for the locations Graz-Mitte and Graz-Süd. In contrast to the models considered in chapter 2 the tree models should give an easy and fast indication for an exceedance 1 day ahead. Before applying tree models to the PM10 data set we indroduce them in the following sections. The implementations in R are performed by using the separate R package **rpart** (recursive partitioning) and **party**, which are one of the many packages in R for implementing tree calculations.

3.1 Introduction to classification and regression trees

Classification and regression trees (CART) are a non parametric method, using a set of p predictor variables $x_1, \ldots, x_p$ to predict and model a categorical or numerical response variable y. The goal is to find a model which predicts the response variable best. The idea of CART models is a recursive partitioning of the data set, which is a stagewise process dividing the data set into smaller and smaller pieces. The idea is to split the data in that way that in each of the descendent subsets the data are purer than in the parent set. The response variable y can be categorical or numerical. For categorical response variables the method produces a classification tree and for continuous variables a regression tree.
CART is a very popular data analysis tool which were introduced in 1984 by Breiman, Friedman, Ohlsen and Stone [5]. In the last decades some refinements of CART and some similar procedures, having the roots in computer science, were developed (e.g. C4.5, GUIDE). In this chapter our discussions follow Berk [1] and Breiman et al. [5].

3.1.1 Construction of a classification tree

A tree model splits the space $\mathcal{X}$ spanned by the predictor variables $x_1, \ldots, x_p$ stagewise into smaller subspaces. The tree is constructed by repeated recursive partitioning the set of data into two descendant subsets, beginning with the space spanned by $x_1, \ldots, x_p$. The process of splitting into subsets (subspaces) performs one partition by a prediction variable x a time, which makes the models easier to interpret. A constructed partition is not reconsidered or revisited in later splitting steps. Each subset is constructed so that the values of the response variable are purer or more homogenous than in the previous subset. This procedure is repeated to each subset separately until a stopping criteria is met or the minimum size of the subset is reached. The terminal subsets form a partition of the space $\mathcal{X}$. Each terminal subset is assigned by a prediction value, which is a class label for classification tree or a numeric number for regression trees. Terminal subsets can be marked with the same class label.
At the beginning we consider the construction of a classification tree with a categorical response variable with 2 classes (e.g. $y \in \{0, 1\}$). In the following the class labels will represent the exceedance and non-exceedance of the critical limit of 50 μ/m^3.
The process of building a tree starts by analyzing the whole data set $\mathcal{X}$. The procedure looks for the best split of the whole data set into two subsets $\mathcal{X}_1$ and $\mathcal{X}_2$ so that each subset is more homogenous than the initial subset. Dividing in 2 subsets means that the procedure finds a threshold in the values of one predictor variable x and group the response variables into 2 subsets according the threshold. In a next step the splitting of the subsets $\mathcal{X}_1$ and $\mathcal{X}_2$ by other thresholds of any predictor variable can be performed. The splitting of the subset $\mathcal{X}_1$ is independent of the splitting of the subset $\mathcal{X}_2$. The same process is applied to all subsequent subsets until all cases are placed in a terminal subset. Each case can only be in one terminal node, because the partitions do not overlap. The CART method performs the division into subsets recursively. The representation of a CART is often shown as an inverted tree (tree diagram) see Figure 3.2. The whole data set $\mathcal{X}$ is contained in the root nodes.
The remaining topics and questions for the complete understanding of the construction of a tree model as follows:

1. What criteria will be used to determine the partitions of subsets?
2. Which criteria for the variable selection is required and how will the variable used to define the new partition.
3. When should we label a node as terminal or continue in splitting it?
4. How to measure the goodness of a tree model.

3.1.2 Splitting criteria

An essential topic for tree models is how to split each node by using the information of the predictor values. For a continuous variable with n distinct values, there are $n - 1$ possible splits on that variable which need to be evaluated. For categorical predictor variables with n

categories $2^{n-1} - 1$ possible splits need to be calculated. CART estimates at the root node all possible splits of all predictor variables and picks the best single split. To define the best split of a subset we focus on the *"impurity"* of a node. The first 3 questions above are linked to this definition. The goal of partitioning the data is to receive resulting subsets which are as *pure* (homogeneous) as possible. Further to have little impurity overall. Consider a binary response variable, then a node has an impurity of zero if all response values are 0 or 1. The maximum of impurity will be reached if the values are equally mixed, 50% of 0 values and 50% of 1 values.

Definition 3.1 (Impurity Function) *The impurity of a node A is a non-negative function ϕ of the probability $P(Y = 1|A)$, with Y a random variable.*

$$I(A) = \phi(P(Y = 1|A)), \tag{3.1}$$

with $\phi \geq 0$, $\phi(P) = \phi(1 - P)$ and $\phi(0) = \phi(1) < \phi(P)$.

The impurity is a non-negative and symmetrical function with a minimum when the node A is pure (containing only 0s or 1s) and a maximum when all classes are equally mixed. In the following we consider 2 different impurity functions which are also discussed in Breiman et al. [5]

Definition 3.2 (Gini Index) *The Gini Index impurity function is given by*

$$\phi(P) = P(1 - P), \tag{3.2}$$

Definition 3.3 (Cross Entropy) *The Cross Entropy impurity function is given by*

$$\phi(P) = Plog(P) - (1 - P)log(1 - P), \tag{3.3}$$

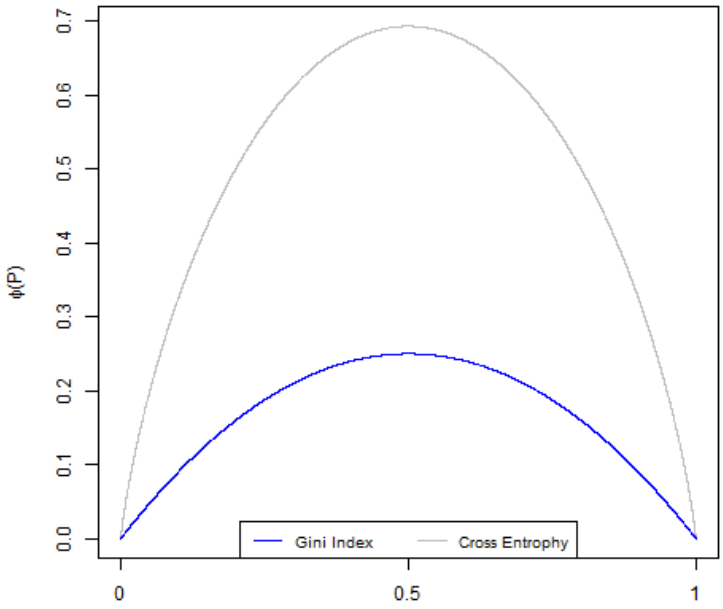

Figure 3.1: Impurity Functions

Both impurity functions fulfill the required properties. The Gini index is generally favored, because Breiman et al. [5] showed that it delivers more desirable results than other impurity functions.
Finding a value in a predictor variable which splits the parent partition so that the resulting partitions have an overall smaller impurity is the goal of the described setup. The CART procedure evaluates all possible splits of all prediction variables and tries to find a value to split a node, which reduces the impurity most. The improvement of the impurity resulting from a split s performed on a node A is the impurity of the parent node A minus the impurity of the left A_L and right A_R daughter node. The decrease of the impurity is given by

$$\Delta I(s, A) = I(A) - P(A_L)I(A_L) - P(A_R)I(A_R)$$

where $I(A)$ is the initial impurity and $P(A_L)$ is the probability of a case falling into the left node and $P(A_R)$ into the right node. The probabilities can be estimated by using the marginal proportions and $P(A_L) + P(A_R) = 1$. The decrease of the impurity is maximized

$$\max_{s,A} \Delta I(s, A).$$

The CART procedure chooses the biggest decrease in impurity to define the new partition. The algorithm will stop if a node is reached where no significant decrease of impurity is possible. A node is declared as terminal for a given threshold , $\beta > 0$, if

$$max_{s,A} \Delta I(s, A) < \beta.$$

As already mentioned another stopping criteria is a minimum count of observations in each terminal node or a maximum count of splits in the tree.
After constructing a tree with a set of terminal nodes, a class to every terminal node is assigned. The assignment of classes is determined by the majority of cases of a class i in the terminal node A. Then the node is labled with class i.
This procedure points the strength of the CART procedure out. After assigning a class to each terminal node the constructed tree makes an estimation of a possible outcome of new observations with unknown response variable very easy. In addition the tree gives us information under which constellation and interactions between variables observations are assigned to a specific class.

3.1.2.1 Model Evaluation

After fitting a tree, in each terminal node the proportion of right classification and failures are calculated. For a given data set one is interested in the quality of the fitted tree and its performance in classifying given observations correctly. Beside the resulting model error the quality in backtesting of data which are not used for growing the tree might be essential. One helpful method for representing the model error of a CART model for data is a *confusion table*. It cross-tabulates the observed classes and the classes assigned by the CART model. The table gives an indication how well a model fits the data which were used to build the tree. In addition a confusion table can be used for evaluating the strength of the model for forecasting purposes of that which were not used to build the model (test data).
In Tabel 3.1 the general structure of a confusion table fur a binary classification model is shown. The letters in the table 3.1 is the number of observations falling in the accord-

	Failure Prediction	Success Prediction	Model Error
Failure	a	b	$b/(a+b)$
Success	c	d	$c/(c+d)$
Use Error	$\frac{c}{a+c}$	$\frac{b}{b+d}$	Overall Error $= \frac{b+c}{a+b+c+d}$

Table 3.1: General structure of a Confusion Table

ing class. The left upper cell represents the observations where the observed and predicted (assigned) value is "failure". The table gives an overview of the overall prediction error $(b+c)/(a+b+c+d)$ as the sum of misclassified cases divided by the total number of observations. This proportion gives an initial quality of the fit. A tree with pure terminal nodes results a table with no misclassified cases. A low misclassification proportion is desirable, but must be compared with the classification of the model where no predictor is used.
The overall error neglects the different importance of the prediction variables. In some cases the accuracy in one variable is more important than in another. For this purpose the table gives an estimation of the the incorrectly classified observations of a class divided by the total amount of observations of a class. The classification error caused by the CART model

(*model error*) is shown in the last column of the table. There are two kinds of failures in the model called *false negatives* b and *false positives* c. They give information on how common it is for this procedure to be wrong. The column proportion shows the so called *use error* of the model. The proportion gives information of times when a particular class is assigned how common it is that the assignment is wrong.
The ratio of the number of false negatives and true positives gives an indication how the results are trading one kind of error for the other. If the proportion of false negatives is x-times larger than false positives means that the model treats false negatives x-times more important than false positives. Such ratios are important for our further interpretations of the results.

3.1.3 Misclassification Costs and Prior Probabilities

In the previous sections the terminal nodes were assessed to each class by majority vote. All cases in a terminal node are classified with the label of the class with the greatest count of cases in the terminal node. Consider a terminal node were the proportion of different classes is nearly equal weighted. Within this terminal node a huge proportion of false negatives or false positives is wrong classified. Such wrong classifications can lead to enormous costs in forecasting and prediction using a modeled tree. Thinking of PM10 forecasting, the wrong prediction of no exceedance of the critical value of 50 μ/m^3 might produce higher costs than the false prediction of an exceedance. Considering costs will be essential for a classification algorithm because the way cases are forecasted or classified will vary depending on the costs introduced. The costs that matters are the costs of classification errors, which will change the results dramatically. Therefore we introduce costs to the CART method, where prior probabilities and loss functions play a key role.
Each categorical response variable has a proportion of observations in each response category. The term "prior" has its origin in Bayesian statistics, where prior refers to the expectations of the data analyst. The beliefs from the data analyst come from past research or other information about what the marginal proportions should be. With the help of prior probabilities the actual proportion of the response variable, which is not reflected in the sample, can be reproduced.

Definition 3.4 (Prior Probabilities) *There are N observations and C classes of the response variable and K terminal nodes. We define π_i, $i = 1, 2, \ldots, C$ as the prior probability of being in class i. For the binary case i would be 1 or 2. These marginal probabilities are sometimes called as prior probabilities.*

Definition 3.5 (Loss Matix) *The loss matrix for the incorrect classification of a case that is really an i as a j is defined by $L(i, j)$. For the binary case the matrix is a 2×2 matrix and captures the cost of classification errors. For correct classification in general the loss is equal to zero which means that the diagonal of the loss matrix is zero.*

We define $\tau(x)$ the true class for an observation x. where x represents the vector of predictor variable values for that observations. A denotes a node in the tree. $\tau(A)$ is the class assigned to node A if A is a terminal node. The number of observations in the sample that are in class i is N_i and N_A represents the number of observations in node A. The number of observations of class i in node A is given by N_{iA}.
The probability of cases appearing in node A is given by $P(A)$ and is equivalent to $\sum_{i=1}^{C} \pi_i P[x \in A|\tau(x) = i]$. The estimation of the probability is given by $\sum_{i=1}^{C} \pi_i (N_{iA}/N_i)$. On can see that the priors can influence the tree structure. The probability of class i given that a case is in node A is denoted by $p(i|A)$ or $P[\tau(x) = i|x \in A]$. The probability equals $\pi_i P[x \in A|\tau(x) = i]/P[x \in A]$ and can be estimated by the number of cases of class i in node A, divided by the total number of cases in that node $\pi_i(N_{iA}/N_i)/\sum_{i=1}^{C} \pi_i(N_{iA}/N_i)$. Here the priors could lead to big differences because the probability of a case with true class i landing in A depends in part on the probability that a case is truely of class i to begin with.

Definition 3.6 (Risk of node A) *The "risk" of a node A is $R(A)$ given by $\sum_{i=1}^{C} p(i|A)L(i, \tau(A))$. $\tau(A)$ is chosen in such a way that the risk is minimized. The risk of a node is a function of the probabilities and the costs.*

Definition 3.7 (Risk of a tree) *The "risk" of a tree T is $R(T)$ which equals $\sum_{j=1}^{K} P(A_j)R(A_j)$, where A_j are the terminal nodes of the tree. The risk of the tree is the sum of the risk associated with each node, weighted by the probability of the cases falling in that node.*

Considering a loss matrix $L(i, j) = 1$ for all $i \neq j$ and the priors are the observed class proportions in the sample, then $p(i|A) = N_{iA}/N_A$ and the riks of the tree $R(T)$ is the proportion of misclassified cases. In an analog way $R(A)$, the risk of a terminal node can be estimated. Scaling up of the loss matrix $L(i, j)$ with $L(i, j) = m$, with m constant, makes no difference to the CART algorithm. If there is no adjustment of the priors the data determines everything which is the same as the costs of classification errors in the loss function is the same and one assumes that the given marginal distribution of the sample are the appropriate priors.
Adjusting the cost in the loss functions leads to different participation and classifications. This method produces a more acceptable ratio of false positives and negatives. The risk of a node is scaled by prior probabilities and the entries of the loss matrix. Suppose there exists a $\tilde{\pi}$ and $\tilde{L}(i, j)$ so that the following holds

$$\tilde{\pi}_i \tilde{L}(i, j) = \pi_i L(i, j).$$

The associated risk of a node stays unchanged as long as the equation holds and it does not matter what the particular values of $\tilde{\pi}$ and $\tilde{L}$ are. The right hand side of the equation represents the weight given to the classification errors for class i in a node. If a change in the weight is desired the prior or the costs or both can be changed. For binary response variables

this method called "altered priors" can be given by

$$\tilde{\pi}_i^* = \frac{\pi_i L_i^*}{\pi_i L_i^* + \pi_j L_j^*}.$$

The values of π_i are the marginal probabilities associated with the empirical prior distribution. L_i^* are the new costs of the loss matrix. All costs in the loss matrix are relative costs. For binary problems the altered prior method is exact and affects the choice of splits. Thinking of an impurity function, which has its maximum if the classes are equal weighted, the altered priors shift the maximum of the impurity function. This method is implemented in the R package **rpart**.

3.1.4 Getting the right sized Tree (Pruning)

In this section we consider the topic of getting the right sized tree. For many analyses a too large and complex tree is not desirable. The stepwise construction of a tree optimizes at each step the impurity over all possible splits of the data. This procedure may lead to an overfitting of the data set, which reduces the accuracy of prediction and the interoperability of the tree. The increase of splits yields a decrease of the misclassification of the tree. For example a tree where each terminal node contains only one data point, the node is classified by the case it contains and the tree error is zero. A possible strategy to avoid such tree, is the setting of a minimum sample size for each terminal node. Another widely used method is called "pruning", which removes undesired branches. Splits of the data sets which do not reduce the heterogeneity sufficiently for the additional complexity are removed. It can be compared with a stepwise linear regression, where additional prediction variables increase R^2 but may have no predictive power. Therefore the consideration of the adjusted R^2 is desirable.
Recall the risk for a tree $R(T) = \sum_{j=1}^{K} P(A_j)R(A_j)$ which is equal to the sum of the terminal nodes of the risk associated with each node times the probability of cases falling in that node. The pruning process starts ideally with a saturated tree. The minimization of the risk of the tree T would not lead to a smaller tree, because the reduction of the risk of a tree is equal to that all terminal nodes are homogeneous and the risk would be zero. To solve the problem of the right sized tree the introduction of a penalty for complexity parameters must be introduced. It will be a tradeoff between bias and variance. Larger trees will have terminal nodes with fewer classification erros, implying less bias. On the other hand a larger tree implies greater instability because there are fewer cases in the terminal nodes. The solution of this problem is given by the minimal cost complexity pruning of trees, which includes a parameter which penalizes the complexity of a tree.

Definition 3.8 (Cost-Complexity Measure) *For a tree T and $\alpha \geq 0$ the cost complexity measure is given by*

$$R_\alpha(T) = R(T) + \alpha|\tilde{T}|.$$

With $|\tilde{T}|$ the number of terminal nodes in tree T.

The cost complexity measure is given by the cost of the classification errors for the tree as a whole and the penalty for complexity. The penalty for each additional terminal node is quantified by α, larger α leads to a heavier penalty for complexity. Breiman et al. [5] proved the following theorem.

Theorem 3.1 *For every value of* α, *there exists a smallest minimizing subtree* $T(\alpha)$

$$R_\alpha(T(\alpha)) = \min R_\alpha(T)$$

The existence of a unique smallest subtree for the cost complexity parameter α implies that there cannot be two subtrees of the same size with the same α.

3.1.5 Missing Data and surrogate variables

Missing values create for nearly all statistical analyses and data algorithms the same difficulties. Missing data sets are very common and lead to a loss of statistical power with the reduction of the sample by ignoring incomplete observations. This is a common mechanism to drop data with missing prediction variables. By missing the data completely at random the main loss is statistical power. If the number of observations with missing response values compared to the total data set is below a given percentage ignoring those cases is a valuable method. Another possibility is to input the data outside of the CART itself. Using a conventional regression of the response with a strongly related predictor, the resulting regression equation can input the missing values.
An additional option to address the problem of missing data within the CART algorithm was introduced by Breiman et al. [5]. The following approach is also available within the R package **rpart**. Recall the selection of a split which will be chosen. The following quantity will be maximized:

$$\Delta I(s, A) = I(A) - P(A_L)I(A_L) - P(A_R)I(A_R),$$

where $I(A)$ is the value of the parent impurity. This term can be calculated without any predictors and is irrespective of missing data. For the calculation of the impurity of the daughter nodes and their construction predictor variables are required. The impurity is evaluated as usual for each optimal split by taking only data sets available. The associated probabilities $P(A_L)$ and $P(A_R)$ are also calculated only over the relevant observations. After the determination of the split the observations have to be assigned to one of the two daughter nodes. After the determination of the split the CART algorithm adjust $P(A_L)$ and $P(A_R)$ so that they sum to $P(A)$. To circumvent the assignment of observations where the predictor values are missing a possible approach is to estimate the missing data by using other independent variables. A variation of this approach is used by ”surrogate variables”. The surrogate variables are found by re-applying a partitioning algorithm to the predictor variable by using the other independent variables. In detail the predictor used for splitting the node in the tree becomes a binary response variable with the two classes determined by the split. The CART

algorithm performs only one step of partitioning of the predictor variable with missing values using the other independent predictor variables. For each independent variable a split is performed and ranked by the proportion of cases of the initial prediction variable that are misclassified. The independent variable with the lowest classification error is used to assign the cases of the missing values of the predictor variable.
This method of dealing with missing data rises some risks if there are a lot of missing values and surrogate variables are used. If a huge number of observations are manufactured an additional error is introduced to the tree model. Berk [1] gives the advice to avoid surrogate variables, because the temptations for misuse are given, and there is no clear threshold for missing data where the use leads to misleading results.

3.1.6 Strengths and Shortcoming of CART Models

The interpretability of the results and the possibility of showing them in a tree diagram are big advantages of the CART algorithm. It gives the possibility of an intuitive insight on the data set and shows interactions and dependencies. The influence of a predictor variable to the reduction of the impurity and its importance for the relationship with the response variable is shown in the tree diagram. The higher above the prediction variable in the tree is the more relevant it is. In Section 2.7 Breiman et al. [5] outline some advantages of the CART procedure:

- It's a very powerful and flexible classification tool which can be applied to any given data set or structure. There are no assumptions about the distribution of the data needed and can be applied to numerical or categorical variables.
- The final output is easy to interpret and can be used for predicting new data. In addition the tree includes classification and misclassification probabilities.
- The stepwise selection of the variables includes implicite information about the importance of a single variable in the model.
- It's very robust with respect to outliers and missclassified data sets.

However there are also some shortcomings of the CART algorithm one has to deal with. There is no formal representation which indicates that the algorithm will find the correct relationship between predictors and response variables from a given data set, even with all predictors provided and well measured. A suboptimal first split (or at one level) might allow better splits on lower levels and might lead to a better overall tree structure and a lower misclassification rate. Predictor variables with a large number of distinct predictor values are favored by the algorithm, therefore bias is included to the models. The flexible algorithm tends to overfit the data set and can be data specific. The specific characteristics and idiosyncrasies of the data set have a huge influence on the final results. The outcome is not general enough to fit new data. As already mentioned in previous sections that pruning may prevent overfitting. Too small numbers of observations in a node can cause instability in the results.

As outlined in Hastie et al. [9] an additional problem of tree models is their high variance. A small change in the data can result a very series of splits, different classification results and terminal nodes. This shortcoming of trees arise from the hierarchical nature of the process. The effect of an error in the top split is propagated down to all of the splits below. The variance can be reduced by bagging which averages many trees and is described in Chapter 4.
Consider response variables that are unbalanced and highly skewed might create some problems for any type of CART models. Sparse data-sets can lead to unstable results or an inability of the software to provide any results at all. Another problem is that rare observations may have a disproportional impact on the result, and a generalization to a bigger data set can be problematic. For the purpose of this work the highly skewed response variable can be problematic. It can be difficult to find predictors that are able to improve the overall fit. Considering a binary response variable with the marginal distribution of 95% of 1s and 5% of 0s. A good and accurate classification can be determined from this information alone. By assigning the 1 category the classification will be correct 95% of the time without using any predictors. Its hard to improve the results or do better than this. But using an adjustment of the priors, by placing heavy weight on the misclassification of rare cases, may improve the results. Its like saying there are more rare to be considered than the data indicate.

3.1.7 Unbiased Recursive Partitioning

As mentioned above the CART algorithm tends to choose predictor variables with more possible splits as a partitioning variable. There is a selection bias build into the tree structure and the results which are summarized in the terminal nodes. In their work Hothorn et al. [12] present a solution of this problem and suggest a tree building algorithm based on hypothesis testing. The procedure seems to be familiar with stepwise regression and has no bias toward certain predictors. The selection of a variable used for a split is determined by the results of a hypothesis test. A test of independence between any of the covariates and the response is performed.
The procedure can be summarized as follows (see Berk [1] p. 137).

1. Perform the global null hypothesis test for the independence between each predictor and the response variable.
2. Stop if the test cannot be rejected.
3. Select as the predictor having the strongest relationship with the response as splitting variable, if the test is rejected.
4. Choose the best split using only the selected predictor.
5. While no further splits are indecated, repeat steps 1-4.

The hypothesis tests are based on the permutation distributions in which values of the response variable are randomly shuffled. For each predictor a permutation test under the null

hypothesis of no association is performed. In addition to the calculation of the p-value of each predictor variable an overall p-value is also computed using methods for multiple tests (e.g. Bonferroni correction). In step 3 of the procedure the predictor with the lowest p-value is chosen if the global null hypothesis is rejected. The predictor with the lowest is equivalent to the predictor with the strongest relationship to the response variable. The split can be determined as usual. If a predictor variable is selected a split similar to the CART algorithm can be performed or by using the permutation tests setup. (Details see Hothorn et al. [12]) The process is performed for every subsequent partition of the data set until the algorithm stops.
In addition Hothorn et al. [12] showed in their work that the performance of prediction accuracy of trees with early stopping is equivalent to the prediction accuracy of pruned trees. This follows that this procedure offers a possible solution to the overfitting problem.
Nevertheless there are some concerns of this algorithm. For example, the result depends on the sample size and there is a permutation variation coming from the response variable. For details see Berk [1].
This procedure is implemented in the R package *party*.

3.1.8 Regression Trees

In this section we focus on the construction of trees with numerical response variables. The main difference of regression trees and classification trees lies in the splitting criterion. Lets consider a data set of N observations consisting of a numerical response variable and p predictor variables. In the classification case the reduction of the impurity

$$\max_s \Delta I(s, A) = \max_{A_L, A_R} I(A) - P(A_L)I(A_L) - P(A_R)I(A_R)$$

is maximized. The impurity of node A in the regression case is given by the sum of squares in the node

$$I(A) = \sum_{i=1}^{N} (y_i - \bar{y}(A))^2.$$

The method and concept used for regression and classification trees is the same. The deviance for a node A is given by

$$D(\tilde{x}, A) = \sum_{i=1}^{n} (y_i - \bar{y}(A))^2 - \left(\sum_{i=1}^{n_{\tilde{x},A}} (y_i - \bar{y}_{1,\tilde{x}}(A))^2 + \sum_{i=n_{\tilde{x}+1,A}}^{n} (y_i - \bar{y}_{2,\tilde{x}}(A))^2 \right)$$

with $\tilde{x}$ being the treshold at which a split is performed. So we are searching for

$$\max_{\tilde{x}} D(\tilde{x}, A).$$

It is not possible to use weights in regression trees, because it's not possible to consider false positives or false negatives. The impurity of the entire tree is given by the sum of overall

terminal nodes. The initial impurity was the sum of square error of the root node. The optimal sized regression tree can be reached by pruning, described in the former sections. The output of a regression tree is assigning the mean of the terminal node to each observation in that node. The assigned mean indicates the classification of an observation and represents the relation of the prediction variables on the numerical response.
A formal representation of a regression tree is shown in Hastie et al. [9]. A regression tree partitions the space spanned by the data set into disjoint regions $R_1, \ldots, R_J$, so that R_j is represented by the jth terminal node of the tree. The regression tree models the response variable y as constant in each region. By assigning a constant γ_j to each region than the relationship between predictors x and response y is

$$x \in R_j \Longrightarrow f(x) = \gamma_j.$$

A tree can be represented as

$$T(x, \Theta) = \sum_{j=1}^{J} \gamma_j 1_{\{x \in R_j\}}$$

with γ_j being the conditional mean of the response values that is assigned to terminal node j and tree parameters $\Theta = \{R_j, \gamma_j\}$.

3.1.9 Classification Tree models for Graz Mitte

The goal in this section is to find a reasonable tree model for predicting the concentration of particulate matter or the exceedance of the critical vale of 50 $\mu g/m^3$. In contrast to the 2 days ahead prediction model developed in Chapter 2 the tree model should be an extension of the current prediction model described in Stadlober et al. [23]. the focus lies on a reasonable prediction model 1 day ahead based on tree models. We will consider binary trees as well as regression trees implemented in the *rpart* package for the software environment **R**. In the following we compare the results applying Hothorn et al. [12] ideas of unbiased recursive trees using the *party* package in **R**. Similar to the regression model in Chapter 2 we will calibrate the model for Graz-Mitte using the seasons 2011/2012-2013/2014 and test the quality of the tree by applying them to the seasons 2014/2015 and 2015/2016.
First we consider a binary tree using the response variable **ueberschreitung** which is 1 if the daily average PM10 concentration exceeds 50 $\mu g/m^3$ and 0 if not. Our request to the tree model is the usability as a prediction model. Therefore we will not use any variables which are not available at midday for predicting the exceedance of the particulate matter concentration for the next day. Excluded of this limitation are meteorologic variables which can be forecasted in a proper way (e.g. precipitation, wind). In Figure 3.2 we see the first tree with the binary response variable **ueberschreitung** fitted to the whole data set of the Site Graz Mitte without any limitations. The resulting tree with 12 spilts seems to be quite complex and confusing. But PM10 concentration at midday one day before (**pm_mittag_lag1**) is a very important variable for splitting the data set. In addition the average temperature at midday one day before (**lute_mittag_lag1**) the binary variables wind velocity (**wind**), inversion

weather condition (**invers**) and precipitation (**nied01**) are used for spanning the tree. The error in the root node is around one fifth. The tree with 12 splits reduces the error by around 40% to a classification error of 12%. The reduction of the relative error seems to be not cost efficient in the sense of pruning and the tree is too large and complex. Pruning the tree with an $\alpha = 0.02$ gives us a very clear and easy tree structure (see: Figure 3.3) with 4 splits.

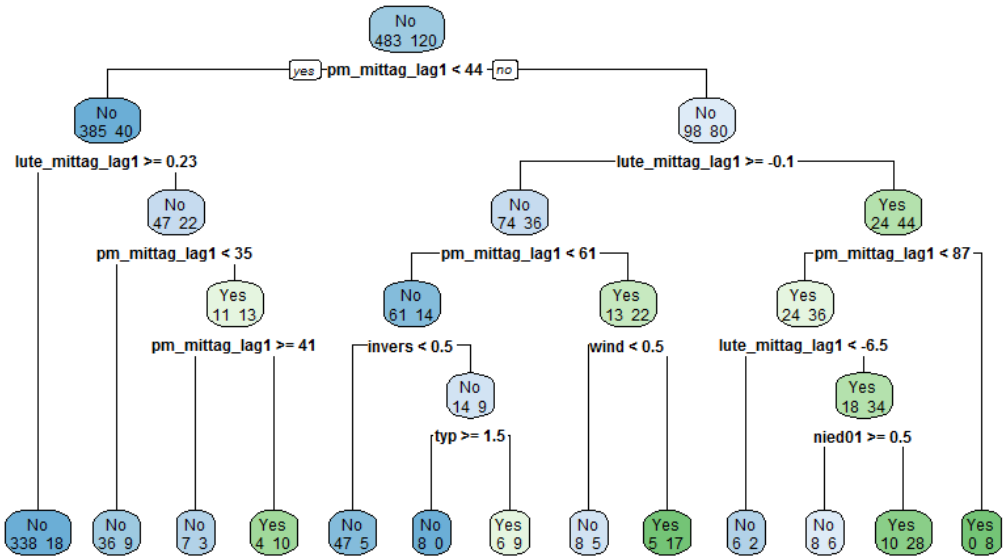

Figure 3.2: Unrestricted Binary Tree for Graz Mitte

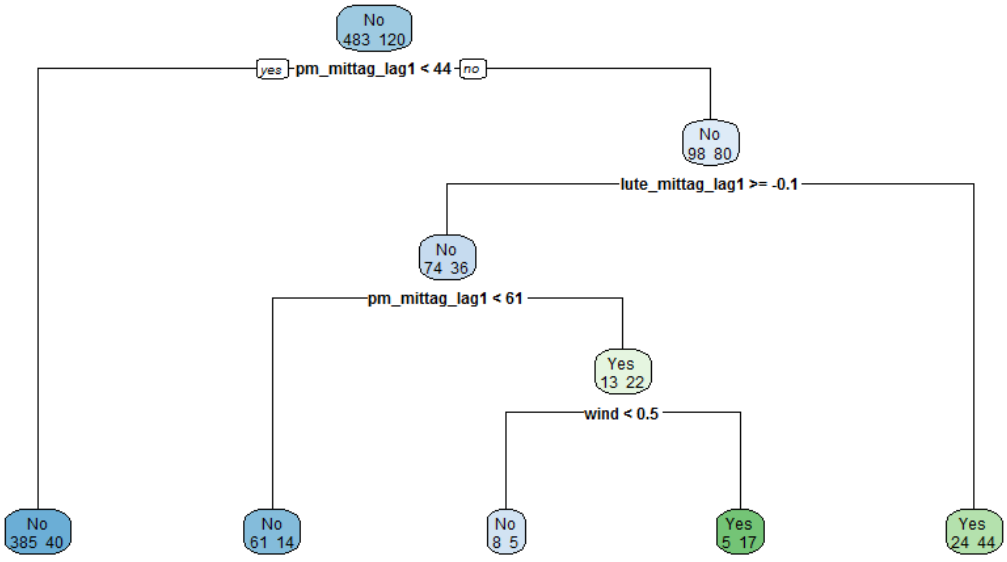

Figure 3.3: Pruned Binary Tree for Graz Mitte with $\alpha = 0.02$

Similar to the unconstrained tree the first split of the data is using the variable **pm_mittag_lag1** at a threshold of 44 $\mu g/m^3$. Air temperature and wind velocity (> 0.5 means low wind) are the remaining variables. Lower temperature and lower wind velocity leads to a classification of an exceedance. This result is in line with the general observations that lower ventilation in the basin area of Graz yields to exceedance of the critical level of 50 $\mu g/m^3$.
Using the trees for forecasting the winter seasons 2014/2015 and 2015/2016 we receive the results shown in tables 3.2 and 3.3

	Unconstrained Tree		Pruned Tree		Total
	Forecast 0-50 $\mu g/m^3$	Forecast 50-100 $\mu g/m^3$	Forecast 0-50 $\mu g/m^3$	Forecast 50-100 $\mu g/m^3$	
Obs. 0-50 $\mu g/m^3$	133	5	135	3	138
Obs. 50-100 $\mu g/m^3$	12	1	11	2	13
Total	145	6	146	5	151

Table 3.2: Results of forecasting winter season 2014/2015 with tree models

	Unconstrained Tree		Pruned Tree		Total
	Forecast 0-50 $\mu g/m^3$	Forecast 50-100 $\mu g/m^3$	Forecast 0-50 $\mu g/m^3$	Forecast 50-100 $\mu g/m^3$	
Obs. 0-50 $\mu g/m^3$	119	10	124	5	129
Obs. 50-100 $\mu g/m^3$	11	12	16	7	23
Total	130	22	140	12	152

Table 3.3: Results of forecasting winter season 2015/2016 with tree models

The forecasting results of both trees for both periods are quite stable and a misclassification rate of 9 respectively 14% for such a simple and transparent model is quite good. But especially the forecast of observations with a particulate matter concentration higher than 50 $\mu g/m^3$ should be improved. Before we introduce a loss matrix in the tree construction, we consider for the sake of completeness a classification tree using unbiased binary recursive partitioning.

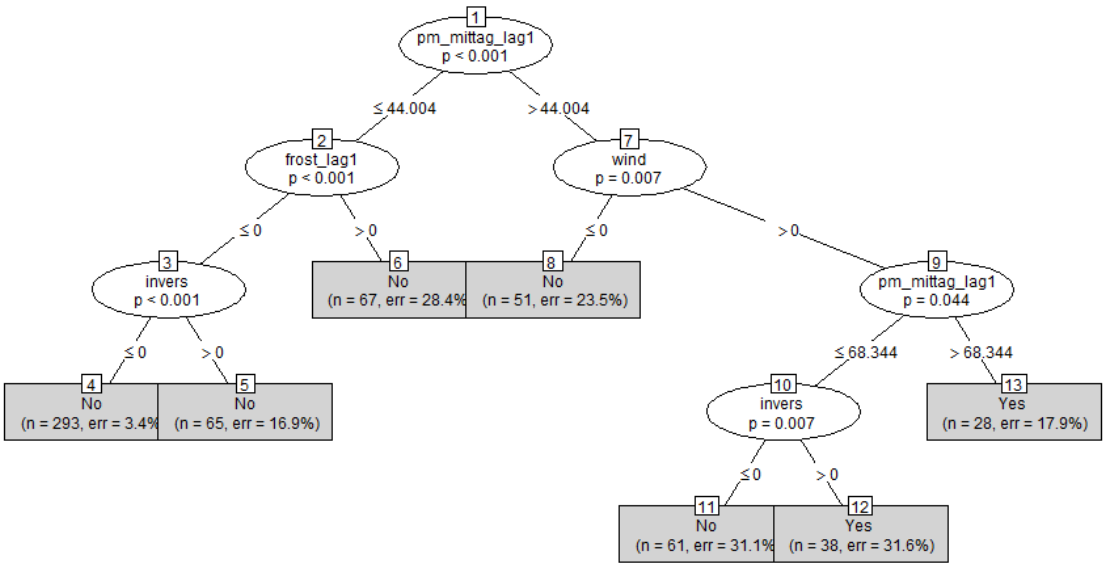

Figure 3.4: Classification Tree for Graz Mitte using unbiased binary recursive partitioning

The conditional inference tree in Figure 3.4, which was constructed using the package **party** in **R**, differs slightly from the tree in Figure 3.2. The variable **invers** seems to be more important. Furthermore the output in Figure 3.4 shows the corresponding p value of the hypothesis tests for each node.

	2014/2015		2015/2016	
	Forecast 0-50 $\mu g/m^3$	Forecast 50-100 $\mu g/m^3$	Forecast 0-50 $\mu g/m^3$	Forecast 50-100 $\mu g/m^3$
Obs. 0-50 $\mu g/m^3$	134	4	121	8
Obs. 50-100 $\mu g/m^3$	11	2	12	11
Total	145	6	133	19

Table 3.4: Results of forecasting for winter seasons 2014/2015 and 2015/2016 with unbiased recursive partitioning

By comparing the forecasting results for the seasons 2014/2015 and 2015/2016 both trees deliver similar results (Table 3.2-3.4). In the following analyses we will use the **rpart**-package only. The results of all trees are quite stable and deliver respectable results for the season 2014/2015 and 2015/2016, but there should be still a possible improvement by adding a loss matrix. Especially the forecast of non-exceedance in the case of a breach of the critical threshold may generate high costs. Therefore we introduce a loss matrix to the tree, where wrong predictions of an exceedance costs 4 times more than a wrong prediction of a non-exceedance of 50 $\mu g/m^3$. Using the generated tree for forecasting the winter seasons 2014/2015 and 2015/2016 we receive a higher total forecasting error of around 20%.(Table

3.5). But the wrong forecasting of exceedance of 50 $\mu g/m^3$ reduces to 9 respectively 3 observations for the winter seasons which is equivalent to 2 respectively 6%. Of course it's a trade-off of total error and a conservative approach of higher costs for wrong exceedance classification. Nevertheless the resulting tree shown in Figure 3.5, might be an additional tool for forecasting critical PM10 concentration for the site of Graz-Mitte. The tree seems to be not too complex and in addition there is only one binary meteorological forecasting variable needed. Only the prediction of an inverse temperature will be observed the next day. All other values for **lute_mittag_lag1** and **pm_mittag_lag1** are already available for the forecast.

	2014/2015		2015/2016	
	Forecast 0-50 $\mu g/m^3$	Forecast 50-100 $\mu g/m^3$	Forecast 0-50 $\mu g/m^3$	Forecast 50-100 $\mu g/m^3$
Obs. 0-50 $\mu g/m^3$	119	19	104	25
Obs. 50-100 $\mu g/m^3$	9	4	3	20
Total	128	23	107	45

Table 3.5: Results of forecasting for winter seasons 2014/2015 and 2015/2016 with loss matrix

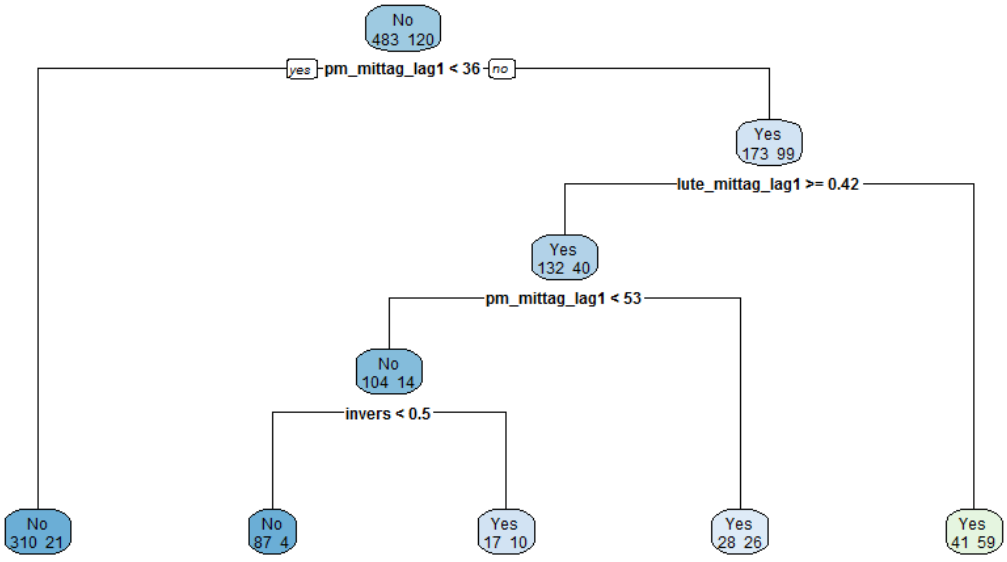

Figure 3.5: Classification Tree for Graz Mitte with loss matrix

3.1.10 Classification Tree models for Graz-Süd

For Graz-Süd we apply the same approach as for Graz-Mitte. First we consider a classification tree with **ueberschreitung** as response variable. Then we are going to prune the tree

and define the complexity parameter with the help of Figure 3.6. This figure shows us the improvement of the cross-validation error by growing the tree and the according complexity parameter. The optimal unconstrained pruned tree has 7 splits. Therefore the optimal complexity parameter for pruning the tree is $\alpha = 0.013699$. The according numbers can be found in the following output.

```
        CP nsplit rel error  xerror     xstd
1 0.300000      0   1.00000 1.00000 0.027670
2 0.068493      1   0.70000 0.73151 0.026053
3 0.023288      2   0.63151 0.66849 0.025411
4 0.020548      4   0.58493 0.61233 0.024744
5 0.017808      5   0.56438 0.59726 0.024548
6 0.013699      7   0.52877 0.58219 0.024346
7 0.010959      9   0.50137 0.57123 0.024194
8 0.010000     10   0.49041 0.56301 0.024077
```

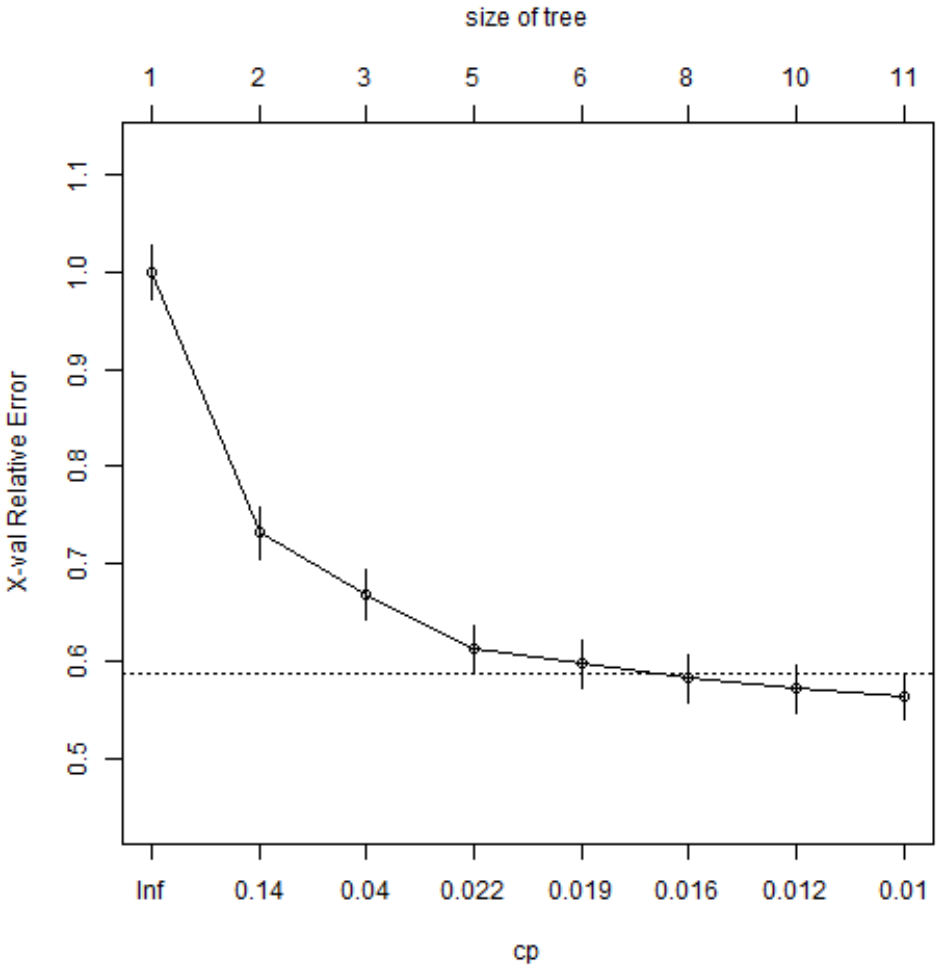

Figure 3.6: Complexity Parameter Plot for Graz-Süd

In comparison to Graz-Mitte the initial misclassification rate in Graz-Süd is much higher. More than 44% of the observations exceed the critical concentration of 50 $\mu g/m^3$. Therefore the fitting of a tree halves the misclassification error for the seasons 2003/2004-2013/2014. The variables for growing the tree (see Figure 3.7) are similar to the variables used for the tree model for Graz-Mitte. The most important variable is the PM10 24 hours average concentration one day before **pm_mittag_lag1**. In addition the binary variable velocity of the **wind**, the metric variable air temperature **lute** and the binary variable of a temperature inversion **invers** for the day of the prediction are included in the tree model. That means that the weather input variables are based on meteorological forecasts. A one day ahead meteorological forecast is in most cases very exact.

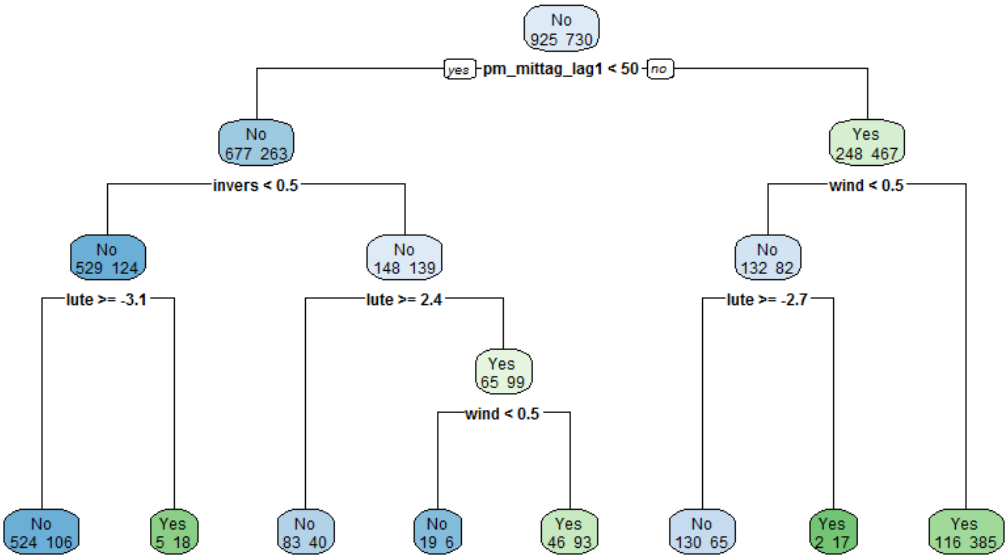

Figure 3.7: Pruned tree for Graz-Süd

Similar to Graz-Mitte we try to evaluate the quality of the model by backtesting it with data for the winter seasons 2014/2015 and 2015/2016. The results are shown in Table 3.6. The results are not satisfying because about 37% of the exceedances ($> 50\mu g/m^3$) are wrongly predicted for winter season 2014/2015 and about 23% for the winter season 2015/2016. In general the misclassification rate for both seasons is around 20%.

	2014/2015		2015/2016	
	Forecast 0-50 $\mu g/m^3$	Forecast 50-100 $\mu g/m^3$	Forecast 0-50 $\mu g/m^3$	Forecast 50-100 $\mu g/m^3$
Obs. 0-50 $\mu g/m^3$	83	16	76	20
Obs. 50-100 $\mu g/m^3$	19	33	13	43
Total	102	39	89	63

Table 3.6: Results of forecasting for winter seasons 2014/2015 and 2015/2016 with the pruned classification tree

The introduction of a loss matrix where a wrong prediction or misclassification of an exceedance of the PM10 concentration costs 2 times more than a misclassification of the a non-exceedance of 50 $\mu g/m^3$. The resulting tree is shown in Figure 3.8.

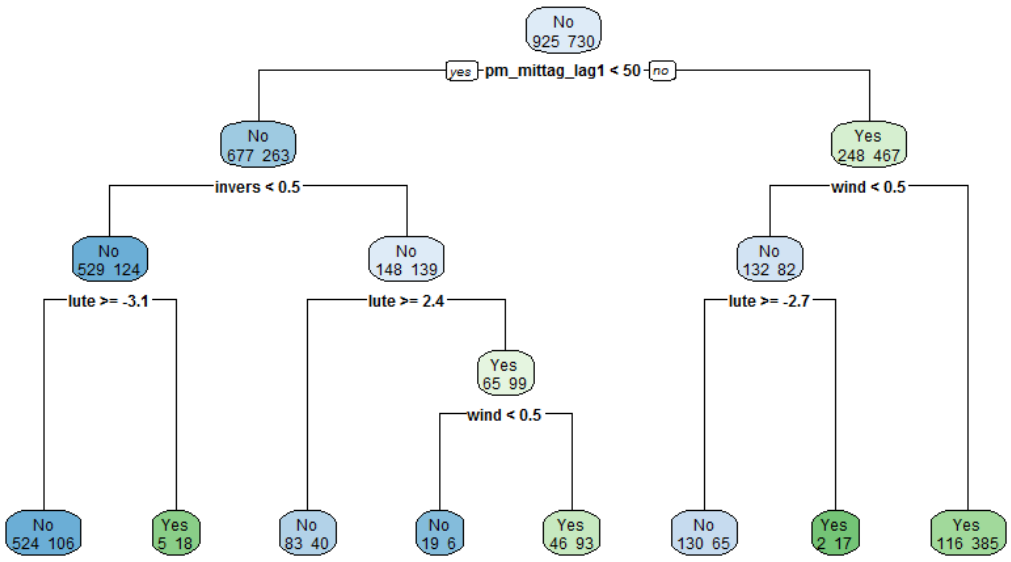

Figure 3.8: Classification Tree for Graz-Süd with loss matrix

Noticeable for this tree is, that the loss matrix does not change the importance of the prediction variables. The used splitting variables are very similar to the selected prediction variables in Figure 3.7.
The backtesting results for the seasons 2014/2015 and 2015/2016 show a slightly improvement against the results in Table 3.6.

	2014/2015		2015/2016	
	Forecast 0-50 $\mu g/m^3$	Forecast 50-100 $\mu g/m^3$	Forecast 0-50 $\mu g/m^3$	Forecast 50-100 $\mu g/m^3$
Obs. 0-50 $\mu g/m^3$	77	22	68	28
Obs. 50-100 $\mu g/m^3$	14	38	8	48
Total	91	60	76	76

Table 3.7: Results of forecasting winter seasons 2014/2015 and 2015/2016 with loss matrix

Increasing the costs for a wrong classification of the exceedance of the critical threshold would lead to an increase of the overall misclassification error, which is not desirable. Setting the cost ratio to 3 to 1 or 4 to 1, would yield an increase of the misclassification error of 40%. The wrong prediction of exceedances would sharply decrease, but the prediction error of non-exceedances would double. Therefore a trade-off of overall error and prediction error of exceedances is needed. This tree could help to get a fast indication if the next days PM10 concentration exceeds 50 $\mu g/m^3$.
In the following we set up a regression tree model for Graz-Süd and compare the results with the binary tree version.

3.1.11 Regression Tree for Graz-Süd

In previous sections we have seen tree models which could be used for predicting exceedance or non-exceedance of the PM10 concentration, but it was not possible to get a value of concentration. Therefore we construct a regression tree for the site Graz-Süd. We will replace the response variable **ueberschreitung** by the metric variable **PM10**. The total sum of square error without any split is 1583625. The maximum regression tree would have 12 splits and a very complex structure. In addition the number of observations in some leaves is rather small, therefore pruning of the tree will be applied. In Figure 3.9 one can see that a complexity parameter $\alpha = 0.29$ delivers the optimal pruned tree with 5 leaves. The root square error of the tree is reduced by around 37.5%.

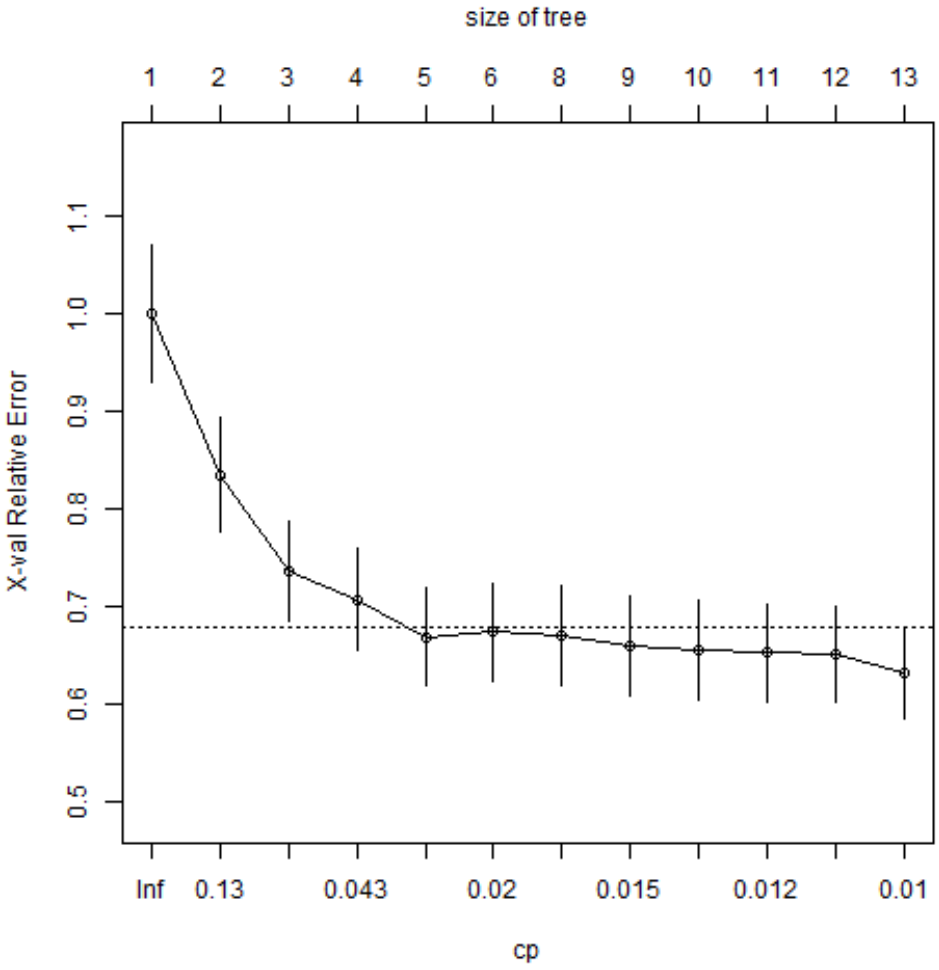

Figure 3.9: Complexity Parameter Plot for Regression Tree Graz-Süd

Similar to all other binary trees is that the most important variable in the regression tree (Figure 3.10) is **PM_mittag_lag1**. The first split at **PM_mittag_lag1** splits the data in 37.5% where **PM_mittag_lag1** larger than 55 $\mu g/m^3$ and 62.5% where **PM_mittag_lag1** is smaller. Based on the correlation between lower wind velocity and the exceedance of the threshold the second important variable is the wind velocity of **wind**. Observations with lower **PM_mittag_lag1** and higher wind speed (wind < 0.5) yield to a lower average response of 33 $\mu g/m^3$. One forth of the data is in this leave. 629 cases have lower **PM_mittag_lag1** concentration than 55 $\mu g/m^3$ and a wind speed lower than $0.6m/s$ and an average response of 49 $\mu g/m^3$. For all cases with a **PM_mittag_lag1** larger than 55 $\mu g/m^3$ the next split performed is the wind speed and for those where the wind velocity is below $0.6m/s$ (binary variable wind > 0.5) are additionally split using **PM_mittag_lag1** concentration. The tree structure is very simple and only two prediction variables are used.

Applying the regression trees to the seasons 2014/2015 and 2015/2016 for backtesting the model, shows a high misclassification error (see table 3.8). We used the regression tree for predicting the metric variable *PM10* and transformed it to a binary variable.

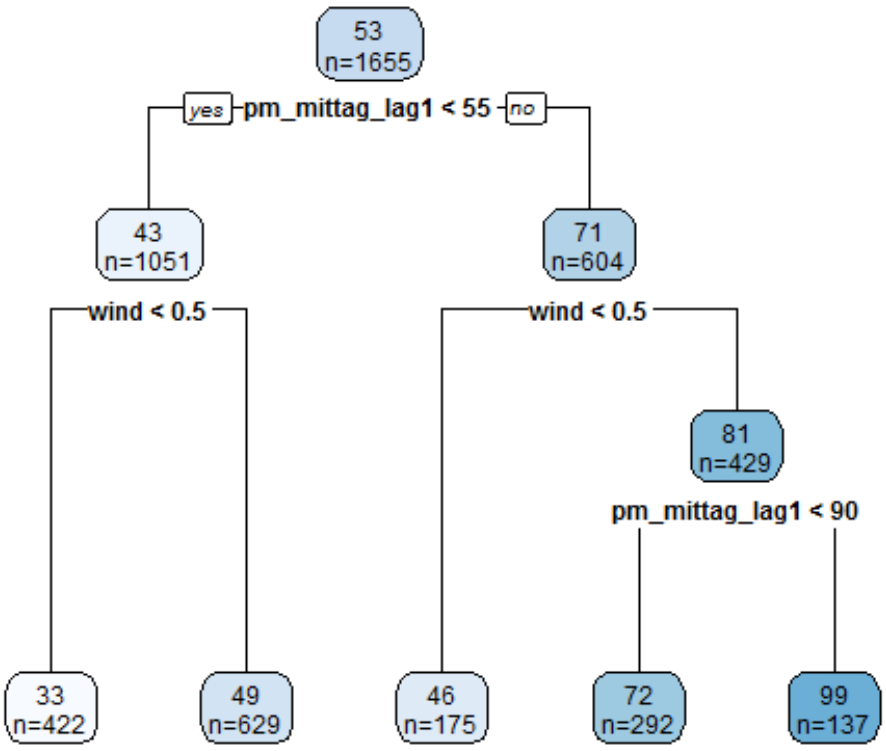

Figure 3.10: Complexity Parameter Plot for Regression Tree Graz-Süd

	2014/2015		2015/2016	
	Forecast 0-50 $\mu g/m^3$	Forecast 50-100 $\mu g/m^3$	Forecast 0-50 $\mu g/m^3$	Forecast 50-100 $\mu g/m^3$
Obs. 0-50 $\mu g/m^3$	93	6	86	10
Obs. 50-100 $\mu g/m^3$	34	18	26	30
Total	127	24	112	40

Table 3.8: Results of forecasting for winter seasons 2014/2015 and 2015/2016 of the regression tree

In general the misclassification error is similar to the error of the binary trees for Graz-Süd. For both seasons 2014/2015 and 2015/2016 we get an error of around one forth which is only negligible higher than the error in Table 3.6, but the amount of false predicitions of exceedances of the critical threshold is much higher and therefore this model is not suitable for forecasting of the PM10 concentration.
Comparing the tree regression tree model with the linear regression model developed by Stadlober et al. [23] the wrong classification of exceedances using the tree model is 3 times higher for the seasons 2014/2015 and 2015/2016. The misclassification error of the binary tree models in both seasons are similar to the current regression model used for prediction. In spite of the simplicity of the model the binary tree model can not replace the current model, because it does not return a PM10 concentration. It could only be a supplementary tool which helps to give a quick indication of exceeding the threshold or not.
In the next section we will focus on models which use many sets of tree models to produce their output and improve the results.

Chapter 4

Prediction Model of PM 10 using tree-based models

In this chapter we consider methods generating multiple versions of results and using these to get an aggregated output. In contrast to the models described in the previous chapter the following methods take the results of several classification or regression trees and aggregate them. The major transaction is from methods producing a single set of results to methods using many sets of results for the final output.
The benefit of the approach of aggregating is that it can avoid over-fitting of data-sets. The averaging procedure tends to exclude results based on idiosyncratic features of a given data set. This can lead to a stabilization and increase of the accuracy of the model as well as a reduction of the variance. This approach offers a solution for the bias variance trade off of the CART algorithm. Such procedures reduce the instability of CART, while the flexibility of tree models remains. Based on the outcome of tree models described in Chapter 3 we use the ”bagging” algorithm going back to Breiman [2] and random forest introduced by Breiman in 2001 [4] for predicting particulate matter concentration in Graz.
The shortcoming of this methods is that the information results of single trees get lost and the descriptive representation of the tree does not exist anymore.

4.1 Bagging

Bagging which stands for ”Bootstrap Aggregating” first used by Breiman [2] for averaging results of different tree models. It’s a combination of taking random bootstrapping samples of a given data set, fitting trees to each sample and aggregate them. The algorithm of bagging for the classification case, of a data set with N observations, can be described as follow (see Berk [1]).

1. Draw a random sample of size N with replacement from the data set. (bootstrap sample)

2. Build a tree without pruning and assign a class to each terminal node.

3. Store the class label of each single case and the prediction values for each observation.
4. Run step 1-3 for a large number of times e.g. 1000.
5. Count the different classifications of each observation over all generated trees.
6. Assign the final category for each observation by majority with one vote per tree over all generated trees.
7. Construct a confusion table using the final classification.

The bagging algorithm can also be applied for the regression case, where in step 6 the result for each case is achieved by averaging the results over all trees. Beside all advantages of the bagging algorithm the biggest shortcoming is that the single tree structure does not exist anymore. Therefore the tree diagram which gives a direct way to consider the predictors is related to the output.

4.1.1 Out-of-bag observations

In the basis algortithm of bagging we draw a bootstrap sample with replacement of size N and build a tree with these observations. This step implies that around one third of the data samples are not used for building the tree (compare Breiman [4]). The data which are left out are called "out-of-bag" (OOB) samples and will be used for testing the constructed tree. The OOB samples imply a model validation and one gets an internal estimation of the model error.
There is no additional need for cross validation or an additional test data set. The OOB samples are dropped down the tree instead of the observations used for constructing the tree and the observed response value is compared with the estimated value.
The OOB testing delivers an unbiased estimation of the prediction error. Breiman [4] mentions that the error estimation is as good as it would be produced with a separate test set. If the number of trees goes to infinity the observed prediction error received by the OOB test converts to the true prediction error (compare Breiman [4]).

4.1.2 PM10 prediction model for Graz-Mitte using bagging

We apply the bagging algorithm to the data set of Graz-Mitte by using the **R**-package **ipred** and the function **bagging**. As an input we use the whole data set with 23 prediction variables consisting of meteorological, PM10 and binary variables linked to the calendarian date. We use 500 bootstrap replications and receive an out-of-bag misclassification error and a general misclassification error around 16%. Using the bagging model for forecasting the winter seasons 2014/2015 and 2015/2016 we receive sightly better results than using tree models only (compare Table 3.3). In Table 4.1 one see the forcasting results of the calibrated model for Graz-Mitte using the bagging algorithm. Similiar to the tree model it's also possible to

	2014/2015		2015/2016	
	Forecast 0-50 $\mu g/m^3$	Forecast 50-100 $\mu g/m^3$	Forecast 0-50 $\mu g/m^3$	Forecast 50-100 $\mu g/m^3$
Obs. 0-50 $\mu g/m^3$	135	3	124	5
Obs. 50-100 $\mu g/m^3$	10	3	14	9
Total	145	6	138	14

Table 4.1: Results of forecasting for winter seasons 2014/2015 and 2015/2016 using the bagging algorithm

add a loss matrix to the bagging procedure. The additional input of the loss matrix to the model (compare Section 3.1.9) yields no improvement of the forecasting results for the seasons 2014/2015 and 2015/2016. One can see that averaging outputs of trees helps enhancing the results.

4.1.3 PM10 prediction model for Graz-Süd using bagging

We apply the similar procedure used for Graz-Mitte also for Graz-Süd. The out-of-bag misclassification and real misclassification error is around 19.5% of the bagging model, which is sightly better than for the tree model in Section 3.1.10. Using the bagging model for forecasting the winter season 2014/2015 one receives a sightly better result of the overall misclassification rate of 21%, but the error of wrong predictions of PM10 concentrations where a breach of the limit is observed increases from 19 to 23. More than 44% of the observed concentrations higher than $50\mu g/m^3$ is wrongly forecasted (see Table 4.2). Therefore we will introduce a loss matrix to the bagging algorithm and try to improve the results. For the winter season 2015/2016 the results of the bagging model are sightly better than results of the tree model (compare Section 3.1.10). By adding a loss matrix where a wrong forecast

	2014/2015		2015/2016	
	Forecast 0-50 $\mu g/m^3$	Forecast 50-100 $\mu g/m^3$	Forecast 0-50 $\mu g/m^3$	Forecast 50-100 $\mu g/m^3$
Obs. 0-50 $\mu g/m^3$	90	9	80	16
Obs. 50-100 $\mu g/m^3$	23	29	12	44
Total	113	38	92	60

Table 4.2: Results of forecasting for winter seasons 2014/2015 and 2015/2016 using the bagging algorithm

of a limit $> 50\mu g/m^3$ costs double than a wrong forecast of no breach yields to no improvement of the model.
In general the sightly improvement of the forecasting error by using the bagging model does not satisfy the loss of the tree structure.

4.2 Random Forest

Now we extend the bagging algorithm by combining it with a randomization of the input variables that are used when considering candidate variables to split a specific node. Instead of looking for the best split among all prediction variables, in random forests at each node only a subset K of the predictors is used. Out of this subset the best split is estimated. In contrast to the bagging algorithm random forests are extended by the randomness of the variable selection for each split. Similar to bagging the predictive power of the random forest is measured using OOB samples. A formal representation of random forest can be found in Breiman [4]. The basic random forest algorithm for the classification case is given by the following steps (see Berk [1]).

1. Draw a random sample of size N with replacement for the data (bootstrap sample).
2. Take a random subset of size K of the predictors.
3. Construct the first split of the tree, using the chosen predictors.
4. Repeat step 2 for each subsequent split in the tree. Do not prune.
5. Drop the OOB samples down the tree and store the assigned class of the observations.
6. Run steps 1-5 for a large number of times, e.g. 1000.
7. Use only the assigned class of the observation when that observation was an OOB sample to calculate by majority vote its classification.

Bagging and Random forest do not overfit due the law of large numbers (proven in Breiman[4]), reduce the variance and produce more stable results. The strength of random forest is dealing with a large number of predictors and even the possibility to get valid models if the number of predictors is larger than the number of observations. In contrast to single trees the randomness allows more different predictors to contribute to the splits of nodes. The large number of runs gives the possibility for each predictor at least several times to define a split. Predictors that otherwise would not have been considered can play a certain role in the model. Therefore the fitting process uses much more information of the data than a CART model.
The absence of the interpretable tree structure and the black box algorithm are the big disadvantages of bagging and random forests.

4.2.1 Margins

Margins are an important concept for the understanding of random forests. The margin is the difference between the proportion of times a case is correctly classified and the proportion of times it is incorrectly classified. Therefore large margins of OOB observations for all data sets are desirable, because this indicates a stable result

4.2.2 Variable Importance

To measure the importance of a predictor variable Breiman [4] suggests to sum the weighted impurity decreases. The following measure is called "mean decrease Gini" if the Gini function is used as impurity function. (compare Louppe [8])

$$Imp_{x_m} = \frac{1}{N_T} \sum\nolimits_T \sum_{t \in T: v(s_t) = x_m} p(t) \Delta I(s_t, t)$$

The importance of the prediction variable x_m for predicting y is the average over all N_T trees in the forest of the weighted impurity decreases $p(t)\Delta I(s_t, t)$ for all nodes t where x_m is used. $p(t)$ is the proportion N_t/N of samples reaching t and $v(s_t)$ is the variable used in split s_t.
The above approach of measuring the average decrease of the impurity each time a given variable is used to define a split is still valid. But this importance measure is based on information used for building the tree. This measure ignores the forecasting skills of the random forest, because it's not calculated for a test data-set.
Therefore Breiman [4] proposed another measure which is based on the reduction in predictive accuracy when a predictor is shuffled.
Dropping down the out-of-bag data in every constructed tree in the forest to get the prediction error of the model $\mathbb{E}$. Then the values of the desired predictor is randomly shuffled and dropped down the tree again. This makes the predictor on the average unrelated to the response and all other predictors. Run this for each predictor x_m, $(m = 1, \ldots, p)$ and compute

$$Imp_{x_m} = \frac{1}{N_T} \sum_{n_t=1}^{N_T} (\mathbb{E}_m^{n_t} - \mathbb{E}^{n_t}) m = 1, \ldots, p$$

where N_T is the number of trees, $\mathbb{E}_m^{n_t}$ is the forecasting error with predictor m is shuffled in the n_t-th tree and $\mathbb{E}^{n_t}$ is the general forecasting error in tree n_t without shuffling.
In the following analysis of the PM10 data set we will identify the most important variables by importance plots which use the measures above.

4.2.3 Partial Dependence

The relationship between the predictor variable and the response variable can be represented by a partial dependence plot which is discussed for example in [1] Berk 2008 and [9] Hastie et al. 2001. The plot is constructed as follows (see Berk [1]):

1. Grow a random forest.
2. Consider the prediction variable of interest e.g.x_m, which has υ distinct values $(\upsilon \leq N)$ in the training data.
3. Construct for each of the values υ of x_m a new data set where x_m only takes on that value while all other prediction variables remains untouched.

4. For each of the υ data sets predict the response using a random forest. Average over each of the υ sets of response predictors to get υ single values.

5. Plot the average prediction against the different values of x_m

6. Run steps 2-5 for each desired predictor.

The resulting plot shows the relationship between the different values of a given predictor and the response averaged within each single value. All other predictors are set constant during the procedure. Thus there is no information about the interaction effects unless the corresponding interaction variable is constructed in advance and used as a predictor.

4.2.4 Random Forest Models for Graz-Mitte

In this section we fit a random forest model for the site Graz-Mitte. First we fit a classification model of the binary variable **überschreitung** with 23 predictors. In the following we construct a model for the metric variable **PM10**. The first fitted random forest consists of 500 trees and selects 4 variables randomly (by default) as candidate for each split. By considering the number of variables to choose randomly for each split in each tree the random forest against the corresponding OOB-error in Figure 4.1 emphasizes the default assumptions of 4 splits.

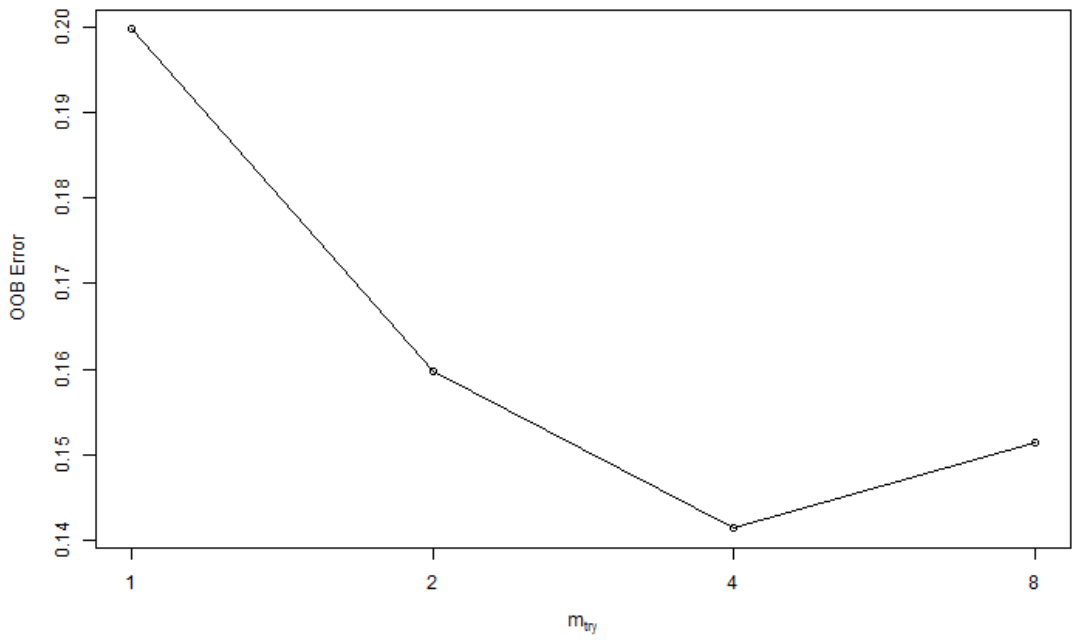

Figure 4.1: Number of randomly choosen variables in each split and the corresponding OOB-error

The number of 500 trees in the random forest seems to be reasonable, because the error rates in dependance of the number of trees in Figure 4.2 show a stabilization of the error rate after 500 trees. There is a high fluctuation of the model error in forests with lower number of trees. We see that using 1000 trees in the models seems to be too much and 300 too less.

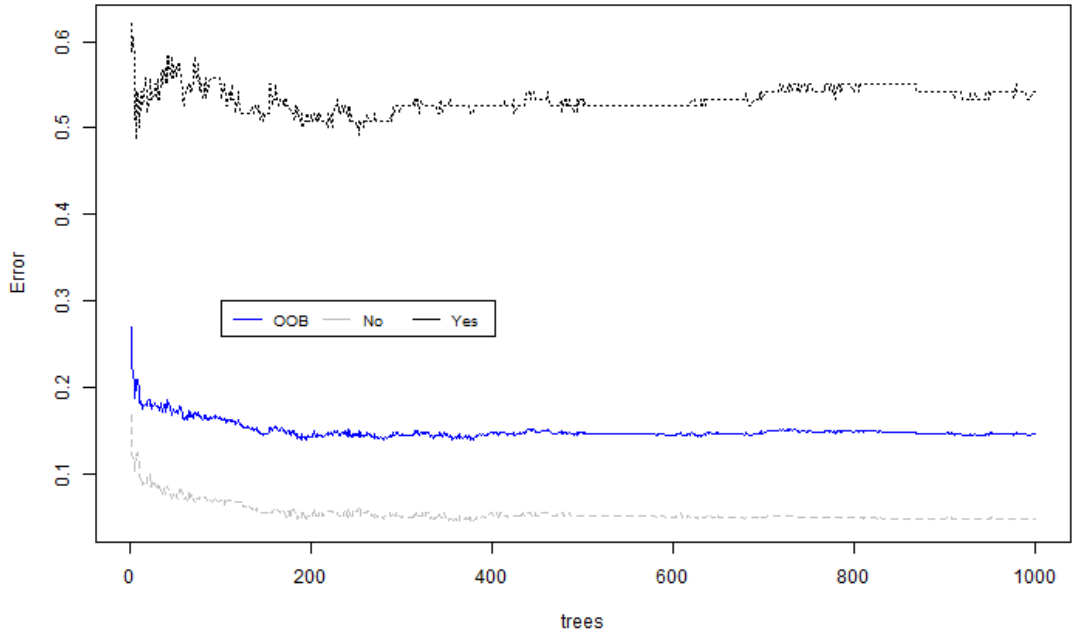

Figure 4.2: Number of tree in the forest and the corresponding OOB-error

To reduce in the further steps the model complexity the variable importance is essential. Figure 4.1 shows the 2 different variable importance measures described in Section 4.2.2. We see that the importance measures deliver slightly different variables. The importance of the variables mainly coincides with the classification tree model in Chapter 3 and the current regression model used as prediction model described in Stadlober et al. [23]. The importance of the variable **pm_mittag_lag1** in the forest model (see Figure 4.3) is in line with the strong auto-correlation in the data shown in Figure 2.3. We see that many calendarian variables are negligible which is not total in line with our expectations.

The OOB-error of 14.31% of the full forest model is slighty lower than the OOB-error of 15.81% of the forest model using only the 6 most important variables. But applying the model for forecasting the exceedance of the critical limit of 50 $\mu g/m^3$ to the winter seasons 2014/2015 and 2015/2016 the full model delivers much better results. The misclassfication error of the full model for the winter season 2014/2015 is only 7% and for 2015/2016 is around 10.5% (see Table 4.3). The shortcoming of this model is similar to the tree and bagging models in the former chapters. It is still underestimation of the PM10 observations

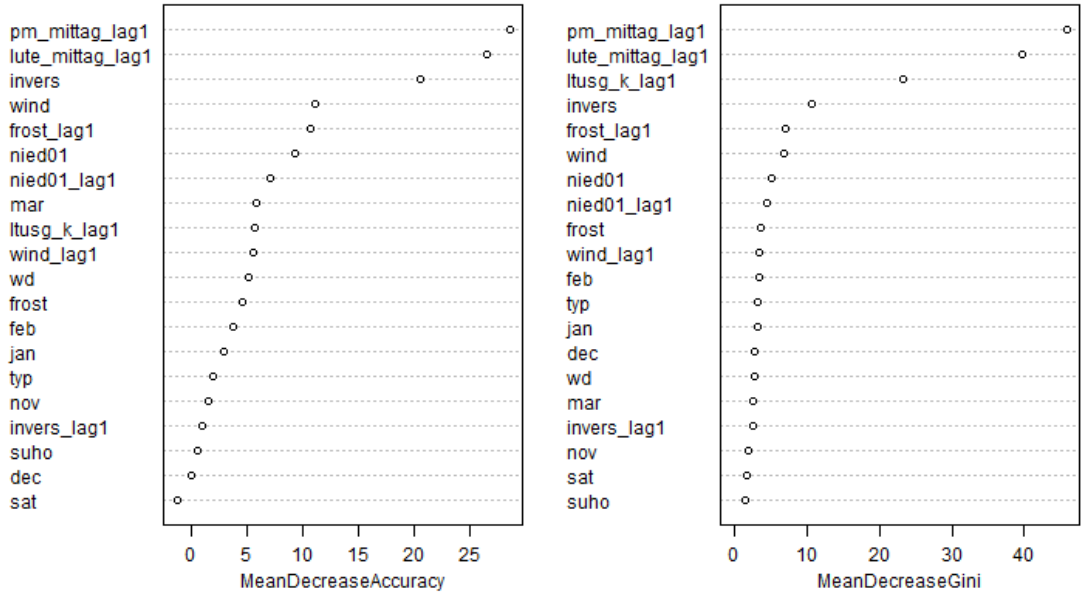

Figure 4.3: Importance functions of the random forest model for Graz-Mitte

above 50 $\mu g/m^3$. An introduction of a loss matrix to the forest model does not decrease the misclassification error for both seasons.

By applying the same steps of a forest model described above for the binary variable

	2014/2015		2015/2016	
	Forecast 0-50 $\mu g/m^3$	Forecast 50-100 $\mu g/m^3$	Forecast 0-50 $\mu g/m^3$	Forecast 50-100 $\mu g/m^3$
Obs. 0-50 $\mu g/m^3$	137	1	126	3
Obs. 50-100 $\mu g/m^3$	10	3	13	10
Total	147	4	139	13

Table 4.3: Results of forecasting for winter seasons 2014/2015 and 2015/2016 with a random forest model

überschreitung to the metric variable PM10 yields to a similar forest structure. The number of trees in the forest is set to 500 and the randomly chosen variables for each split is 4. The random forest model with regression trees describes only 51.15% of the variance in the data and the RMSE of 13.3 which yields to weaker results in forecasting than the regression model in Stadlober et al [23].

Now we consider the relationship between the predictor variables between 2 of the most important predictors **pm_mittag_lag1**and **lute_mittag_lag1** using the partial dependence plot. Figure 4.4 shows a linear relationship between the PM10 concentration and the concentra-

tion of PM10 at midday the day before and an almost negative linear relationship of the temperature and the PM10 concentration. These two plots emphasize the high autocorrelation in the PM10 data and that the probability of exceeding the critical limit rises with lower temperatures.

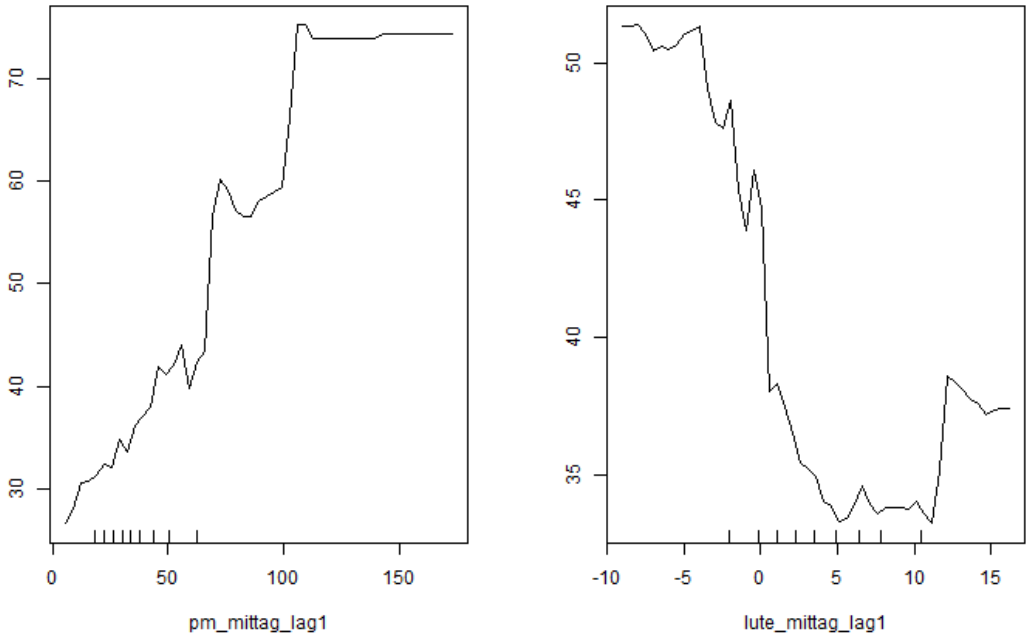

Figure 4.4: Partial dependence plots of 2 top contributing predictors for Graz-Mitte

4.2.5 Random Forest Models for Graz-Süd

For the site Graz-Süd we apply the same analysis steps as for Graz-Mitte. The number of trees in the forest is set to 500 and the variables chosen for each split are estimated by 4. The OOB-error of the calibrated forest model is 18.7%, which is lower than the misclassification error of the tree models for Graz-Süd in Chapter 3. By considering the plot (Figure 4.5) of the different importance functions we see that the result of the most important variables is similar to the results of Graz-Mitte.

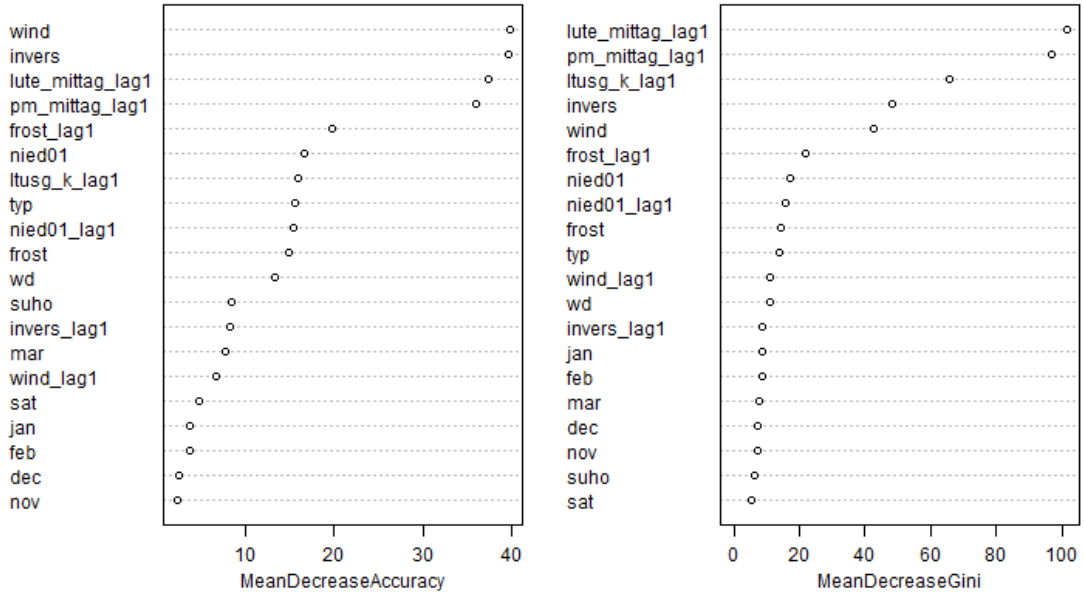

Figure 4.5: Importance functions of the random forest model for Graz-Süd

The forecasting results for the season 2014/2015 and 2015/2016 of the forest model are in the first season worse than the results of the tree model and in the second season slightly better than the pruned tree (see Table 4.4).

	2014/2015		2015/2016	
	Forecast 0-50 $\mu g/m^3$	Forecast 50-100 $\mu g/m^3$	Forecast 0-50 $\mu g/m^3$	Forecast 50-100 $\mu g/m^3$
Obs. 0-50 $\mu g/m^3$	90	9	76	20
Obs. 50-100 $\mu g/m^3$	24	28	9	47
Total	114	37	85	67

Table 4.4: Results of forecasting for winter seasons 2014/2015 and 2015/2016 with a random forest model

The high rate of wrong misclassifications of observations with PM10 concentrations higher than 50 $\mu g/m^3$ is a result of the margins within the calibrated model.

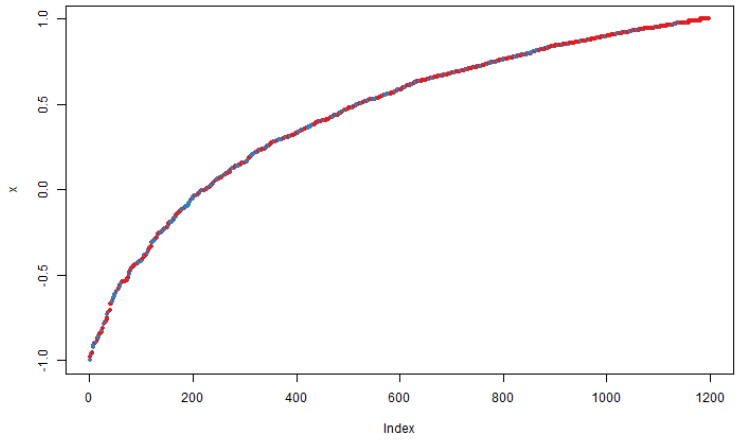

Figure 4.6: Margins of the random forest model for Graz-Süd

In Figure 4.6 we see that nearly a third of the observations has margin below 0.5, and 200 obersvations a margin below 0 which can lead to instable results for forecasting the PM10 concentration.
One can see that the bagging and forest models yield sightly better results than the tree models and stabilizes the results. Similar to the tree model in the regression case the predictive power is much lower than the current linear regression model for PM10 described in Stadlober et al. [23].

Chapter 5

Prediction Model of PM 10 using Neural Networks

Within this chapter we apply a widely used machine learning method for the prediction of daily PM10 average concentration. Similar to Perez [19] we use neural networks for modeling the particulate matter concentration of Graz. This black-box method uses concepts borrowed from an understanding of human brains in order to model arbitrary functions. It uses a network of artificial neurons or nodes to solve learning problems. During the last decades, neural networks become more popular because of the increase of computational power. The big advantage of neural networks is that a combined model for Graz-Mitte and Graz Süd can be constructed.

5.1 Neural networks

Before applying neural networks to the PM10 data of the sites Graz-Mitte and Graz-Süd we give a briefly introduction to neural networks, where we mainly follow Hastie et al. [9] and Lantz [15]. A neural network is a non-linear statistical mdel and can be seen as a two-stage regression model. The neural network is typically represented as a network diagram see Figure 5.1.

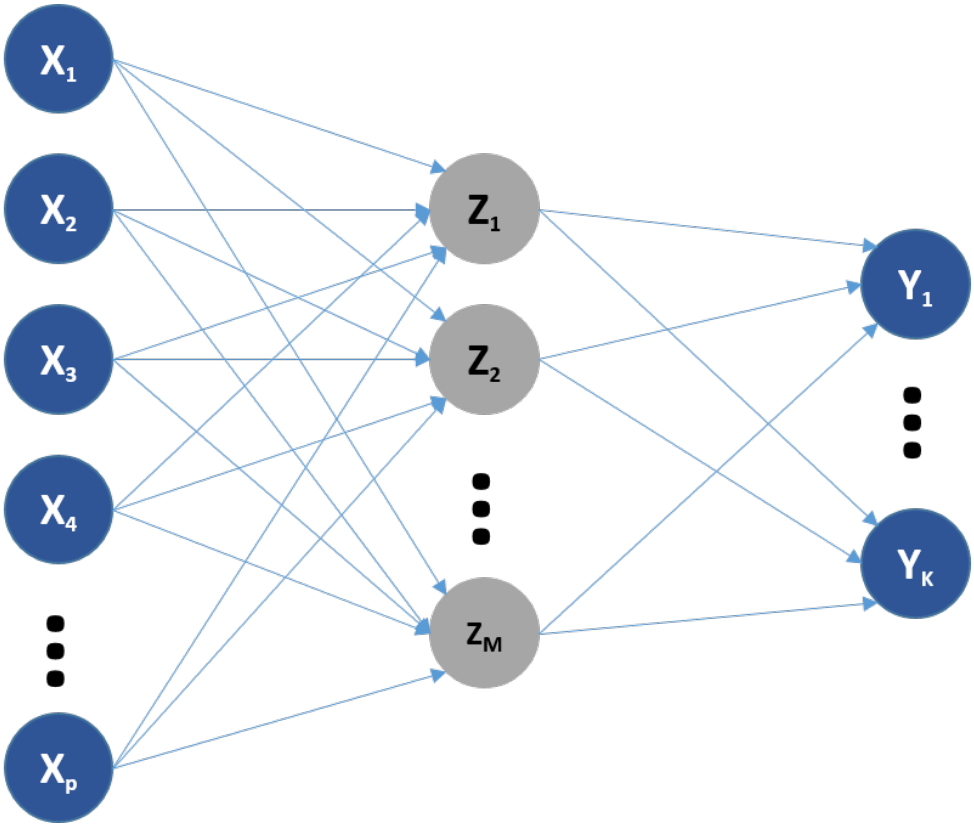

Figure 5.1: Sigmoid $\sigma(\upsilon) = 1/(1 + e^{-\upsilon})$ activation function

The advantage of neural networks in comparison to regression models is that there are several possible output units. A regression has only one possible output unit at the right hand side of the figure Y_1. Considering a K-class classification, there are K units at the right hand side, with k target measurements Y_k, $k = 1, \ldots, K$ where each is coded as a binary variable for the k-th class.
The layer which is marked with Z is created from linear combinations of the input and is called hidden layer. The target Y_k is modeled as a function of linear combinations of Z_m

$$Z_m = \sigma(\alpha_{0m} + \alpha_m^T X, \quad m = 1, \ldots, M$$

$$T_k = \beta_{0k} + \beta_k^T Z, \quad k = 1, \ldots, K$$

$$Y = f_k(X) = g_k(T), \quad k = 1, \ldots, K$$

where $Z = (Z_1, Z_2, \ldots, Z_M)$ and $T = (T_1, T_2, \ldots, T_K)$. The function σ is called activation function and is usually chosen to be *sigmoid* $\sigma(\upsilon) = 1/(1 + e^{-\upsilon})$. (see Figure 5.2).

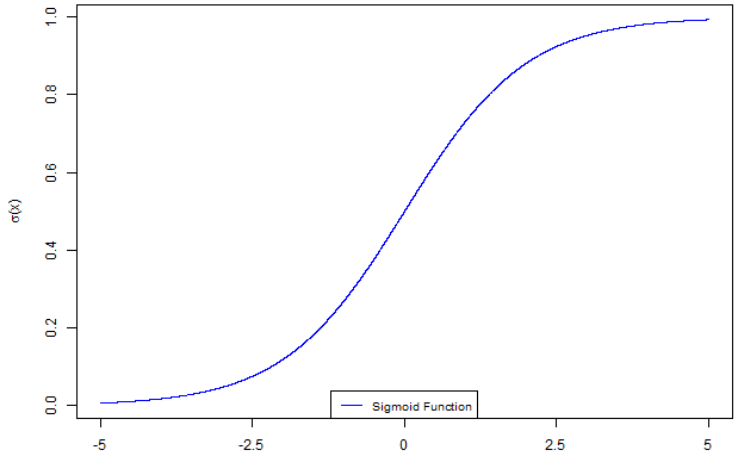

Figure 5.2: Sigmoid $\sigma(v) = 1/(1 + e^{-v})$ activation function

In the following we will observe an additional bias added to the neural network which can be seen as an additional input feature. This bias is captured by the intercept (α_{0m} and β_{0k}) in the model.
The output function $g_k(T)$ allows a final transformation of the vector of outputs T. For regression models typically $g_k(T)$ is the identity function and for K-class classification often the softmax function

$$g_k(T) = \frac{e^{T_k}}{\sum_{l=1}^{K} e^{T_l}}.$$

used. The hidden layer Z_m is not directly observed. In the further analysis we see that neural networks can have more than one hidden layer. The hidden layer Z_m is a basis expansion of the original input using the according weights α_m and the activation function σ. The difference of the basis expansion techniques using for a regression model is that the parameters in the neural networks are learned from the data. If the activation function is the identity function the neural network collapses to a linear model in the inputs. We see in Figure 5.2 that the non-linearity of transformation depends on the norm of the weights α. If $||\alpha||$ is very small the unit will be operating in the linear part of the activation function.

5.1.1 Network topology

An essential attribute of neural networks is their topology and structure of interconnected neurons. The topology of the network defines the complexity of the task that can be learned by the network. Larger and more complex networks are capable to identify more complex facts.
The network shown in Figure 5.1 is defined as a mulit-layer network with one hidden layer. The number of hidden layers can vary for the different degree of complexity. We see that

the network is fully connected which means that every node is connected with every node in the next layer. Another essential input to the network is the number of nodes in each layer. There is no reliable rule to determine the number of neurons in the hidden layer. The number depends on the number of input nodes the amount of data sets and the complexity of the learning task. Similar to tree models a trade-off of complexity and stability for prediction must be met. Too complex networks need a huge computational power and may be poor in the predictive power of new data.
The training of networks and adjusting the connection weights is very intensive, but can be solved by the algorithm which is known as backpropagation (compare Hastie et al. [9]). This algorithm makes neural networks common in the field of data mining and machine learning. In the following we use the **R**-package **neuralnet** to create neural networks for Graz-Mitte and Graz-Süd.

5.2 Neural network for Graz-Mitte

Similar to the tree based models we will train the neural networks using the seasons 2010/2011-2013/2014 for training the network and validating its prediction power by applying it to season 2014/2015 and 2015/2016. As mentioned above there is no concrete rule for determining the number of hidden layers as well as nodes in the hidden layer. Therefore we apply several neural networks with different amount of nodes in the hidden layer and we vary the amount of hidden layers between 1 and 2. After calibrating the neural network we compare the predictive power for the seasons 2014/2015 and 2015/2016 of each single network. We compare the misclassification error of the exceedance of the critical limit of 50 $\mu g/m^3$ and the quality of the forecast using quality function II from Section 2.1.4. As input to the neural network we use data which are available at midday before the desired day of prediction and binary meteorological variables of the day of the prediction. The meteorological variables for day $t+1$ are **frost, invers** and **nied01**. The indention for using those variables is, that a yes or no prediction of those input variables should be possible for one day ahead. There are 19 different input variables to the neural network for Graz-Mitte and the target variable is the average PM10 concentration. The number of the nodes in the first hidden layer is limited with 12 nodes, and the nodes in the second hidden layer is limited with the number of nodes of the first hidden layer. Using this procedure we receive 81 different neural networks starting with a simple network with only one node in the first hidden layer up the the most complex model with 12 nodes in the first hidden layer and 11 nodes in the second one. In Table 5.1 we see the forecasting results for the winter season 2014/2015 for 78 different networks. Comparing the results with the forecast for winter season 2015/2016 (see Table 5.2) we see that an increase of nodes in the hidden layers is no guarantee for better forecasting results and the stability of the neural network. On can see that for both seasons a neural network with 2 hidden layers and 2 nodes in the first layer and 1 node in the second layer delivers good forecasting results. There are more complex networks which deliver better results of each single season but for both this neural networks looks reasonable. The final network is shown in Figure 5.3. The

results are similar to the results of the tree and random forest models.
For modeling the binary variable **überschreitung** we will apply the similar procedure to get a reasonable model. We will reduce the number of possible nodes in the first hidden layer to 8 and in the second to the half of the amount in the first hidden layer. The results of forecast of the neural networks are very poor for both seasons. Therefore those models will not discussed anymore in the following.

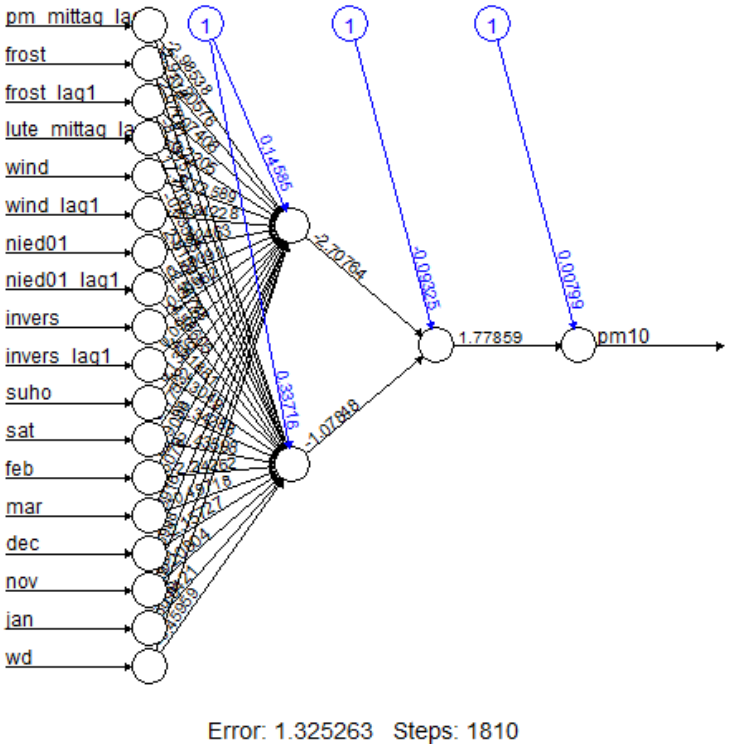

Figure 5.3: Neural network for Graz-Mitte

5.3 Neural network for Graz-Süd

To the data set of Graz-Süd we apply the same algorithm as for Graz-Mitte. In contrast to the results of Graz-Mitte it is not obvious, that a higher number of nodes in the hidden layer results in worse forecasting results. The neural network with 7 nodes in the first hidden layer and 1 node in the second hidden layer gives a valid forecasting result for both winter seasons (compare Table 5.3 and Table 5.4). While the overall misclassification error is low, the quality of more than 10% in the winter season 2015/2016 of the forecasting results is very bad. But the quality of a huge part is excellent and good. For the model with 8 nodes in the first hidden layer and 7 nodes in the second hidden layer the algorithm could not find an optimal solution, therefore no results can be calculated.
By comparing the observed PM10 concentration with the forecasted values for the winter

season 2014-2015 in Figure 5.4 we see that the models overestimate much more observations than the regression model. For the winter season 2015/2016 the results looks very simliar. (Figure 5.5).

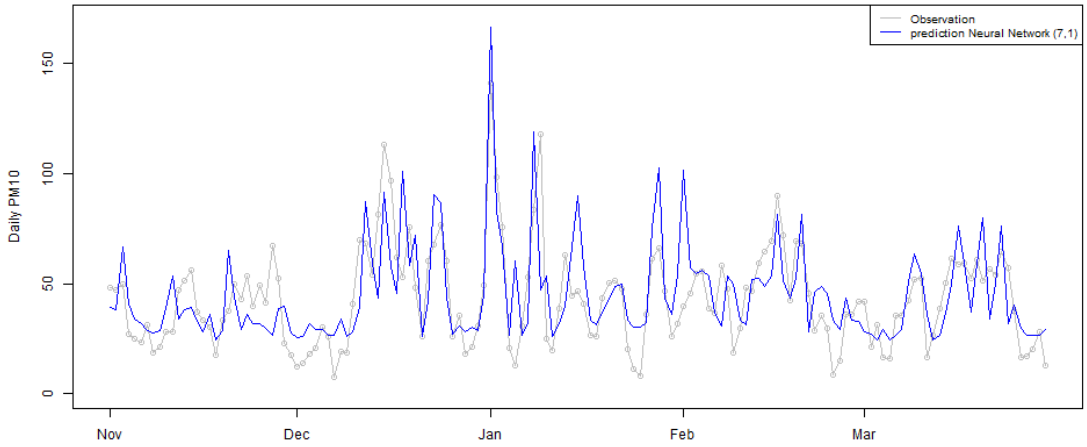

Figure 5.4: Comparison of observed and predicted values for the winter season 2014/2015

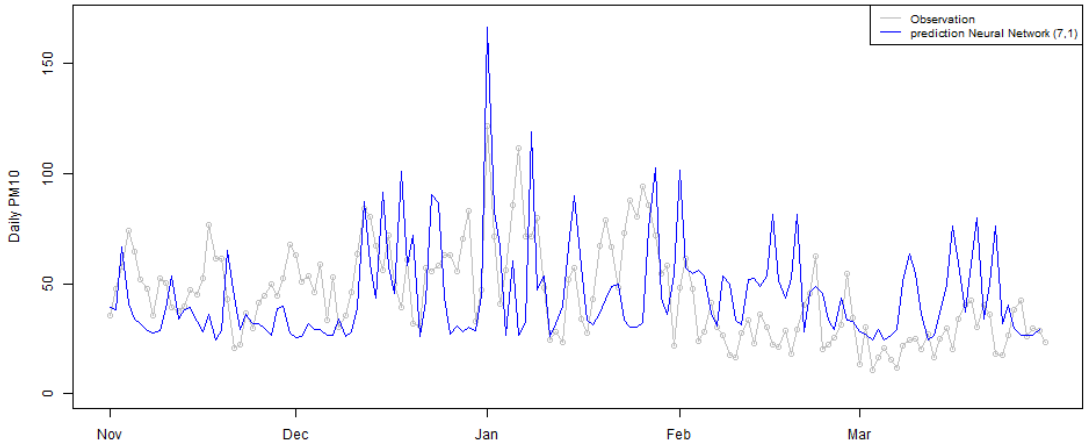

Figure 5.5: Comparison of observed and predicted values for the winter season 2015/2016

5.4 Conclusion

While the neural network algorithm delivers remarkable results for both sites by modeling the average **PM_10** concentration directly give the tree models and forest models good indications for a possible exceendance of the critical limit. The tree and forest model are very

simple and can be used in addition to the current model. The neural network uses much more input variables and is very complex to deal with. As we can see the stability of the neural networks is not always given. Therefore the network models can not replace the current linear regression model used for the one-day ahead prediction.

Table 5.1: Forecasting Results of the Neural networks for the Winter Season 2014/2015 in Graz-Mitte

#	# nodes in Hidden Layer		# of wrong classifications			Quality Function II				
	1	2	0-50 $\mu g/m^3$	> 50 $\mu g/m^3$	Total	excellent	good	satisfying	bad	very bad
1	1	0	2	9	11	119	21	1	0	10
2	2	0	2	9	11	121	17	2	0	11
3	**2**	**1**	**2**	**8**	**10**	**121**	**20**	**1**	**0**	**9**
4	3	0	0	9	9	120	18	2	1	10
5	3	1	6	8	14	118	21	1	1	10
6	3	2	8	8	16	115	20	2	3	11
7	4	0	7	9	16	110	25	4	1	11
8	4	1	1	10	11	120	18	2	0	11
9	4	2	6	8	14	116	22	4	0	9
10	4	3	1	8	9	118	22	2	1	8
11	5	0	7	8	15	105	29	5	2	10
12	5	1	2	9	11	126	13	1	0	11
13	5	2	1	7	8	120	22	0	0	9
14	5	3	2	8	10	116	21	5	0	9
15	5	4	4	9	13	116	21	2	1	11
16	6	0	3	8	11	118	19	2	1	11
17	6	1	3	8	11	114	25	1	0	11
18	6	2	5	8	13	111	25	3	1	11
19	6	3	10	7	17	111	25	4	2	9
20	6	4	10	9	19	116	14	5	3	13
21	6	5	4	9	13	117	19	3	0	12
22	7	0	7	7	14	118	18	2	3	10
23	7	1	5	9	14	112	23	2	2	12
24	7	2	9	8	17	115	22	2	2	10
25	7	3	11	7	18	105	26	6	2	12
26	7	4	2	9	11	116	23	1	0	11
27	7	5	7	9	16	111	25	2	2	11
28	7	6	4	9	13	105	33	3	0	10
29	8	0	5	8	13	111	26	2	1	11
30	8	1	7	9	16	106	28	1	5	11

#	# nodes in Hidden Layer		# of wrong classifications			Quality Function II				
			0-50	> 50						
	1	2	$\mu g/m^3$	$\mu g/m^3$	Total	excellent	good	satisfying	bad	very bad
31	8	2	8	8	16	101	30	6	4	10
32	8	3	11	9	20	95	33	3	1	19
33	8	4	8	8	16	100	34	4	3	10
34	8	5	7	8	15	107	28	4	2	10
35	8	6	10	8	18	108	23	4	6	10
36	8	7	9	6	15	109	29	3	2	8
37	9	0	12	9	21	98	32	7	3	11
38	9	1	7	8	15	113	21	6	0	11
39	9	2	9	9	18	108	26	3	3	11
40	9	3	9	7	16	107	27	5	1	11
41	9	4	7	9	16	111	23	1	3	13
42	9	5	9	8	17	98	39	2	1	11
43	9	6	7	8	15	109	24	5	3	10
44	9	7	13	8	21	102	23	11	1	14
45	9	8	15	7	22	110	18	9	4	10
46	10	0	11	8	19	101	30	5	3	12
47	10	1	11	8	19	98	34	5	3	11
48	10	2	7	7	14	116	21	1	1	12
49	10	3	5	8	13	109	30	0	2	10
50	10	4	14	7	21	97	30	7	2	15
51	10	5	19	8	27	98	25	9	3	16
52	10	6	17	8	25	91	35	6	6	13
53	10	7	7	8	15	105	27	7	1	11
54	10	8	9	7	16	102	31	3	3	12
55	10	9	6	8	14	100	32	8	0	11
56	11	0	10	9	19	96	32	5	7	11
57	11	1	12	9	21	111	20	4	1	15
58	11	2	8	8	16	104	30	4	1	12
59	11	3	3	9	12	115	24	1	1	10
60	11	4	11	10	21	106	23	5	3	14
61	11	5	10	6	16	110	25	5	2	9
62	11	6	15	8	23	89	34	6	7	15
63	11	7	16	9	25	99	22	12	5	13
64	11	8	8	8	16	108	26	4	4	9
65	11	9	10	6	16	106	26	9	0	10
66	11	10	11	7	18	97	31	9	0	14
67	12	0	10	9	19	95	31	10	5	10

#	# nodes in Hidden Layer		# of wrong classifications			Quality Function II				
			0-50	> 50						
	1	2	$\mu g/m^3$	$\mu g/m^3$	Total	excellent	good	satisfying	bad	very bad
68	12	1	10	8	18	97	33	6	4	11
69	12	2	5	7	12	117	19	3	1	11
70	12	3	10	9	19	107	24	4	2	14
71	12	4	7	7	14	110	25	3	5	8
72	12	5	6	7	13	110	26	5	3	7
73	12	6	10	7	17	103	27	4	3	14
74	12	7	12	8	20	97	30	4	3	17
75	12	8	13	7	20	102	27	6	3	13
76	12	9	13	8	21	95	32	6	3	15
77	12	10	9	8	17	102	30	4	4	11
78	12	11	11	7	18	96	33	3	3	16

Table 5.2: Forecasting Results of the Neural networks for the Winter Season 2015/2016 in Graz-Mitte

#	# nodes in Hidden Layer		# of wrong classifications			Quality Function II				
			0-50	> 50						
	1	2	$\mu g/m^3$	$\mu g/m^3$	Total	excellent	good	satisfying	bad	very bad
1	1	0	5	12	17	114	18	7	2	11
2	2	0	5	11	16	118	17	3	3	11
3	**2**	**1**	**4**	**12**	**16**	**115**	**18**	**5**	**2**	**12**
4	3	0	7	15	22	111	17	5	3	16
5	3	1	7	13	20	111	17	4	3	17
6	3	2	11	12	23	107	21	5	3	16
7	4	0	13	12	25	101	25	5	6	15
8	4	1	5	15	20	108	19	7	1	17
9	4	2	9	9	18	109	23	4	3	13
10	4	3	3	16	19	112	17	4	1	18
11	5	0	8	15	23	107	19	5	3	18
12	5	1	7	16	23	110	21	1	3	17
13	5	2	13	16	29	104	21	7	2	18
14	5	3	11	10	21	105	26	3	7	11
15	5	4	10	14	24	98	28	5	2	19
16	6	0	15	15	30	95	23	9	7	18
17	6	1	5	16	21	105	23	3	4	17

#	# nodes in Hidden Layer		# of wrong classifications			Quality Function II				
	1	2	0-50 $\mu g/m^3$	> 50 $\mu g/m^3$	Total	excellent	good	satisfying	bad	very bad
18	6	2	16	15	31	91	29	7	8	17
19	6	3	12	11	23	98	30	8	2	14
20	6	4	20	13	33	98	20	7	5	22
21	6	5	7	15	22	101	25	5	2	19
22	7	0	8	15	23	100	26	4	4	18
23	7	1	13	16	29	103	20	6	4	19
24	7	2	8	13	21	108	17	6	4	17
25	7	3	7	14	21	101	24	8	3	16
26	7	4	9	16	25	97	28	5	2	20
27	7	5	6	17	23	103	24	4	2	19
28	7	6	6	15	21	107	20	6	2	17
29	8	0	11	12	23	99	27	7	3	16
30	8	1	13	13	26	98	25	5	5	19
31	8	2	8	15	23	101	25	3	3	20
32	8	3	19	15	34	88	27	7	5	25
33	8	4	5	11	16	103	21	11	4	13
34	8	5	14	12	26	95	27	6	4	20
35	8	6	9	17	26	103	20	8	2	19
36	8	7	11	18	29	93	33	2	4	20
37	9	0	13	14	27	95	27	4	7	19
38	9	1	14	11	25	98	26	8	5	15
39	9	2	13	14	27	94	28	6	5	19
40	9	3	11	14	25	92	31	7	5	17
41	9	4	16	13	29	99	22	4	5	22
42	9	5	12	13	25	99	25	2	6	20
43	9	6	10	14	24	95	27	5	4	21
44	9	7	8	13	21	98	29	5	3	17
45	9	8	10	11	21	97	30	6	4	15
46	10	0	10	16	26	91	32	4	5	20
47	10	1	20	15	35	91	24	3	8	26
48	10	2	8	17	25	101	25	4	2	20
49	10	3	5	16	21	99	29	1	5	18
50	10	4	20	14	34	78	37	5	7	25
51	10	5	14	10	24	92	29	7	8	16
52	10	6	16	11	27	98	26	6	3	19
53	10	7	7	16	23	105	22	5	1	19
54	10	8	8	18	26	98	25	4	3	22

#	# nodes in Hidden Layer		# of wrong classifications			Quality Function II				
			0-50	> 50						
	1	2	$\mu g/m^3$	$\mu g/m^3$	Total	excellent	good	satisfying	bad	very bad
55	10	9	15	14	29	90	27	9	5	21
56	11	0	10	13	23	94	30	7	3	18
57	11	1	15	9	24	105	22	6	7	12
58	11	2	14	14	28	102	20	3	6	21
59	11	3	7	15	22	103	25	2	6	16
60	11	4	16	13	29	96	25	8	3	20
61	11	5	11	16	27	91	31	5	4	21
62	11	6	14	14	28	89	26	11	8	18
63	11	7	14	9	23	92	34	9	4	13
64	11	8	10	15	25	96	24	5	5	22
65	11	9	24	11	35	86	26	10	8	22
66	11	10	14	14	28	98	20	5	7	22
67	12	0	19	12	31	89	27	9	3	24
68	12	1	14	13	27	96	27	4	4	21
69	12	2	7	18	25	99	27	4	1	21
70	12	3	9	18	27	97	22	6	3	24
71	12	4	15	14	29	91	19	16	3	23
72	12	5	16	14	30	82	34	9	4	23
73	12	6	14	13	27	86	35	7	7	17
74	12	7	17	13	30	94	25	3	6	24
75	12	8	14	13	27	90	29	8	6	19
76	12	9	18	11	29	85	30	9	10	18
77	12	10	19	11	30	89	29	9	9	16
78	12	11	13	14	27	88	32	2	4	26

Table 5.3: Forecasting Results of the Neural networks for the Winter Season 2014/2015 in Graz-Süd

#	# nodes in Hidden Layer		# of wrong classifications			Quality Function II				
			0-50	> 50						
	1	2	$\mu g/m^3$	$\mu g/m^3$	Total	excellent	good	satisfying	bad	very bad
1	1	0	10	7	17	101	33	4	3	10
2	2	0	9	8	17	92	42	5	2	10
3	2	1	12	8	20	96	34	8	3	10
4	3	0	9	8	17	102	32	6	1	10

#	# nodes in Hidden Layer		# of wrong classifications			Quality Function II				
			0-50	> 50						
	1	2	$\mu g/m^3$	$\mu g/m^3$	Total	excellent	good	satisfying	bad	very bad
5	3	1	11	9	20	99	35	4	1	12
6	3	2	11	9	20	101	29	5	3	13
7	4	0	10	8	18	102	32	5	1	11
8	4	1	16	7	23	93	37	8	3	10
9	4	2	11	7	18	103	31	6	2	9
10	4	3	11	7	18	96	36	7	1	11
11	5	0	9	8	17	108	26	4	2	11
12	5	1	10	9	19	95	35	4	5	12
13	5	2	8	9	17	103	31	6	0	11
14	5	3	10	9	19	98	36	4	2	11
15	5	4	16	8	24	79	49	8	3	12
16	6	0	12	8	20	97	36	5	3	10
17	6	1	11	8	19	100	33	4	3	11
18	6	2	17	8	25	91	37	8	3	12
19	6	3	13	5	18	99	31	9	3	9
20	6	4	8	9	17	96	38	4	0	13
21	6	5	15	7	22	84	45	10	2	10
22	7	0	14	7	21	99	34	7	3	8
23	**7**	**1**	**10**	**8**	**18**	**103**	**28**	**6**	**3**	**11**
24	7	2	12	8	20	101	29	7	3	11
25	7	3	12	8	20	93	40	4	4	10
26	7	4	9	8	17	99	33	5	2	12
27	7	5	10	9	19	98	33	4	2	14
28	7	6	12	9	21	104	24	6	6	11
29	8	0	11	8	19	100	32	5	3	11
30	8	1	14	8	22	98	32	4	7	10
31	8	2	9	8	17	103	28	4	4	12
32	8	3	13	7	20	101	31	7	2	10
33	8	4	16	7	23	95	36	3	6	11
34	8	5	13	9	22	96	33	5	4	13
35	8	6	14	8	22	99	30	7	4	11
36	9	0	8	9	17	100	35	3	2	11
37	9	1	11	7	18	91	44	4	2	10
38	9	2	11	9	20	97	35	3	4	12
39	9	3	11	8	19	90	40	7	2	12
40	9	4	13	8	21	96	37	2	4	12
41	9	5	10	9	19	105	27	4	2	13

#	# nodes in Hidden Layer		# of wrong classifications			Quality Function II				
			0-50	> 50						
	1	2	$\mu g/m^3$	$\mu g/m^3$	Total	excellent	good	satisfying	bad	very bad
42	9	6	16	5	21	93	37	6	4	11
43	9	7	11	7	18	99	32	7	2	11
44	9	8	17	7	24	102	26	4	7	12
45	10	0	12	8	20	108	23	3	5	12
46	10	1	10	7	17	104	30	2	5	10
47	10	2	10	8	18	102	33	4	1	11
48	10	3	13	8	21	100	28	6	2	15
49	10	4	11	7	18	106	24	6	4	11
50	10	5	11	11	22	104	24	4	3	16
51	10	6	17	7	24	92	32	6	3	18
52	10	7	12	8	20	101	31	7	0	12
53	10	8	9	8	17	98	33	6	4	10
54	10	9	12	6	18	103	27	6	7	8
55	11	0	16	5	21	96	36	5	6	8
56	11	1	18	8	26	95	30	9	4	13
57	11	2	16	9	25	91	37	6	4	13
58	11	3	11	8	19	97	35	4	4	11
59	11	4	7	9	16	109	27	1	3	11
60	11	5	12	9	21	101	29	5	4	12
61	11	6	16	6	22	95	32	10	4	10
62	11	7	17	8	25	100	27	10	3	11
63	11	8	16	7	23	91	36	5	7	12
64	11	9	18	8	26	92	32	9	5	13
65	11	10	13	8	21	92	35	6	5	13
66	12	0	13	6	19	107	22	7	5	10
67	12	1	8	9	17	103	32	2	2	12
68	12	2	13	8	21	94	36	5	4	12
69	12	3	14	8	22	99	30	4	6	12
70	12	4	9	10	19	96	36	3	4	12
71	12	5	11	7	18	101	30	5	2	13
72	12	6	17	7	24	89	34	14	3	11
73	12	7	19	9	28	91	32	7	10	11
74	12	8	16	9	25	97	26	7	4	17
75	12	9	17	10	27	93	29	9	4	16
76	12	10	13	5	18	94	39	5	5	8
77	12	11	14	9	23	100	28	6	5	12

Table 5.4: Forecasting Results of the Neural networks for the Winter Season 2015/2016 in Graz-Süd

#	# nodes in Hidden Layer		# of wrong classifications			Quality Function II				
	1	2	0-50 $\mu g/m^3$	> 50 $\mu g/m^3$	Total	excellent	good	satisfying	bad	very bad
1	1	0	18	7	25	91	36	11	6	8
2	2	0	17	8	25	87	40	9	6	10
3	2	1	24	8	32	87	31	14	10	10
4	3	0	19	7	26	93	34	8	9	8
5	3	1	19	7	26	91	35	8	8	10
6	3	2	23	6	29	89	33	10	12	8
7	4	0	18	7	25	92	36	8	7	9
8	4	1	20	7	27	89	36	13	6	8
9	4	2	22	7	29	89	31	13	10	9
10	4	3	23	6	29	83	37	14	11	7
11	5	0	19	6	25	92	34	7	9	10
12	5	1	14	9	23	94	33	6	8	11
13	5	2	20	6	26	91	34	13	6	8
14	5	3	19	8	27	87	34	11	10	10
15	5	4	24	7	31	83	35	17	8	9
16	6	0	19	7	26	89	35	7	9	12
17	6	1	22	7	29	92	28	15	7	10
18	6	2	19	5	24	90	37	11	4	10
19	6	3	25	6	31	85	33	15	9	10
20	6	4	16	9	25	92	35	9	5	11
21	6	5	24	6	30	80	37	18	9	8
22	7	0	22	7	29	87	34	12	10	9
23	**7**	**1**	**14**	**8**	**22**	**91**	**35**	**6**	**4**	**16**
24	7	2	24	7	31	89	34	12	9	8
25	7	3	27	9	36	83	32	16	8	13
26	7	4	23	10	33	83	33	12	10	14
27	7	5	20	11	31	92	23	10	12	15
28	7	6	28	5	33	86	30	14	13	9
29	8	0	18	9	27	92	31	8	7	14
30	8	1	21	7	28	98	21	14	7	12
31	8	2	19	10	29	88	31	9	6	18
32	8	3	22	6	28	88	31	13	5	15
33	8	4	15	8	23	96	31	8	5	12
34	8	5	22	5	27	92	27	12	10	11

#	# nodes in Hidden Layer		# of wrong classifications			Quality Function II				
	1	2	0-50 $\mu g/m^3$	> 50 $\mu g/m^3$	Total	excellent	good	satisfying	bad	very bad
35	8	6	28	6	34	83	34	15	11	9
36	9	0	17	7	24	100	25	6	9	12
37	9	1	25	9	34	82	37	8	10	15
38	9	2	16	9	25	99	28	6	6	13
39	9	3	27	7	34	79	37	11	11	14
40	9	4	20	10	30	87	30	16	7	12
41	9	5	27	10	37	82	35	8	9	18
42	9	6	16	6	22	96	28	6	9	13
43	9	7	20	8	28	86	36	6	10	14
44	9	8	19	6	25	89	36	9	6	12
45	10	0	21	6	27	92	33	8	9	10
46	10	1	19	8	27	88	34	10	7	13
47	10	2	19	11	30	93	27	9	6	17
48	10	3	29	9	38	81	29	15	7	20
49	10	4	26	7	33	85	31	12	10	14
50	10	5	22	8	30	77	44	13	4	14
51	10	6	23	4	27	91	26	9	7	19
52	10	7	20	6	26	92	35	7	9	9
53	10	8	24	11	35	84	31	9	14	14
54	10	9	18	7	25	93	33	9	5	12
55	11	0	26	8	34	90	26	9	14	13
56	11	1	24	6	30	87	32	13	8	12
57	11	2	21	8	29	89	32	10	6	15
58	11	3	15	10	25	93	33	5	7	14
59	11	4	19	11	30	94	27	8	7	16
60	11	5	21	12	33	83	36	12	4	17
61	11	6	18	8	26	86	38	9	5	14
62	11	7	26	6	32	92	27	10	6	17
63	11	8	24	8	32	83	30	13	11	15
64	11	9	23	7	30	87	30	13	7	15
65	11	10	23	7	30	81	31	13	12	15
66	12	0	20	10	30	89	26	11	12	14
67	12	1	17	9	26	93	30	9	5	15
68	12	2	17	9	26	86	37	10	4	15
69	12	3	24	7	31	86	32	12	12	10
70	12	4	20	8	28	85	37	10	8	12
71	12	5	23	11	34	79	34	10	12	17

#	# nodes in Hidden Layer		# of wrong classifications			Quality Function II				
	1	2	0-50 $\mu g/m^3$	> 50 $\mu g/m^3$	Total	excellent	good	satisfying	bad	very bad
72	12	6	21	4	25	88	34	9	13	8
73	12	7	34	6	40	80	27	13	19	13
74	12	8	18	7	25	88	37	6	6	15
75	12	9	19	8	27	87	39	5	7	14
76	12	10	24	8	32	83	31	13	8	17
77	12	11	20	9	29	90	31	7	8	16

Bibliography

[1] Berk R. A. *Statistical Learning from a Regression Perspective*. Springer, New York, 2008.

[2] Breiman L. Bagging Predictors. *Machine Learning*, 24:123–140, 1996.

[3] Breiman L. Technical Note: Some Properties of Splitting Criteria. *Machine Learning*, 24:41–47, 1996.

[4] Breiman L. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[5] Breiman L., Friedman J. H., Olshen A. R., Stone J. C. *Classification and Regression Trees*. Wadswoerth, California, 1984.

[6] Chen Y., Shi R., Shu S., Gao W. Ensemble and enhanced PM10 concentration forecast model based on stepwise regression and wavelet analysis. *Atmospheric Environment*, 74:346–359, 2013.

[7] Domanska D., Wojtylak M. Explorative forecasting of air pollution. *Atmospheric Environment*, 92:19–30, 2014.

[8] G. Louppe G., Wehenkel L., Sutera A., Geurts P. Understanding variable importances in forests of randomized trees. *Advances in Neural Information Processing Systems*, page 431439, 2013.

[9] Hastie T., Tibshirani R., Friedman J. *The Elements of Statistical Learning*. Springer, New York, 2001.

[10] Hörmann S., Pfeiler B., Stadlober E. Analysis and Prediction of Particulate Matter PM10 for the Winter Season in Graz. *Austrian Journal of Statistics*, 34(4):307–326, 2005.

[11] Hörmann S., Stadlober E. Modellierung und Prognose der Feinstaubkonzentration in Graz-Mitte für den Zeitraum Herbst/Winter. Technical report, Institut für Statistuk, TU-Graz, 2004. `https://www.stat.tugraz.at/stadl/papers/host04.pdf`.

[12] Hothorn T., Hornik K., Zeileis A. Unbiased recursive partitioning: a conditional inference framework. *Journal of Computational and Graphical Statistics*, 15(3):651–674, 2006.

[13] Kim H., Loh W-Y. Classification Trees With Unbiased Multiway Splits. *Journal of the American Statistical Association*, 96:598–604, 2001.

[14] Koo Y-S., Kim J-S., Cho J-S., Jang Y-K. Performance evaluation of the updated air quality forecasting system for Seoul predicting PM10. *Atmospheric Environment*, 58:59–69, 2012.

[15] Lantz B. *Machine Learning with R*. Packt Publishing Ltd., Birmingham, 2013.

[16] Loh W-Y. Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1:14–23, 2011.

[17] Loh W-Y., Shih Y-S. Spit Selection Methods for Classification Trees. *Statistica Sinica*, 7:815–840, 1997.

[18] Muhamad M., Ul-Saufie Z. A. , Deni M. S. Three Days Ahead Prediction of Daily 12 Hour Ozone (O_3) Concentrations for Urban Area in Malaysia. *Journal of Environmental Science and Technology*, 8:102–112, 2015.

[19] Perez P. Combined model for PM10 forecasting in a large city. *Atmospheric Environment*, 60:271–276, 2012.

[20] Pires M. C. J., Martins G. F., Sousa V. I. S., Alvim-Ferraz M. C. M., Pereira C. M. Prediction of the Daily Mean PM10 Concentrations Using Linear Models. *American Journal of Environmental Sciences*, 4:445–453, 2008.

[21] Poggi J-M. ,Portier B. PM10 forecasting using clusterwise regression. *Atmospheric Environment*, 45:2005–2014, 2011.

[22] Saide E. P., Carmichael R. G., Spak N. S., Gallardo L., Osses E. A., Mena-Carrasco A. M., Pagowski M. Forecasting urban PM10 and PM2.5 pollution episodes in very stable nocturnal conditions and complex terrain using WRFeChem CO tracer model. *Atmospheric Environment*, 45:2769–2780, 2011.

[23] Stadlober E., Hörmann S., Pfeiler B. Quality and performance of a PM10 daily forecasting model. *Atmospheric Environment*, 42:1098–1109, 2008.

[24] Stadlober E., Hübnerova Z., Michalek J., Kolar M. Forecasting of Daily PM10 Concentration in Brno and Graz by different Regression Approaches. *Austrian Journal of Statistics*, 41(4):287–310, 2012.

[25] Stadlober E., Pfeiler B. Ermittlung von saisonbereinigten PM10-Mittelwerten des Winterhalbjahres auf Basis der Daten 2003/2004 - 2014/2015. Technical report, Institut für Statistuk, TU-Graz, 2016. Endbericht Februar 2016.

[26] Strobl C., Boulesteix A-L., Augustin T. Unbiased split selection for classification trees based on the Gini Index. Technical report, Istitut für Statistik, LMU München, 2005.

[27] Therneau M.T., Atkinson J.E. An Introduction to Recursive Partitioning Using the RPART Routines. Technical report, Mayo Foundation, 2015.

[28] Ul-Saufie Z. A., Yahaya S. A., Ramli A. N., Rosaida N., Hamid A. N. Future daily PM10 concentrations prediction by combining regression models and feedforward backpropagation models with principle component analysis (PCA). *Atmospheric Environment*, 77:621–630, 2013.

[29] Westerlund J., Urbain J-P., Bonilla J. Application of air quality combination forecasting to Bogota. *Atmospheric Environment*, 80:22–28, 2014.

[30] Yusof F. N., Ramli A.N., Yahaya S. A.,Sansuddin N., Ghazali A. N., al Madhoun W. Monsoonal differences and probability distribution of PM10 concentration. *Enviromental Monitoring and Assessment*, 163:655–667, 2010.

[31] Zhang Y., Bocquet M., Mallet V., Seigneur C., Baklanov A. Real-time air quality forecasting, part I: History, techniques, and current status. *Atmospheric Environment*, 60:632–655, 2012.

[32] Zhang Y., Bocquet M., Mallet V., Seigneur C., Baklanov A. Real-time air quality forecasting, part II: State of the science, current research needs, and future prospects. *Atmospheric Environment*, 60:656–676, 2012.