KAUNAS UNIVERSITY OF TECHNOLOGY
FACULTY OF FUNDAMENTAL SCIENCE
DAPERTMENT OF APPLIED MATHEMATICS

Julija Vlasova

SPATIO-TEMPORAL ANALYSIS OF WIND POWER PREDICTION ERRORS

Master project

Supervisors:
prof. A. Aksomaitis
prof. H. Madsen

Kaunas, 2007

KAUNAS UNIVERSITY OF TECHNOLOGY
FACULTY OF FUNDAMENTAL SCIENCE
DEPARTMENT OF APPLIED MATHEMATICS

Head of the Department:
prof. dr. J. Rimas

SPATIO-TEMPORAL ANALYSIS OF WIND POWER PREDICTION ERRORS

Master project in Applied Mathematics

Supervisors:
prof. A. Aksomaitis
prof. H. Madsen

Reviewer
dr. Virginijus Radziukynas

Student
FMMM 5 gr. stud.
J. Vlasova

Kaunas, 2007

ABSTRACT

Nowadays there is no need to convince anyone about the necessity of renewable energy. One of the most promising ways to obtain it is the wind power. Countries like Denmark, Germany or Spain proved that, while professionally managed, it can cover a substantial part of the overall energy demand. One of the main and specific problems related to the wind power management — development of the accurate power prediction models. Nowadays State-Of-Art systems provide predictions for a single wind turbine, wind farm or a group of them. However, the spatio-temporal propagation of the errors is not adequately considered.

In this paper the potential for improving modern wind power prediction tool WPPT, based on the spatio-temporal propagation of the errors, is examined. Several statistical models (Linear, Threshold, Varying-coefficient and Conditional Parametric) capturing the cross-dependency of the errors, obtained in different parts of the country, are presented. The analysis is based on the weather forecast information and wind power prediction errors obtained for the territory of Denmark in the year 2004.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Secure energy supplies are vital for the well functioning and competitiveness of the European economy. However, most observers agree that the era of cheap and unlimited supplies of fossil energy is over. The EU's dependency on imported fossil fuel energy has become a threat to economic stability. The rise in prices has coincided with other important trends: fears over real level of known and realizable oil and gas reserves, fears over the security of supply coming from politically unstable regions and fear over adequacy of refining capacity to satisfy demand, particularly in the medium to long-term. The world has moved from the era of energy security to one of energy insecurity, which coincides with rising concern over climate change, where limiting carbon from fossil fuels is at the heart of climate change policies [14].

The way out from the problematic situation - renewable energy. Europe is the world leader in renewable energy and in the most promising and mature renewable technology. Wind energy will not only be able to contribute to securing European energy independence and climate goals in the future, it could also turn a serious supply problem into an opportunity for Europe in the forms of technological research, exports and employment. Wind energy currently provides 2.4% of European electricity, a figure that could rise to 12% by 2020 and over 22% in 2030 according the plan for EU renewable energy development [14]

1.1 Problem formulation and objectives of the work

Forecasts of wind power generation are more and more frequently used in various management tasks related to integration of wind generation in power systems. In order to be able to absorb a large fraction of wind power in the electrical systems, reliable short-term (say 36 hours) predictions of the future wind power generation are needed.

The quality of the forecast is very important, and a reliable estimate of the uncertainty of the forecast is known to be essential.

Today the forecast of wind power generation is provided without a proper consideration to the spatio-temporal dependencies observed in the wind power generation field. The state-of-art prediction systems typically provide forecasts for a single wind farm or a larger region with a number of wind turbines or wind farms. However, the spatio-temporal relations are not adequately considered.

In this master project the errors from WPPT (Wind Power Prediction Tool), which is one of the leading systems, will been examined. The primary aim of the work is to investigate whether there is a potential for improving predictions based on the spatio-temporal dependency of errors. If the potential presents, then it is of high interest to investigate it by capturing the nature of error propagation, creating new models and methods for improving the short-term wind power predictions.

1.2 Outline of the thesis

Chapter 2 introduces the data used in the thesis and describes the way of managing it before the modeling part.

In Chapter 3 the pre-modeling analysis of the data structure is provided. The results of the correlation study are presented and discussed.

Chapter 4 deals with the modeling part. It consists of the four sections (4.1-4.4), each corresponding to the particular model type (Linear, Threshold, Varying-coefficient and Conditional parametric models respectively). Every section begins with Modelling and Estimation sub-sections, followed by Application and Results. The first two sections are theoretical, presenting the general form of the models, common estimation procedures, while the last two introduce the details of applying the models to the data and discuss the obtained results.

Chapter 5 describes validation methods used in this study for checking the adequacy of the performance of the fitted models.

In Chapter 6 the overall conclusions according to the thesis objectives are presented and the experiences from the implementation of the models are summarized into a number of recommendations for the possible future work.

Chapter 2

Data Description

2.1 Data

The data selected for this work comes from 24 wind farms owned by ELSAM where Wind Power Prediction Tool (WPPT) has been used to make forecasts of the power production based on information from the Danish Meteorological Institute (DMI) HIRLAM model. The power production is recorded at one hour intervals matching the temporal resolution of the HIRLAM predictions. A new version of WPPT was installed in the fall of 2003, as several parameters are estimated adaptively some time is needed for the model to burn in. Therefore, it was decided to disregard data from before 01-01-2004. For this project we use data from the first seven months of 2004.

The HIRLAM data is delivered in a 40 by 42 grid covering Denmark and surroundings in particular a large part of the North Sea. Every six hours a new 48 hour forecast with one hour steps is calculated. It takes 2-3 hours to calculate the forecast so in practice a 46 hour forecast is produced. Wind prediction in different levels of the HIRLAM model are available, but in this work we use wind at height 10 and 3000 meters a.g.l. (above ground level) only. The motivation for this choice is that the 10m wind minds local landscape characteristics while the 3000m is considered as a level where the wind is "undisturbed". However, after correlation analysis of the data was performed (see Section 3 for details) the 10 meters a.g.l. wind showed better performance and was chosen for the modelling step.

The WPPT model contains two parts: one, describing the power curve including dependence on wind speed and direction, and a dynamical part, which includes the estimates of autoregression. While calculating the prediction errors, it was chosen to use the output from the combined performance of the two parts. The errors were created as the difference between the power predictions and productions normalized by the installed wind power.

2.2 Managing the data

In order to reduce the influence of local behavior and to concentrate on global phenomena it was decided to group the data according to the location and correlation in errors in $lag\,0$. Following groups were selected:

Group No	farms included
1	wab, war, wbs, wrb, whoe, wtj
2	wbi, wfj, wve, wvm
3	wgv, woek, wnv, wnr, wkm, whh
4	wto, wrv, wtu, who
5	wdr, wdg

Table 2.1: Wind farm groups

The location of the mentioned farms and groups is given in the Figure 2.1. As can be noticed, only data from the 22 farms is used. The remaining two farms were excluded at this point of managing the data since the correlation study showed that they can not be reasonably pooled into the same groups together with the other farms.

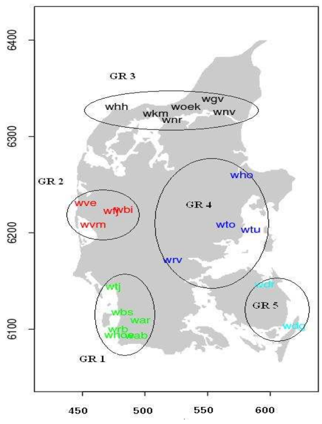

Figure 2.1: Selected groups of wind farms

The group errors were calculated as an average of the errors within the groups. In the same way the group wind speed was calculated. For the directions in groups the geometrical approach was chosen. Wind direction at each of the farms was presented as a vector. The resultant vector (corresponding angle) was taken to represent the wind direction for that group.

Chapter 3

Identification of the Model Structure

In the next three chapters we introduce respectively: model identification, estimation and validation which are the three stages in model building process ([11]). Current chapter deals with the first of them. Tools like (Auto-) Correlation and Partial (Auto-) Correlation Functions will be presented, and the results analyzed. The following notation is introduced:

ACF - Auto-Correlation Function
PACF - Partial Auto-Correlation Function
CCF - Cross-Correlation Function
PCCF - Partial Cross-Correlation Function

The main purpose of the correlation analysis is to identify the structure of the data. We apply this technique to choose the most proper modelling direction. The main questions at this point are:

1. Is there a significant linear dependency within and between the groups?

2. Do the variables wind direction and wind speed influence the strength of this dependency?

Below, we try to deliver the answers.

3.1 Dependency within the groups

Firstly we investigate the influence that the previous values of each time series have on its current state. Namely, we apply here ACF and PACF. Remember that autocorrelation is the correlation of the process against a time-shifted version of itself. Assigning our time series as X_t

which is assumed to be stationary, the formula for the ACF in lag k is given below:

$$ACF(k) = Cor[X_t, X_{t-k}] = \frac{E[(X_t - \mu)(X_{t-k} - \mu)]}{\sigma^2} \tag{3.1}$$

where μ is the mean of the time series X_t and σ is its standard deviation. The coefficient takes values in the interval $[-1, 1]$. Obviously for $k = 0$ it is equal to unity. The problem with ACF is that since the estimates of autocorrelations are correlated within themselves, the pattern at the lower lags can be propagated on the larger lags. In order to remove the influence of the lower lags, we use PACF. This is given by the following formula:

$$PACF(k) = Cor[X_t, X_{t-k}|X_{t-1}, ..., X_{t-k+1}] = Cor(R_{X_t}, R_{X_{t-k}}) \tag{3.2}$$

which shows the "real" dependency between X_t and X_{t-k} excluding the influence of all the observations in between. It can be treated as correlation of R_{X_t} and $R_{X_{t-k}}$, which are residuals after regressing respectively X_t and X_{t-k} on $X_{t-1}, ..., X_{t-k+1}$. Plots 3.1 and 3.2 show the ACF and PACF for Group 5.

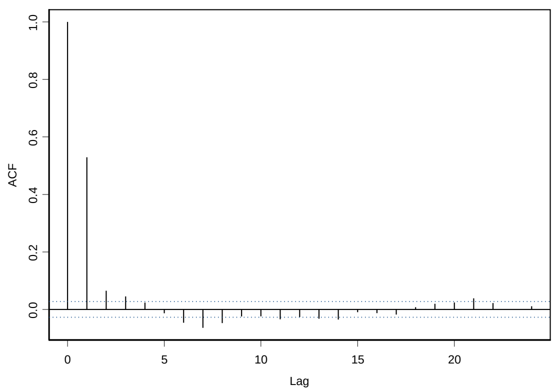

Figure 3.1: ACF for Group 5

3.2 Dependency between the groups

Once the Auto-correlation of the group errors is identified, we proceed with investigation of the cross-dependencies among the data. The common practice is to check if the errors at different farm groups are cross-correlated. Especially the information about the dependency in lags larger than 0 would be of great importance, regarding the future model forecasting ability. If such pattern is discovered, we can speak about the errors propagation within the groups.

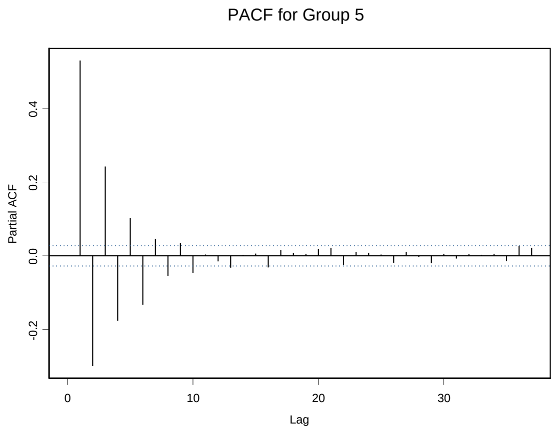

Figure 3.2: PACF for Group 5

Simple Correlation- and Partial Correlation Coefficients will be used here to identify the association between the variables. The formulas are given below respectively:

$$CCF(k) = Cor[Y_t, X_{t-k}] = \frac{E[(Y_t - \mu_Y)(X_{t-k} - \mu_X)]}{\sigma_Y \sigma_X} \tag{3.3}$$

$$PCF(k) = Cor[Y_t, X_{t-k}|X_t, X_{t-1}, ..., X_{t-k+1}] = Cor(R_{Y_t}, R_{X_{t-k}}) \tag{3.4}$$

Here μ_X and μ_Y are the means of X_t and Y_t respectively, while σ_X and σ_Y - standard deviations of the relevant time series. Analogically to PACF, PCF can be presented as the correlation between residuals R_{Y_t} and $R_{X_{t-k}}$ which are obtained after regressing Y_t and X_{t-k} on $X_t, X_{t-1}, ..., X_{t-k+1}$.

After examining the Auto- and Cross-Correlation for all the groups, the results are presented only for the Group 5, which seems to be the most promising one (Table 3.1). Note that this is not the correlation matrix but rather the table containing the correlation coefficients between Group 5 at time t and and the remaining Group values in the past up to the time t. Note that the last raw contains Auto-correlation values.

Group	t	t-1	t-2	t-3	t-4	t-5
1	0.1755	0.2818	0.3072	0.2166	0.1191	0.0696
2	0.1915	0.1866	0.1689	0.1631	0.1382	0.0795
3	0.1578	0.1481	0.1140	0.0814	0.0743	0.0597
4	0.2893	0.3198	0.2601	0.1391	0.0503	0.0183
5	1.0000	0.5267	0.0589	0.0403	0.0194	-0.0162

Table 3.1: Cross- and Auto-Correlation for Farm Group 5

Similar table was constructed for the Partial Correlation results (Table 3.2).

Group	t-1	t-2	t-3	t-4	t-5
1	0.2791	0.1147	-0.0105	-0.0605	-0.0594
2	0.1885	0.0439	0.0137	-0.0062	-0.0324
3	0.1424	0.0103	-0.0163	-0.0119	-0.0138
4	0.3219	0.0326	-0.0713	-0.0802	-0.0581
5	0.5294	-0.1935	-0.0855	-0.0584	-0.0539

Table 3.2: Partial Cross- and Auto-Correlation for Farm Group 5

From the Tables 3.1-3.2 it can be inferred that the largest influence on Group 5 are coming from Groups 1 and 4. The corresponding plots of those dependencies are presented in Figures 3.3-3.4.

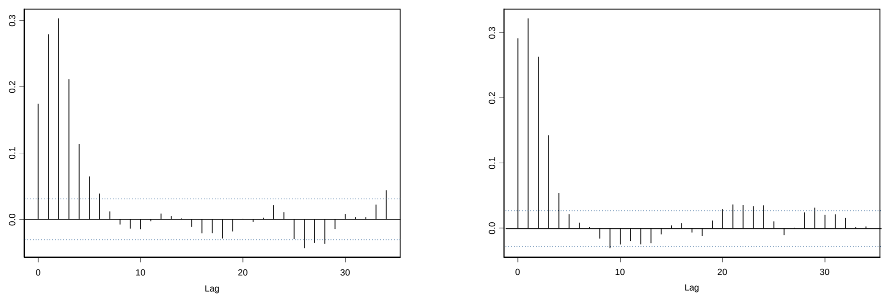

Figure 3.3: CCF for the Groups 5 and 1 (left) and Groups 5 and 4 (right)

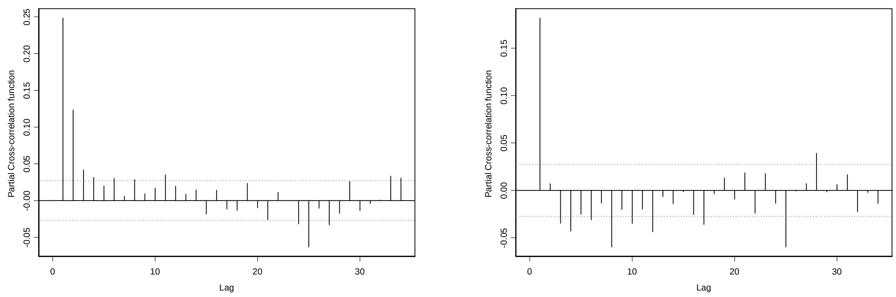

Figure 3.4: PCCF for the Groups 5 and 1 (left) and Groups 5 and 4 (right)

3.3 Dependency on the wind direction

The next step is to check how the wind direction influences the dependency. For further analysis Group 5 was chosen as a variable to be explained and other groups were decided to play the role of explanatory variables. However, similar results could be obtained trying to explain the errors of other groups.

In order to examine whether the wind direction has any influence on the dependency between the groups, we divide the data into four intervals according to the wind direction at Group 5 at time t. The division is performed by constructing four intervals: (0-90], (90-180], (180-270], (270-360] which are adequate to the cardinal directions (North-East, South-East, South-West, North-West, respectively). Cross-Correlation between the Group 5 and the remaining groups were calculated for each interval. Below, only the most significant results for Groups 5 and 1 as well as Groups 5 and 4 are presented (Table 3.3 and Table 3.4).

	Regime			
Lag	(0-90]	(90-180]	(180-270]	(270-360]
0	0.0457	0.1472	0.2240	0.1580
1	0.0499	0.2856	0.3597	0.2361
2	0.0672	0.3103	0.4213	0.2219
3	0.0358	0.1810	0.3218	0.1542
4	-0.0166	0.0985	0.2193	0.0519
5	0.0115	0.1130	0.1347	-0.0099

Table 3.3: Directional correlation for Groups 5 and 1

	Regime			
Lag	(0-90]	(90-180]	(180-270]	(270-360]
0	0.1390	0.3200	0.2615	0.3460
1	0.2212	0.2691	0.2570	0.4514
2	0.1788	0.2049	0.2075	0.3762
3	0.1288	0.1555	0.0978	0.1831
4	0.1014	0.0965	0.0158	0.0485
5	0.0252	0.0735	0.0102	-0.0157

Table 3.4: Directional correlation for Groups 5 and 4

It is easy to observe that in direction $(180-270]$ and $(270-360]$ the dependency is stronger than in the remaining intervals and stronger than the overall cross correlation coefficient for relevant groups. This result will be essential in the further parts of this paper.

3.4 Dependency on the wind speed

Another potential explanatory variable to be examined is the forecasted wind speed. We will check whether it influences the current correlation coefficients in a similar fashion as in Section 3.3. Namely, the data is grouped into five intervals according to the wind speed [m/s] at Group 5 in the time t as shown: 1 interval- [0,4), 2 interval- [4,6), 3 interval- [6,8), 4 interval- [8,10), 5 interval- [10,25)

Again, the cross-correlation coefficients were calculated for each interval separately. Figure 3.5 shows how the correlation coefficients vary among the lags for different wind speed levels. The data taken into account comes form direction [180-270) only. From the picture one can

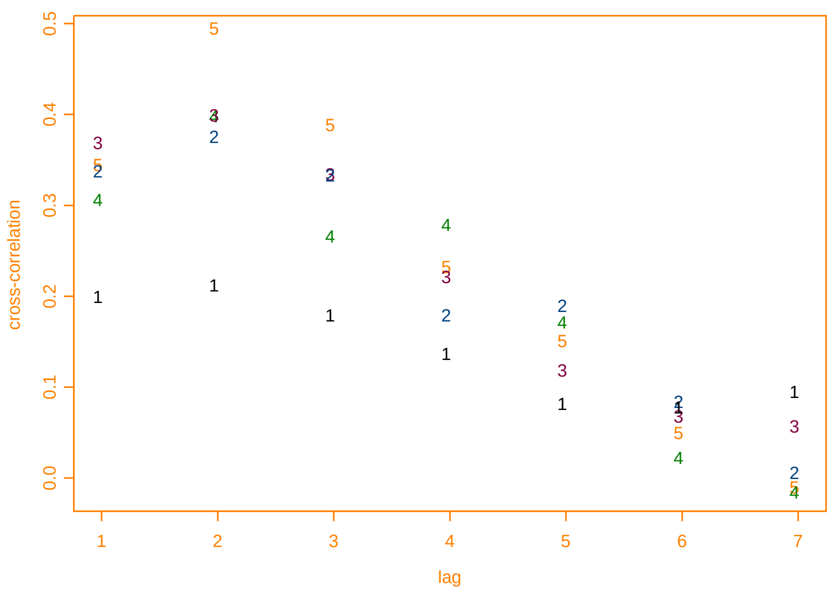

Figure 3.5: Cross-correlation for Groups 5 and 1 in direction (180,270] for different speed levels

conclude that there is a certain tendency: the higher the wind speed is, the larger dependency can be observed. However this analysis has some restrictions. One should keep in mind that the number of observations is not the same among the intervals. Each of them contains 200-400 data points which makes the results difficult to compare. The further analysis of the wind speed influence will be held in Chapter 4.

3.5 Dependency on the time of weather forecast updates

Another information, which might explain the cross-dependency between the groups, is the time when the last weather forecast was made. Since the forecasts are updated every 6 hours, it is possible that the correlation between the errors is dependent on the fact whether those errors were obtained while making predictions based on the same forecast or not. For this purpose we introduce a new variable $I(t, t-j)$, which indicates if the newest available forecasts for the time moments t and $(t-j)$ were made at the same time (then $I(t, t-j) = 0$) or not (then $I(t, t-j) = 1$). Figure 3.6 shows how the cross-correlation coefficients in lags between Groups 5 and 1 change for different value of I. The data taken into account comes form direction [180-270) only.

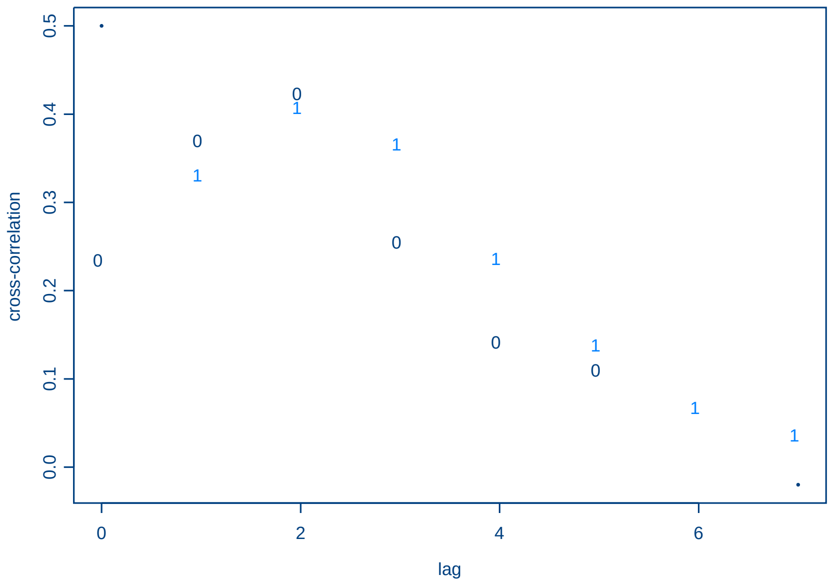

Figure 3.6: Cross-correlation for Groups 5 and 1 in direction (180,270] when $I(t, t-j) = 0$ ('0') and $I(t, t-j) = 1$ ('1')

From the Figure 3.6 one can see that the influence of I is not unambiguous. However, since even a small improvement of error predictions, achieved in this study, might be of a great importance, a slight growth of correlation coefficient for lags 3-4 noticed in Figure 3.6 will not be disregarded in the model building procedures.

3.6 Results

From the analysis carried out above one can clearly infer the dependency of the investigated process on four factors. Firstly, it is the influence of its own previous values, secondly, the dependency on the errors of remaining groups, and, finally, on wind direction and speed. Though the Autocorrelation seems to play the major role (value $= 0,5267$), after dividing the process according to the wind direction, we can observe fairly big influence of Group 1 (0,4213 in direction

$(180-270]$ at $lag = 2$) and Group 4 (0,4514 in direction $(270-360]$ at $lag = 1$). Time of weather forecast updates showed a slight influence on correlation between the groups at certain lags. This dependency does not look as promising as the four factor mentioned above, however the final inference will be made while building the models.

Chapter 4

Models

After identifying the structure of the data, the modeling part is to be performed. This takes place in the following chapter which is the core of this paper. All the considered models aim at making one hour error predictions. However, it is assumed that analogous methodology could be applied for longer term predictions. Since the highest Cross-Correlation coefficient was discovered between errors in Groups 5 and 4 (see Section 3.6), it was decided to use the errors of Group 5 as a dependent variable and assign it as Y_t. Errors of Groups 1-4 will be denoted as $X_{1,t}, ..., X_{4,t}$ and called the explanatory variables. This notation is throughout this paper.

This chapter is divided into four sections, each corresponding to the particular model type. As a general rule, every section begins with two theoretical sub-sections: Modelling and Estimation, followed by Application and Results in which models are used in practice, and finally results presented and compared.

As a starting point, the Linear Regression is considered. The Threshold Models governed by the external signal extend the topic by letting the coefficients vary among some selected regimes. In the third section the Varying Coefficients models will be introduced, including Kernel estimation, followed by the Conditional-parametric Models (Section 5.4).

4.1 Linear Regression

In this section several Linear Regression Models are considered. Among them, the special attention is payed to the AR and ARX Models. In the first case, the output depends on the previous output only, while the latter model allows the dependency on the external signal.

The section starts with the theoretical background regarding the model structure and a brief description of the famous Least Squares Estimation. Afterwards a few models are chosen and applied. Finally the results are introduced and with respect to them, the best model is selected.

4.1.1 Modeling

Firstly, we consider a very general case of Linear Regression having a form:

$$Y_t = \boldsymbol{X}_t^T \boldsymbol{\beta} + \epsilon_t. \tag{4.1}$$

Here, $\boldsymbol{\beta}$ is the set of coefficients, $\{\epsilon_t\}$ is a noise sequence and the column vector $\boldsymbol{X}_t$ contains all the possible explanatory variables including external signals at the different time points and the previous output. If only the latter component is used, the model reduces to autoregressive process of order p (called also AR(p)-process) ([11]). By the definition, it is given by:

$$Y_t = \beta_0 + \sum_{l=1}^{p} \beta_l Y_{t-l} + \epsilon_t, \tag{4.2}$$

with p being the order of the process. The name autoregressive indicates that (4.2) defines a regression of Y_t on its own past values.

ARX Model, which incorporates the stimulus signal is another important model from the linear family to be considered. The actual equation for the model is as follows:

$$Y_t = \beta_0 + \sum_{l=1}^{p} \beta_l Y_{t-l} + \sum_{i=1}^{n} \sum_{j=1}^{k_i} \beta_{i,j} X_{i,t-j} + \epsilon_t, \tag{4.3}$$

where the dependent variable Y_t is explained by its p previous values, and in addition by n external input variables, each up to lag k_i. All the coefficients may be put into a vector $\boldsymbol{\beta} = [\beta_0 ... \beta_p, \beta_{1,1} ... \beta_{n,k}]$, and the $\{\epsilon_t\}$ again stands for the noise sequence.

Below both of the above models will be applied. Before that the Ordinary Least Squares (OLS) estimation method will be briefly described.

4.1.2 Estimation

The main idea behind the Least Squares method is to minimize the residual sum of squares (RSS), which means to find an estimate $\hat{\boldsymbol{\beta}}$ of a real value $\boldsymbol{\beta}$, for which expression

$$RSS \equiv \sum_{t=1}^{T} \left(Y_t - \boldsymbol{X}_t^T \boldsymbol{\beta}\right)^2 \tag{4.4}$$

will be the smallest ([7]). Here, T is the length of the analyzed time series. Let us now rewrite (4.1) in a matrix form as:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{4.5}$$

where

$$\boldsymbol{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_T \end{pmatrix} \boldsymbol{X} = \begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_T^T \end{pmatrix} \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_T \end{pmatrix}$$

With this notation, the OLS estimate can be expressed as follows:

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \boldsymbol{Y}. \tag{4.6}$$

We disregard here all the computational aspects of arriving at the formula above. For a deeper description and properties of the LS estimates consult [7].

4.1.3 Application

Before fitting the model to data, one needs to decide which explanatory variables to choose. We begin with the univariate Linear Regression. The natural choice is to select the input variable which has the highest correlation coefficient with the dependent variable. Since our model aims at obtaining one hour predictions, all the explanatory variables must be considered at least in one hour delay with respect to Y_t. Following the results from Section 3.2 it was decided to use the errors of Group 1 (X_1) as an input. Model Lin 1 is presented below:

$$Y_t = \beta_0 + \beta_1 X_{1,t-1} + \beta_2 X_{1,t-2} + ... + \epsilon_t. \tag{4.7}$$

In order to decide on the lag number we check the partial cross-correlation results from Section 3.2. The "visual analysis" suggests two-lag dependency. However, here another criterion is applied: starting from the bigger model (10 lags), the amount of lags is gradually decreased and the results compared. As a simple rule we decided to disregard a variable if the decrease of R-squared value is smaller than 0.005. Furthermore, we eliminate the variables in case of big

P-value which means that according to the t-statistics the value of the corresponding coefficient does not significantly differ from zero. Following this method, three-lag structure was decided:

$$Y_t = \beta_0 + \beta_1 X_{1,t-1} + \beta_2 X_{1,t-2} + \beta_3 X_{1,t-3} + \epsilon_t. \tag{4.8}$$

The next step is to check what happens when one wants to broaden the range of explanatory variables. The errors of the remaining groups might also be used in the model. Again, we start with checking the correlation coefficients (Table 3.1). It can be seen that there is no significant dependency between groups 5 and 3. The correlation coefficient is only 0.1481 at time t-1 and decreases while the lag grows. The dependency on the Group 2 seems more promising, however it lays very close to Group 1. Incorporating it into the model does not improve the fit, once Group 1 errors are already included. As a result we disregard potential variables X_2 and X_3 in the model building process. It is not the case while considering X_4. The correlation in lag 1 is 0.3198 which is considered significant and furthermore, the layout of the group excludes the possibility of repeating information revealed by other group errors. After applying the procedure described above regarding the variable selection, it has been chosen to use (apart from the previously decided 3 lags of X_1) the variables $X_{4,t-1}$ and $X_{4,t-2}$ which leads us to Model Lin 2:

$$Y_t = \beta_0 + \beta_1 X_{1,t-1} + \beta_2 X_{1,t-2} + \beta_3 X_{1,t-3} + \beta_4 X_{4,t-1} + \beta_5 X_{4,t-2} + \epsilon_t. \tag{4.9}$$

Model Lin 3 consists, on the contrary, only of autoregressive explanatory variables. As mentioned in the previous section, these kind of models with k lag dependency are well known as the AR(k) models, which are regressed on their own k past values. The model can be formulated:

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \beta_3 Y_{t-3} + ... + \epsilon_t. \tag{4.10}$$

The ACF and PACF plots described in Section 3.1 suggest to use order k = 10. Akaike Information Criterion (AIC) results in 13-lag structure, however this criterion often causes model overparametrisation. Also in this case the "elimination method" was applied, leading to usage of only 7 lags. In the forthcoming section the results of all 3 cases are presented and compared.

The last of the Linear Models (Lin 4) includes all the explanatory variables described above, which are the lagged values of errors in Groups 1 and 4, as well as autoregressive part of Group 5. The model structure was chosen after checking many models using criteria like R-squared, t-statistics and P-value while eliminating insignificant variables. Finally we arrived at the following model:

$$Y_t = \beta_0 + \sum_{i=1}^{7} \beta_i Y_{t-i} + \sum_{j=1}^{3} \beta_{1,j} X_{1,t-j} + \sum_{l=1}^{2} \beta_{4,l} X_{4,t-l} + \epsilon_t. \tag{4.11}$$

4.1.4 Results

The models are compared using two criteria: R-squared (R^2) and Root Mean Squared Error (RMSE). Firstly, we describe them shortly. R-squared is the relative predictive power of the model. By definition it is a fraction of the total squared error, which is explained by the model

$$R^2 = 1 - \frac{SSerror}{SStotal} = \frac{SStotal - SSerror}{SStotal} \tag{4.12}$$

where SS stands for Sum of Squares. We can easily conclude that once the goodness of fit improves, SSerror decreases, and as a result R^2 approaches value 1. However, one has to be aware of the problems that can occur while using this measurement: once the number of parameters grows, the sample R-squared converges to 1 at the upper boundary. It is essential then to keep a reasonable amount of variables.

Secondly, the formula of RMSE is given below

$$RMSE = \sqrt{\frac{1}{n}\sum_{1=1}^{n}(y_i - \hat{y}_i)^2} = \sqrt{\frac{1}{n}\sum_{1=1}^{n}e_i^2} \tag{4.13}$$

which is just what it reads: the square root of the averaged errors squared (e_i). Table 4.1 gives the results expressed by those two measurements. It is clearly seen that the largest contribution

Model Name	No of Lags (X_1, X_4, Y)	R^2	RMSE
Lin 1	3, 0, 0	0.1189	0.0761
Lin 2	3, 2, 0	0.1744	0.0736
Lin 3	0, 0, 7	0.4213	0.0618
Lin 3	0, 0, 10	0.4271	0.0613
Lin 3	0, 0, 13	0.4272	0.0613
Lin 4	3, 2, 7	0.4794	0.0585

Table 4.1: Linear Models Results

toward explaining the variability of errors in Group 5 has the Autoregressive part of the model. However, 13 lags (indicated by the Akaike Criterion) seem not to be the best choice. After decreasing the number of lags to 10 the R^2 and RMSE remain almost the same. Finally, we decide to use order 7 of the autoregressive part since by further increasing it, does only imply the increase in R^2 value 0.5%. By adding the cross-components to the model we obtain a further improvement of 5%. In the end, according to R^2 value, it is possible to explain almost 48% of variation of Y_t with Model Lin 4.

4.2 Threshold Models

In this section we introduce the idea of what we call directional correlation and directional regression. The main purpose is to discover and capture the dependency between wind direction and propagation of prediction errors among the groups. The idea is supported by the intuitive, physical knowledge. Namely, we claim that if the wind direction is compatible with the direction of the vector, having its beginning in the group A and ending in B, the error dependency should be higher than in other cases. As before we will start with a theoretical section and proceed with the application and results.

4.2.1 Modeling

Threshold Models extend the idea of the Linear Models by letting coefficients vary among the regimes. Regimes are defined by the threshold values, which are the upper bounds of the intervals in which the given 'sub-model' is active. In the literature (see e.g [2], [12], [20]), one can find numerous classes of Threshold Models. Here the distinction between the models is made according to how they are governed: by the previous output, by some observable external signal or by an unobservable random variable.

Define intervals $R_1 \cup ... \cup R_k = \Re$ and $R_i \cap R_j = \emptyset, i \neq j$. Each interval is given by $R_i = (r_{i-1}, r_i]$. The values $r_0, ..., r_k$ are called thresholds. The general form of the models is examined below:

$$Y_t = \beta_0^{(J_t)} + \sum_{i \in G^{(J_t)}} \sum_{j \in L_{x_i}^{(J_t)}} \beta_{i,j}^{(J_t)} X_{i,t-j} + \sum_{l \in L_y^{(J_t)}} \beta_l^{(J_t)} Y_{t-l} + \epsilon_t, \tag{4.14}$$

where

$$J_t = \begin{cases} 1 & U_t \in R_1, \\ 2 & U_t \in R_2, \\ \vdots & \\ k & U_t \in R_k, \end{cases}$$

where U_t is an external signal which determines the regime switch, t is the time index, Y_t is the output variable, $\boldsymbol{X}_t$ are input variables, $\{\epsilon_t\}$ is zero mean white noise, L_y and L_{x_i} are sets of non-negative integers defining the autoregressive and input lags in the model, G is a set of positive integers defining which input variables to include into the model, J_t indicates the regime at time t. $\beta_{j,i}$ are coefficients to be estimated.

If U_t corresponds to variable Y_{t-k} which is the output variable in lag k,and the influence of the variable X_t is omitted, than the model is called Self-Exciting Threshold Auto Regressive (SETAR(k)) where the current regime is determined by the previous values of the output. However, the main emphasis in this section will be put on another type of the Threshold Models, governed by the external signal. It corresponds to the situation when the modeled process is influenced by another variable which is observable and known. It is very common approach in the weather related phenomena. For sake of complexity, one has to mention here the third type of models, when the change of regime depends on some variable we can't observe and sometimes we don't know. Very often we introduce this variable as a Markov Chain. A broad and complex description of Threshold Models can be found in [20].

4.2.2 Estimation

Estimation process focuses on two aspects: threshold values and the coefficients to be estimated. In this paper we consider the case when threshold values are known in advance. The motivation for that will be given in Application Section. Since the regimes are known, the estimation problem is solved by fitting different linear models to the data in each of the regimes. The technique used for that is again Least Squares, as described in the previous section (4.1.2).

4.2.3 Application

Let us assign the forecasted wind direction at the Group 5 at time t as U_t. The time series Y_t is divided according to U_t values into 4 vectors. The structure of the regimes is shown below:

$$J_t = \begin{cases} 1 \quad for \quad U_t \in [0, 90), \\ 2 \quad for \quad U_t \in [90, 180), \\ 3 \quad for \quad U_t \in [180, 270), \\ 4 \quad for \quad U_t \in [270, 360). \end{cases}$$

The decision of fixing the threshold values was dictated by easiness in interpreting the influence of wind direction which in this case is compatible with the geographical cardinal directions. Furthermore, other divisions were checked and the improvement in model fit was considered insignificant or none.

As in case of Linear Regression Models (Section 4.1.3), we will distinguish between 4 different models. In order to keep the unique names and a reasonable consistency, we substitute Lin by Thr. Furthermore following notation was adapted (here, relevant to Model Lin 1):

$$Thr\,1 = \begin{cases} Thr\,1.90 & for \quad J_t = 1, \\ Thr\,1.180 & for \quad J_t = 2, \\ Thr\,1.270 & for \quad J_t = 3, \\ Thr\,1.360 & for \quad J_t = 4 \end{cases}$$

which means, that each of the models is 'built' of linear models. As a result we obtain a four-regime model determined by the external signal (forecasted wind direction for groups at time t). Linear regression is fitted for each of the regimes. The choice of explanatory variables is made separately for each regime using the same rule as in Section 4.1.3. As mentioned before, we expect the dependency between errors to increase, when the wind direction is relevant to the direction of vector having its beginning in Group A, ending in Group B. Keeping it in mind, each regime should involve different combination of explanatory variables according to layout of the wind farms toward Group 5.

Staying in accord to the convention taken in Section 4.1.3, few models will be considered: Models Thr 1 capturing only one group dependency and Thr 2 which consists of all the cross-components. As a model relevant to Lin 3 from Section 4.1.3, Thr 3 will be recognized as a generalized SETAR (Self Exciting Threshold Auto Regressive Model). Finally, we arrive at the full version of the model in Thr 4 which will include both, the influence of the previous output and the external signal.

Thr 1 includes (as Lin 1) only the time lagged values of X_1 as shown below:

$$Y_t = \beta_0^{(J_t)} + \beta_1^{(J_t)} X_{1,t-1} + \beta_2^{(J_t)} X_{1,t-2} + ... + \epsilon_t. \tag{4.15}$$

Table 4.2 shows the final model structure.

Model Regime	Lag number in Group 1
Thr 1.90	1st,3rd,4th,5th
Thr 1.180	1st-3rd, 5th
Thr 1.270	1st-4th
Thr 1.360	1st and 3rd

Table 4.2: Model Thr 1 Structure

The second of the directional models will include all the possible cross-group-variables that were relevant to each regime. The model is given by:

$$Y_t = \beta_0^{(J_t)} + \sum_{i \in G^{(J_t)}} \sum_{j \in L_{x_i}^{(J_t)}} \beta_{i,j}^{(J_t)} X_{i,t-j} + \epsilon_t, \tag{4.16}$$

where G and L_{x_i} are the finite sets of indexes representing the number of the group and the corresponding lags, respectively. Again, we will decide on them separately for each regime, what has been shown in the Table 4.3.

Model Regime	Lags in Group 1	Group 2	Group 3	Group 4
Thr 2.90	-	-	-	1st,4th,5th
Thr 2.180	1st, 2nd, 5th	-	-	1st
Thr 2.270	1st-4th	-	-	1st and 4th
Thr 2.360	1st and 3rd	-	-	1st, 2nd, 5th

Table 4.3: Model Thr 2 Structure

As one can see, this structure is different from the one of Model Lin 2 which included three lags of Group 1 and two of Goup 4 only. Model Thr 3, which consists only of the previous output, is very similar to Open Loop Threshold AR (TARSO) described in [12]. The model is presented below:

$$Y_t = \beta_0^{(J_t)} + \sum_{i=1}^{k^{(J_t)}} \beta_1^{(J_t)} Y_{t-i} + \epsilon_t \tag{4.17}$$

While fitting the model, the following structure has been chosen:

Model Regime	AR lags
Thr 3.90	6
Thr 3.180	9
Thr 3.270	10
Thr 3.360	6

Table 4.4: Model Thr 3 Structure

The final and the most efficient form of the directional model is the one that captures both AR- and Cross-dependencies. The model is given by equation (4.14), and Table 4.5 reveals the structure of it.

Model Regime	Lags in Group 1	Group 2	Group 3	Group 4	AR
Thr 4.90	-	-	4th	1st	10
Thr 4.180	1st	-	-	1st	5
Thr 4.270	1st, 2nd, 4th	-	-	1st	6
Thr 4.360	1st and 3rd	-	-	1st	6

Table 4.5: Model Thr 4 Structure

The choice of the variables seems to be reasonable if the position of the farm groups is taken into account (see Figure 2.1), e.g. for the direction (270,360] which corresponds to the North-West wind, influence of the Groups 1 and 4 is significant. Maximum lags taken for Groups 1 and 4 conform with the directional distance from Group 5 in this regime. By the directional distance in this case we consider a distance between the groups projected on the representative direction of the corresponding regime. The respresentative direction of the regimes are:

$$\begin{cases} 45 & \text{for Regime 1,} \\ 135 & \text{for Regime 2,} \\ 225 & \text{for Regime 3,} \\ 315 & \text{for Regime 4.} \end{cases}$$

4.2.4 Results

In this section, the results of the Threshold Models will be introduced and compared with the simple Linear Regression (Table 4.6).

Model No	R^2	linear R^2	RMSE	linear RMSE
Thr 4.90	0.1412	0.1189	0.0752	0.076
Thr 4.180	0.2154	0.1744	0.0718	0.072
Thr 4.270	0.4363	0.4213	0.0608	0.062
Thr 4.360	0.4991	0.4794	0.0574	0.059

Table 4.6: Threshold Models Results

Furthermore, R^2 and RMSE values for all the regimes of Thr 4 are shown in Table 4.7. It appears that dividing the data according to directions is fairly successful. For the overall model we obtain more than 2% improvement in R^2 value, compared with the Linear Regression (Table (4.6)), while taking only data of directions $[180, 270]$ and $[270, 360]$, the R^2 reaches and exceedes level of 0.55. Root Mean Square Error Criterion also shows the advantage of the Treshold Models. It is

Model Regime	R^2	RMSE
Thr 4.90	0.384	0.05763
Thr 4.180	0.4614	0.04352
Thr 4.270	0.4932	0.06832
Thr 4.360	0.549	0.05579

Table 4.7: Model Thr 4 Results

due to the fact, that quite a lot of wind data is available in those directions and the position of Group 1 and Group 4 is relevant. On the contrary, considerably less observations are gathered in directions from intervals $[0, 90]$ and $[90, 180]$ and, what is more important, there is no groups of wind farms situated in those directions towards Group 5. That is the motivation for the further work to consist of the data from directions $[180, 270]$ and $[270, 360]$ only.

4.3 Varying-coefficient models

Here we try to improve the directional model presented in the previous section by including additional weather forecast information into the model building procedure. For this purpose a further extension of linear ARX-models - varying-coefficient ARX-models are to be considered.

4.3.1 Modeling

The varying-coefficient ARX-models are the models which are linear in regressors, but their coefficients are allowed to change smoothly with the value of other variables (called 'effect-modifiers'). The model has the form ([8], [11]):

$$Y_t = \sum_{i \in L_y} Y_{t-i}\beta_{0,i}(U_{0,t-i}) + \sum_{j \in G}\sum_{i \in L_x^j} X_{j,t-i}\beta_{j,i}(U_{j,t-i}) + \epsilon_t, \tag{4.18}$$

where t is the time index, Y_t is the output variable, $\boldsymbol{X}_t$ and $\boldsymbol{U}_t$ are input variables, $\{\epsilon_t\}$ is zero mean white noise, L_y and L_x^j are sets of non-negative integers defining the autoregressive and input lags in the model, G is a set indicating groups to be considered, $\beta_{j,i}(\cdot)$ are unknown but smooth functions to be estimated. Extension to multivariate $U_{j,t}$ is straightforward.

4.3.2 Estimation

The aim is estimation of the functions $\boldsymbol{\beta} = [\boldsymbol{\beta}_{0,i\in L_y}, \dots, \boldsymbol{\beta}_{m,i\in L_x^m}]$ within the space spanned by the observations of input variables. Please, note that $\boldsymbol{\beta}_{0,i\in L_y}$ is a vector of coefficient functions for corresponding lags of the autoregressive part. Analogically, $\boldsymbol{\beta}_{m,i\in L_x^m}$ is a vector of coefficient functions for the corresponding lags of the m^{th} explanatory variable. Model (4.19) as it stands is too general for most applications, where no restrictions are imposed on coefficient functions. The estimation approach considered in this paper is based on grouping the input data. The main idea behind this solution is grouping the data into n intervals according to the values of $\boldsymbol{U}_t = [U_{0,t}, U_{1,t}, \dots, U_{m,t}]$. First it is assumed that inside the intervals coefficient functions $\boldsymbol{\beta}$ are constant. Then, instead of dealing with estimation of continuous functions, the problem reduces to the estimation of coefficient matrix:

$$\boldsymbol{\beta}^* = \begin{bmatrix} \boldsymbol{\beta}_{0,1}^{*i\in L_y} & \boldsymbol{\beta}_{1,1}^{*i\in L_x^1} & \dots & \boldsymbol{\beta}_{m,1}^{*i\in L_x^m} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{\beta}_{0,n}^{*i\in L_y} & \boldsymbol{\beta}_{1,n}^{*i\in L_x^1} & \dots & \boldsymbol{\beta}_{m,n}^{*i\in L_x^m} \end{bmatrix}$$

$\boldsymbol{\beta}^{*i}_{(\cdot),k}$ corresponds to the value of parameters in the k^{th} interval for the lag i. In this case model (4.19) is modified to:

$$Y_t = \sum_{i \in L_y} Y_{t-i} \beta^{*i}_{0,k(U_{0,t-i})} + \sum_{j \in G} \sum_{i \in L_x^j} X_{j,t-i} \beta^{*i}_{j,k(U_{j,t-i})} + \epsilon_t, \tag{4.19}$$

Least Squares technique is used for finding estimates of $\boldsymbol{\beta}^*$. At this point the estimation procedure could stop, assuming that the parameter functions are constant within the constructed intervals. However, another possible step to take is estimation of the changing behavior of coefficients inside the constructed interval. For this purpose the following assumption is made: the value of coefficient estimated for the whole interval is assumed to be equal to the value of the coefficient function β at one of the points from the analyzed interval. Usually the middle point of the interval is chosen for that purpose. Which means the following assumption: $\beta_{j,i}(u) = \beta^{*i}_{j,k(U_{j,t-i})}$, where u is a middle point of the k^{th} interval. Then to find the values of $\boldsymbol{\beta}$ smoothing technique is used. Below, the main idea of smoothing together with the applied methods (kernel smoothing and splines) are described.

Smoothing

Given the data (x_i, y_i), where $y_i \approx f(x_i), i = 0, \ldots, n$ it is sometimes needed to find $f(x)$ values, when $x = t$, $t \in [x_0, x_n]$ Quite often $f(x)$ is unknown or very difficult to calculate [18]. Then depending on the precision of y_i estimates there are two ways to solve the problem:

- Interpolation.
- Smoothing.

Interpolation is used when $y_i = f(x_i)$ or errors are that small that it is possible to neglect them. Smoothing is applicable when y_i are just approximate estimates of $f(x_i)$.

Kernel smoothing   A kernel smoother uses an explicitly defined set of local weights, defined by the kernel, to produce the estimate of each target value. Usually a kernel smoother uses weights that decrease in a smooth fashion as one moves away from the target point.

The weight given to the jth point in producing the estimate at x_0 is defined by

$$S_0 j = \frac{c_0}{\lambda} d\left(\left|\frac{x_0 - x_j}{\lambda}\right|\right) \tag{4.20}$$

where $d(t)$ is an even function decreasing in $|t|$. Parameter λ is the window-width, also known as a bandwidth, and the constant c_0 is usually chosen so that the weights sum to unity, although there are slight variations on this. A natural candidate for d is the standard Gaussian density: this gives the so-called Gaussian kernel smoother.

As far as computational aspects are concerned one can visualize the action of kernel smooth as sliding the weight function along x-axis is short steps, each time computing the weighted mean of y. The smooth is thus similar to a convolution between the kernel and an empirical step function defined on the data. Typically the kernel smooth is computed as

$$s(x_0) = \frac{\sum_{i=1}^{n} d\left(\frac{x_0 - x_i}{\lambda}\right) y_i}{\sum_{i=1}^{n} d\left(\frac{x_0 - x_i}{\lambda}\right)} \tag{4.21}$$

Splines Splines are piecewise polynomial functions that are constrained to join at points called knots, which divide the range of x into regions. In addition, it is customary to force the piecewise polynomial to join smoothly at the knots. The theory of splines is dissertated in [9], [5] and computational aspects are carefully described in [18]. Below we present the main idea of spline theory together with computational steps which were followed while analyzing the data.

Smoothing splines Polynomial regression has limited appeal due to the global nature of its fit, while in contrast the kernel smoothers have an explicit local nature. Splines offer a compromise by representing the fit as a piecewise polynomial.

Consider the following problem: among all functions $f(x)$ with two continuous derivatives it is needed to find one which minimizes penalized residual sum of squares:

$$\sum_{i=1}^{n} \{y_i - f(x_i)\}^2 + \lambda \int_a^b \{f''(t)\}^2 dt \tag{4.22}$$

where λ is a fixed constant and $a \leq x_1 \leq ... \leq x_n \leq b$. The first term measures closeness to the data and the second term penalizes curvature of the function. Large values of λ produce smoother curves while smaller values produce more wiggly curves. At one extreme, as $\lambda \to \infty$, the penalty term dominates, forcing $f''(x) = 0$ everywhere and thus the solution is the least-squares line. At the other extreme, as $\lambda \to 0$,the penalty term becomes unimportant and the solution tends to an interpolating twice-differentiable function.

It is known that the problem stated above has an explicit, unique minimizer and that the minimizer is a natural cubic spline with knots at the unique values of x_i. [9]

A natural cubic spline is a variant of cubic spline with an additional constraint requiring that the function is linear beyond the boundary knots.

Computational aspects For computational convenience we will use the unconstrained B-spline basis and write $s(x) = \sum_1^{n+2} \theta_j B_j(x)$, where θ_j are coefficients and the B_j are the cubic

B-spline basis functions [9],[18]. To solve (4.22) we replace f by s and perform the integration. Defining matrices $\boldsymbol{B}$ and $\boldsymbol{\Omega}$ by

$$B_{ij} = B_j(x_i), \tag{4.23}$$

$$\Omega_{ij} = \int B_i''(x) B_j''(x) dx. \tag{4.24}$$

We can rewrite the residual sum of squares (4.22) as:

$$(y - B\theta)^T (y - B\theta) + \lambda \theta^T \Omega \theta. \tag{4.25}$$

Setting the derivative of (4.25) with respect to θ equal to 0 gives

$$\hat{\theta} = \left(B^T B + \lambda \Omega\right)^{-1} B^T y. \tag{4.26}$$

4.3.3 Application

We try to improve the Threshold Model fitted in the previous section by assuming that its coefficients are not constants but smooth functions of additional weather forecast information. Here the coefficients are treated as functions of a forecasted wind speed level and forecast-age characteristic (Model VarCoef 1 shown in (4.27)).

$$Y_t = \sum_{i \in L_y} Y_{t-i} \beta_{0,i}(s_{5,t-i}, I_{t,t-i}) + \sum_{j \in G} \sum_{i \in L_x^j} X_{j,t-i} \beta_{j,i}(s_{j,t-i}, I_{t,t-i}) + \epsilon_t, \tag{4.27}$$

where the structure of the model is similar to the one of Threshold Model presented before (see Table 4.5). Variable $s_{j,t}$ indicates wind speed level at time moment t near the group j. $I_{t,t-i}$ is an indicator function showing weather the newest available forecasts for time moments t and $(t-i)$ were made at the same time (then $I_{t,t-i} = 0$) or not (then $I_{t,t-i} = 1$).

For dividing wind speed into intervals a vector of knots $[0, 4, 6, 8, 10, 30]$ was chosen after consulting power curve data. For kernel smoothing a normal kernel with $bandwidth = 5$ was used.

4.3.4 Results

The results of varying coefficient model for the Regime 4 in comparison with the analogous Regime of the Threshold model are shown in the Table 4.8.

Smoothing type	R^2	RMSE
Kernel	0.5582	0.0552
Spline	0.5532	0.0555
No smoothing	0.5829	0.0536
Threshold model for direction [270,360)	0.5490	0.0558

Table 4.8: Varying coefficient model results for direction [270,360)

Summarizing, the varying coefficient model gave slightly better results than a threshold model of the similar structure. However, it is much more complex and time demanding to estimate than the threshold model. Another remark is the fact that the results are better without applying the smoothing step in estimation procedure.

The coefficient function plots can be found in Figures $4.1 - 4.6$. One can easily observe that the influence of the forecast age does not influence the coefficients as much as the wind speed does. Graphical representation of the coefficients dependency on the wind speed also shows some curvature which might be caused more by the fact of insufficient amount of data points used for tracking the behaviour inside the intervals, than by the real nature of the phenomena.

The arguments discussed above show that probably the improvement could be even bigger if different estimation approach was applied. Obviously, some useful information is lost by simple division into intervals and averaging. Also, some improvement could be made in the choice of the 'effect-modifiers'. So far influence of the wind speed was taken into account separately for each of the groups and lags. It is probable that a better choice would be to consider an adjusted wind sped level information, summarizing the overall wind speed situation near the points of interest. Those corrections are to be considered in the next section.

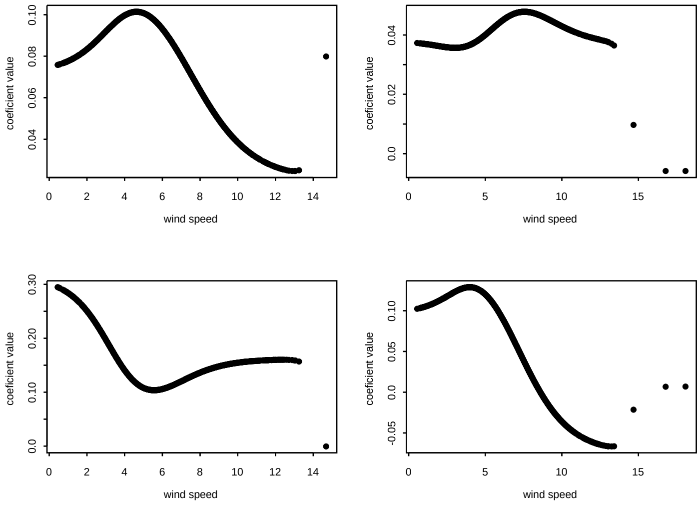

Figure 4.1: Coefficients $\beta_{1,1}$ (left column) and $\beta_{1,3}$ (right column) from the model VarCoef 1 dependency on the wind speed level. Top row - when $I(t, t-j) = 0$, bottom row - when $I(t, t-j) = 1$. Kernel smoothing is used.

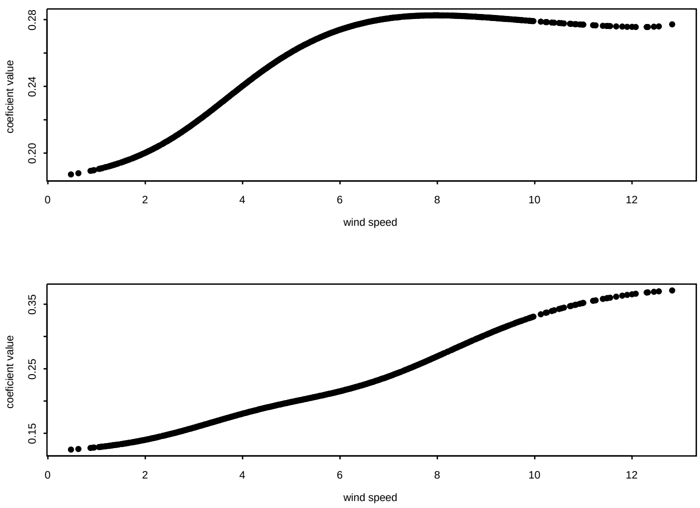

Figure 4.2: Coefficient $\beta_{4,1}$ from the model VarCoef 1 dependency on the wind speed level. Top row - when $I(t, t-j) = 0$, bottom row - when $I(t, t-j) = 1$. Kernel smoothing is used.

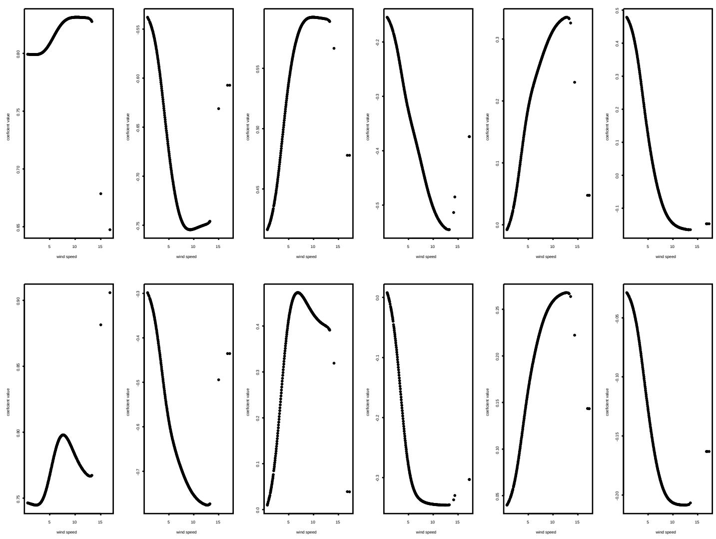

Figure 4.3: Coefficients $\beta_{0,\cdot}$ from the model VarCoef 1 dependency on the wind speed level. Top row - when $I(t, t-j) = 0$, bottom row - when $I(t, t-j) = 1$. Columns correspond to the lags taken. Kernel smoothing is used.

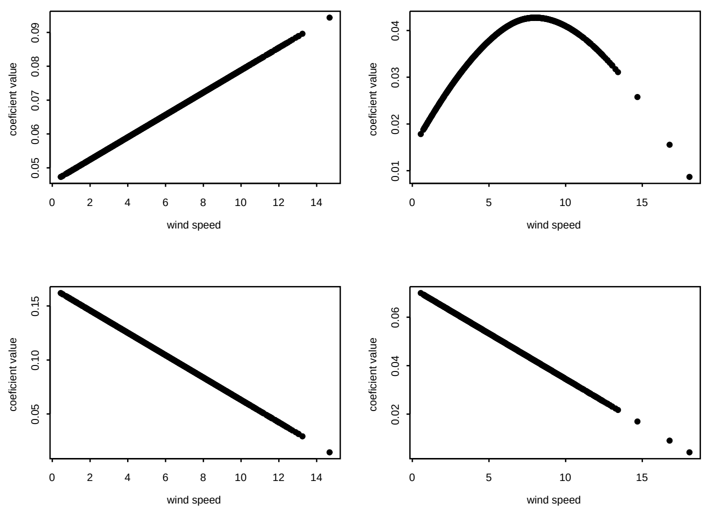

Figure 4.4: Coefficients $\beta_{1,1}$ (left column) and $\beta_{1,3}$ (right column) from the model VarCoef 1 dependency on the wind speed level. Top row - when $I(t, t-j) = 0$, bottom row - when $I(t, t-j) = 1$. Smoothing splines are used.

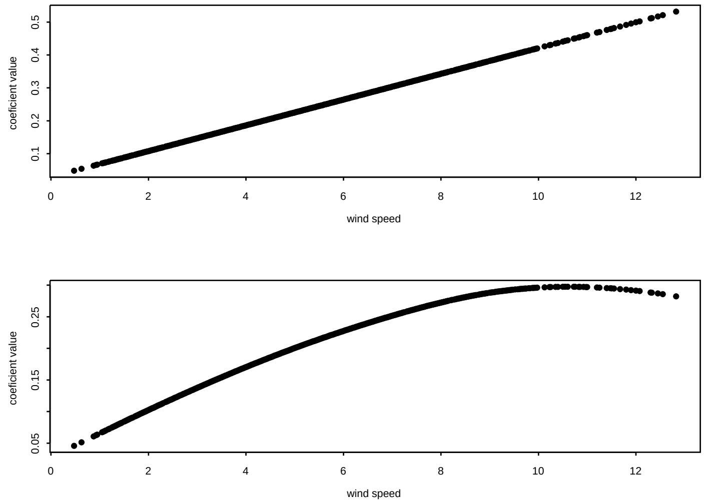

Figure 4.5: Coefficient $\beta_{4,1}$ from the model VarCoef 1 dependency on the wind speed level. Top row - when $I(t, t-j) = 0$, bottom row - when $I(t, t-j) = 1$. Smoothing splines are used.

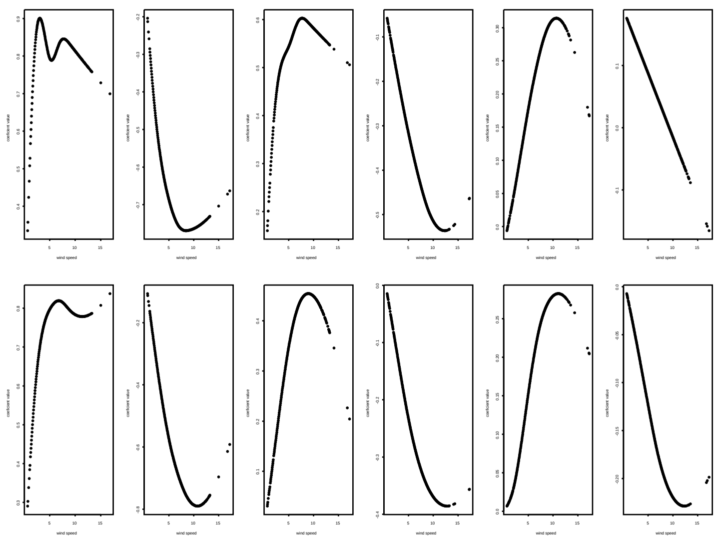

Figure 4.6: Coefficients $\beta_{0,.}$ from the model VarCoef 1 dependency on the wind speed level. Top row - when $I(t, t-j) = 0$, bottom row - when $I(t, t-j) = 1$. Smoothing splines are used. Columns correspond to the lags taken.

4.4 Conditional parametric models

Varying coefficient models, when all coefficients depend on the same variables, are denoted as conditional parametric models. The broader description can be found in [1] and [4]. The model has the form :

$$Y_i = \boldsymbol{z}_i^T \boldsymbol{\theta}(\boldsymbol{x}_i) + e_i; \quad i = 1, \dots, N, \tag{4.28}$$

where Y_i is a measure of the response, $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$ are the explanatory variables, e_i are independent normal variables with $E(e) = 0$ and $Var(e) = \sigma^2$. $\boldsymbol{\theta}(\cdot)$ is a vector of unknown smooth functions, which estimation we are interested in and $i = 1, ..., N$ indicates the observation number.

4.4.1 Estimation

The parameters of the above model are estimated locally. In this work two cases are presented: first, when we simply replace $\boldsymbol{\theta}(\cdot)$ by the locally constant vector, and second, approximating it by the local polynomial. Both cases are described below.

Before estimation   Before starting estimation, we have to select the weight function, bandwidth and polynomial degree. All of those factors can substantially influence the quality and properties of our estimate. Below we describe some methods which should help to choose them in the most optimal way. Let us denote $\boldsymbol{x}$ as a single point out of those, for which the function $\boldsymbol{\theta}(\cdot)$ would be estimated and the $\hat{\boldsymbol{\theta}}(\boldsymbol{x})$ as a proper estimate of this function. This notation will be valid for the whole following chapter.

Selecting the weight function   Let us assume that $\boldsymbol{\theta}(\cdot)$ is continuous. In most cases we would like the weight function $W(\boldsymbol{x})$ to give the largest values to the observations close to $\boldsymbol{x}$ and decaying smoothly while the distance increases. Also, considering the computation speed, we wish the weighting function to take nonzero values only in some bounded interval. In such cases we can just ignore the observations which do not belong to the interval giving them weights equal to 0. If $\boldsymbol{x}$ is multidimensional, we should scale the elements, e.g. divide by the standard deviation. The most common weight functions are Box, Triangle, Tricube and Gaussian. They are presented in Table 4.9. Its worth remembering that the smoother the weight function is, the smoother estimate we will obtain. Below we describe briefly the case of a spherical kernel:

$$w_i(\boldsymbol{x}) = W\left(\frac{||\boldsymbol{x}_i - \boldsymbol{x}||}{h(\boldsymbol{x})}\right). \tag{4.29}$$

Name	Weight function
Box	$W(u) = \begin{cases} 1, & u \in [0;1) \\ 0, & u \in [1;\infty) \end{cases}$
Triangle	$W(u) = \begin{cases} 1-u, & u \in [0;1) \\ 0, & u \in [1;\infty) \end{cases}$
Tricube	$W(u) = \begin{cases} (1-u^3)^3, & u \in [0;1) \\ 0, & u \in [1;\infty) \end{cases}$
Gauss	$W(u) = \exp(-u^2/2)$

Table 4.9: Weight Functions.

where $||\boldsymbol{x}_i - \boldsymbol{x}||$ is an Euclidean distance between $\boldsymbol{x}_i$ and $\boldsymbol{x}$, and $h(\boldsymbol{x})$ is a scalar called the bandwidth.

Selecting the bandwidth Choosing an appropriate bandwidth can strongly improve the smoothness of the estimate $\hat{\boldsymbol{\theta}}(\cdot)$. As usual in such cases our aim is to have both, the smallest possible variance and bias. The most common choices are fixed and the nearest neighbor bandwidth. Usually the fixed bandwidth performs worse, since the changes of density of the data can lead to high variability. One should also pay extra attention to the observations laying in the tails of the distribution. If we apply the same bandwidth in the boundary regions as the interior data, the variation will increase. But it is not only the bandwidth that decides about smoothness of our estimates. The selection of polynomial order plays a big role as well.

Choice of polynomial degree Another factor that we have to choose before proceeding the estimation is the polynomial degree. The rule is easy: the higher order of the polynomial the better fit we obtain which means the smaller bias. We loose, though, the smoothness and our estimate becomes more noisy which leads to increased variation. The choice of polynomial degree and the bandwidth should be considered together. Once we decide to fit a higher order polynomial, we can extend the bandwidth and take broader neighborhood. In this way we will maintain the smoothness. Interesting results can be obtained by applying locally constant fitting (0 order polynomial). However, locally quadratic fitting will perform better, when the approximated function is not monotonic.

Local constant estimates After choosing the weight function, the bandwidth and the polynomial degree, the estimation is performed. In case of 0 degree polynomial, we replace $\boldsymbol{\theta}(x_i)$ in (1) by the constant vector $\boldsymbol{\theta}(x)$. To estimate $\hat{\boldsymbol{\theta}}(\boldsymbol{x})$ we fit the model locally to $\boldsymbol{x}$ using weighted least squares. In particular, when $z_i = 1$ for all i the method reduces to kernel regression.

Local polynomial estimates As mentioned above, in situation of substantial curvature of regression surface, the local polynomial would deliver the better results than the local constant fitting. Let's, like in [15] denote $\boldsymbol{\theta}_j(\cdot)$ to be the j'th element of $\boldsymbol{\theta}(\cdot)$ and $\boldsymbol{P}_d(\boldsymbol{x})$ a vector contained terms of d-oder polynomial at $\boldsymbol{x}$, e.g. $\boldsymbol{P}_2(\boldsymbol{x}) = [1\ x_1\ x_2\ x_1^2\ x_1x_2\ x_2^2]^T$ when $\boldsymbol{x} = [x_1\ x_2]^T$. The estimation is performed by fitting the model

$$Y_i = \boldsymbol{u}_i^T \boldsymbol{\phi}(\boldsymbol{x}) + e_i; \quad i = 1, \ldots, N, \tag{4.30}$$

locally to x. Where

$$\boldsymbol{u}_i^T = \left[z_{1i}\boldsymbol{P}_{d(1)}^T(\boldsymbol{x}_i) \ldots z_{ji}\boldsymbol{P}_{d(j)}^T(\boldsymbol{x}_i) \ldots z_{pi}\boldsymbol{P}_{d(p)}^T(\boldsymbol{x}_i)\right] \tag{4.31}$$

and

$$\hat{\boldsymbol{\phi}}^T(\boldsymbol{x}) = [\hat{\boldsymbol{\phi}}_1^T(\boldsymbol{x}) \ldots \hat{\boldsymbol{\phi}}_j^T(\boldsymbol{x}) \ldots \hat{\boldsymbol{\phi}}_p^T(\boldsymbol{x})], \tag{4.32}$$

$\boldsymbol{z}_i = [z_{1i} ... z_{pi}]^T$ and $\hat{\boldsymbol{\phi}}_j(\boldsymbol{x})$ is a vector of local constant estimates evaluated at $\boldsymbol{x}$ corresponding to the relevant elements of $\boldsymbol{u}_i^T$. Finally, we can write our estimate as follows:

$$\hat{\theta}_j(\boldsymbol{x}) = \boldsymbol{P}_{d(j)}^T(\boldsymbol{x})\, \hat{\boldsymbol{\phi}}_j(\boldsymbol{x}); \quad j = 1, \ldots p. \tag{4.33}$$

At the beginning we assumed that the residuals e_i are normally distributed with mean 0 and variance σ^2. Note that while estimating $\hat{\boldsymbol{\theta}}_j(\boldsymbol{x})$ we use observations only in the neighborhood of $\boldsymbol{x}$ so we can allow heteroschadasity of residuals among the whole data set. Variance should be constant, though, within the each neighborhood of $\boldsymbol{x}_i$.

Software In the further part of this chapter we implement LFLM (Local Fitting of Linear Model) software by Henrik Aalborg Nielsen from Danish Technical University. More details about the software can be found in [15]. The brief description of the application method and the results are given below.

4.4.2 Application

Here, two cases are considered: firstly the coefficients are treated as functions of a forecasted wind speed level (Model Cond 1 shown in (4.34)), and, secondary, forecast age characteristic is added into model building procedure (Model Cond 2 shown in (4.35)).

$$Y_t = \sum_{k=1}^{4} \left(\sum_{j \in L_x^k} \theta_j^k(s_t) X_{k,t-j} \right) + \sum_{j \in L_y} \theta_j^5(s_t) Y_{t-j} + \epsilon_t, \tag{4.34}$$

$$Y_t = \sum_{k=1}^{4} \left(\sum_{j \in L_x^k} \theta_j^k(s_t, I_{t,t-j}) X_{k,t-j} \right) + \sum_{j \in L_y} \theta_j^5(s_t, I_{t,t-j}) Y_{t-j} + \epsilon_t, \tag{4.35}$$

where, analogously to the way it was assigned in the previous section, k indicates a number of group, L_y and L_x^k define the autoregressive and input lags in the model. The structure of the model is similar to the one of Thr 4 presented before (see Table 4.5).

Variable s_t indicates wind speed level at time moment t. Firstly, it was decided to take wind speed near Group 5 at time moment t as a representative of s_t. Of course, in this case the information about wind speed levels near other groups at different time moments was lost and not involved into the model. Having it in mind it was decided that s_t should be a summary of wind speed information near all the farms included into the model. For making such summary weighted linear composition was chosen. The weights were selected according to the correspondent coefficients of simple linear regression of Y_t on the rest of variables from the model.
$I_{t,t-j}$ is an indicator function showing weather the newest available forecasts for time moments t and $(t-j)$ were made at the same time (then $I_{t,t-j}$=0) or not (then $I_{t,t-j}$=1).

The polynomial degree while applying LFLM software was set to 1 since this value corresponds the best to the nature of coefficient dependency on explanatory variables.

4.4.3 Results

As mentioned before the main focus in this work will be on the data from the directions $[180, 270)$ and $[270, 360)$. The results from the other directions will be mentioned but not analyzed in details.

Firstly the Cond 1 was applied. Summary of the obtained results in comparison with Thr 4 can be found in Table 4.10

The improvement in R^2 and RMS values is significant, especially for directions $[0, 90)$ and $[180, 270)$. Those results look very promising.

Cond 1	R^2	RMS	Thr 4	R^2	RMS
Regime 1	0.485	0.053	Regime 1	0.384	0.075
Regime 2	0.483	0.043	Regime 2	0.461	0.044
Regime 3	0.556	0.064	Regime 3	0.493	0.068
Regime 4	0.577	0.054	Regime 4	0.549	0.056

Table 4.10: Cond 1 Model Results

Simple diagnostics of the Regimes 3-4 from the Cond 1 are shown in Figure 4.7. One can see that the model does not describe the data well in the tails. After checking the correlation of residuals (Figure 4.8) a bootstrapping technique was applied to check the level of uncertainty associated with the estimates of coefficients. The results are shown below in Figures 4.9-4.10. In can be observed that when the wind speed is very high or very low, the uncertainty level is much higher. This is related to insufficient amount of data available for those intervals.

As far as the coefficients dependency on the wind speed level is concerned, in can be noticed, that the influence of AR- part decreases when the wind speed level grows. The influence of the other groups, on the contrary, grows. This is logical, especially having in mind that this growth is most significant for the groups which are located following current wind direction from group 5 (i.e for Group 1 for the regime 3 and for the Group 4 in the regime 4).

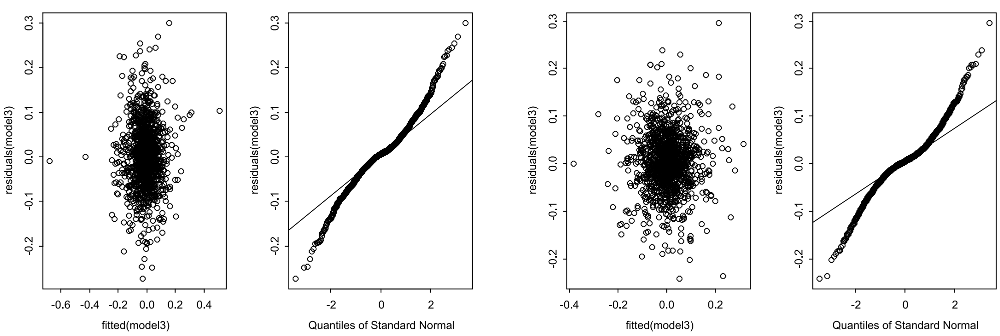

Figure 4.7: Simple diagnostics for the fit corresponding to model Cond 1 regimes 3 (left) and 4 (right)

Second step was to introduce forecast characteristic into the data analysis by applying Cond 2. Summary of the obtained results in comparison with Cond 1 can be found in Table 4.4.3

The improvement in R^2 and RMS values can be noticed especially for the direction $[0, 90)$. 4% improvement noticed for the direction $[180, 270)$ also looks promissing. For other directions the growth in R^2 is not that big. Looking back at the results obtained from Cond 1 one can notice that in both case (Cond 1 and Cond 2) bigger improvements were achieved for directions $[0, 90)$

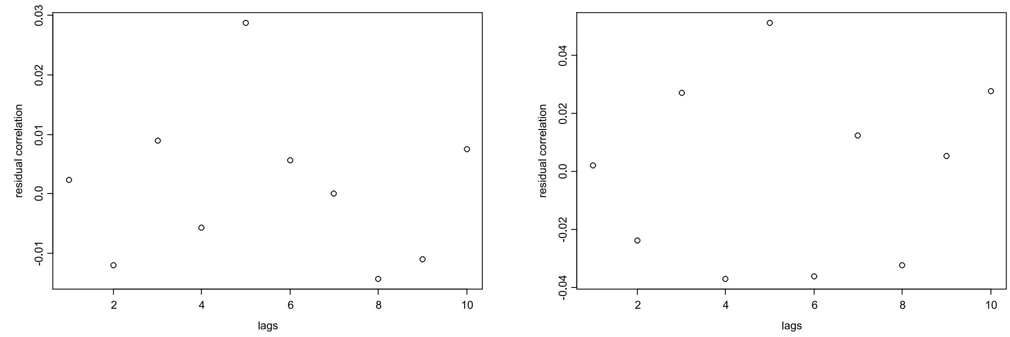

Figure 4.8: Correlation of model Cond 1 regimes 3 (left) and 4 (right) residuals in lags

Cond.2 model	R^2	RMS	Cond.1	R^2	RMS
Regime 1	0.556	0.049	Regime 1	0.485	0.053
Regime 2	0.493	0.042	Regime 1	0.483	0.043
Regime 3	0.590	0.062	Regime 1	0.556	0.064
Regime 4	0.593	0.053	Regime 1	0.577	0.054

Table 4.11: Cond.2 Model Results

and $[180, 270)$ than for $[90, 180)$ and $[270, 360)$. One of the possible reasons is that groups 1 and 4, which are included into both of the models, supply more information about the errors when the wind is in directions $[0, 90)$ and $[180, 270)$, since for the remaining regimes the directional distance between those groups is much smaller, then it is logical to assume that some of the information is repeated.

Simple diagnostics of the regimes 3-4 from the Cond 2 are shown in Figure 4.11. The pattern of the tail fit is similar to the one observed for Cond 1. The residuals test can be found in (Figure 4.12). The estimates of coefficients are presented in Figures 4.13-4.14. One can see that when the wind speed is very high or very low, the uncertainty level blows up. Similar to the case discussed for Cond 1 this is related to insufficient amount of data available for those intervals.

Talking about the coefficients dependency on the wind speed level the same pattern of decrease in the influence of AR- part can be noticed. If to analyze how coefficients depend on the forecast characteristic, then from the plots no significant difference can be found. Judging from the estimated coefficient plots and having in mind quite small growth in R^2 for the main directions of interest one can decide that including forecast characteristic into the model does not pay-off.

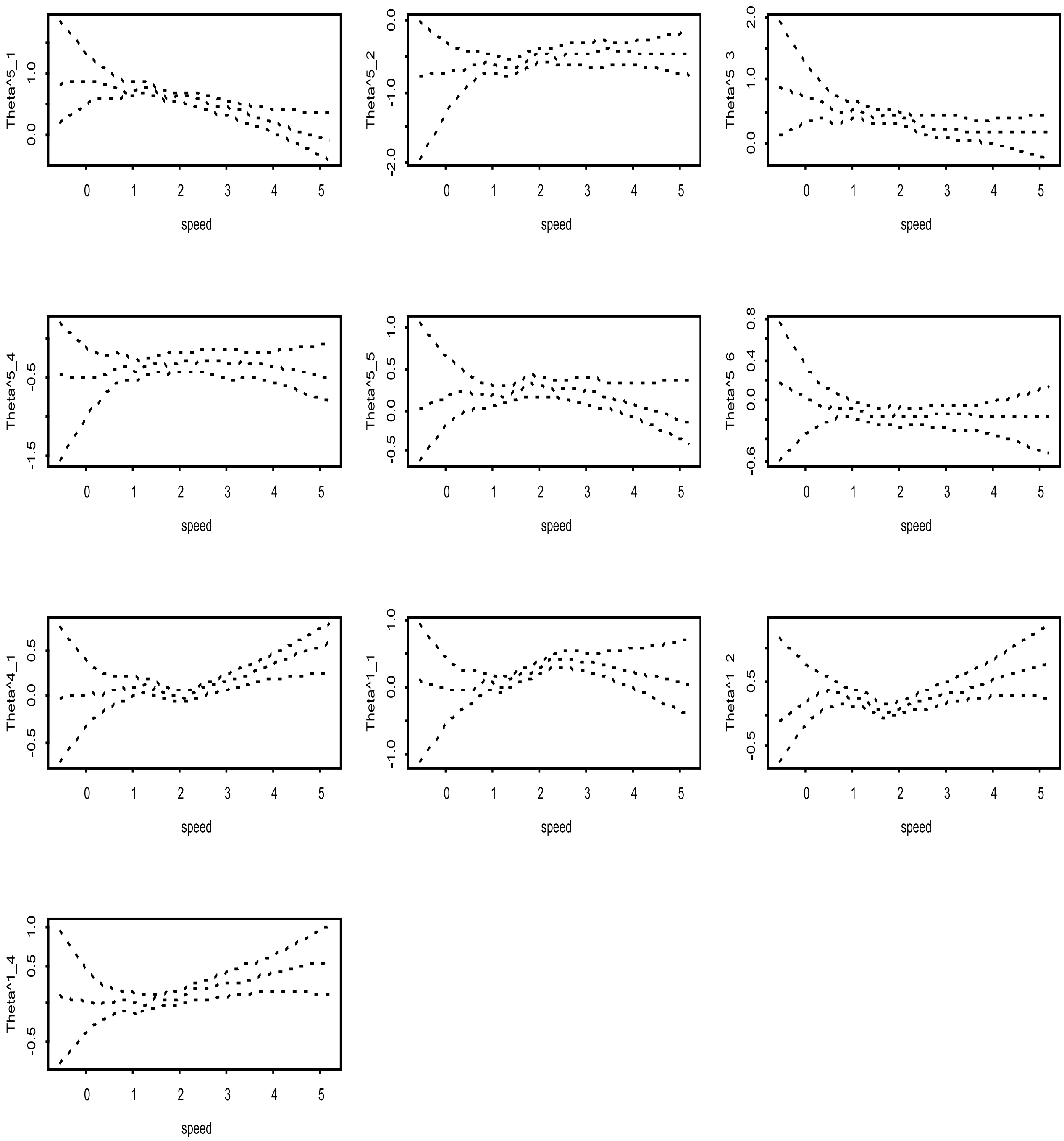

Figure 4.9: Obtained coefficients and 95% standard normal intervals based on 200 bootstrap replicates of regime 3 in model Cond 1. Here Thetaˆ k_j stands for θ_j^k and speed level for s

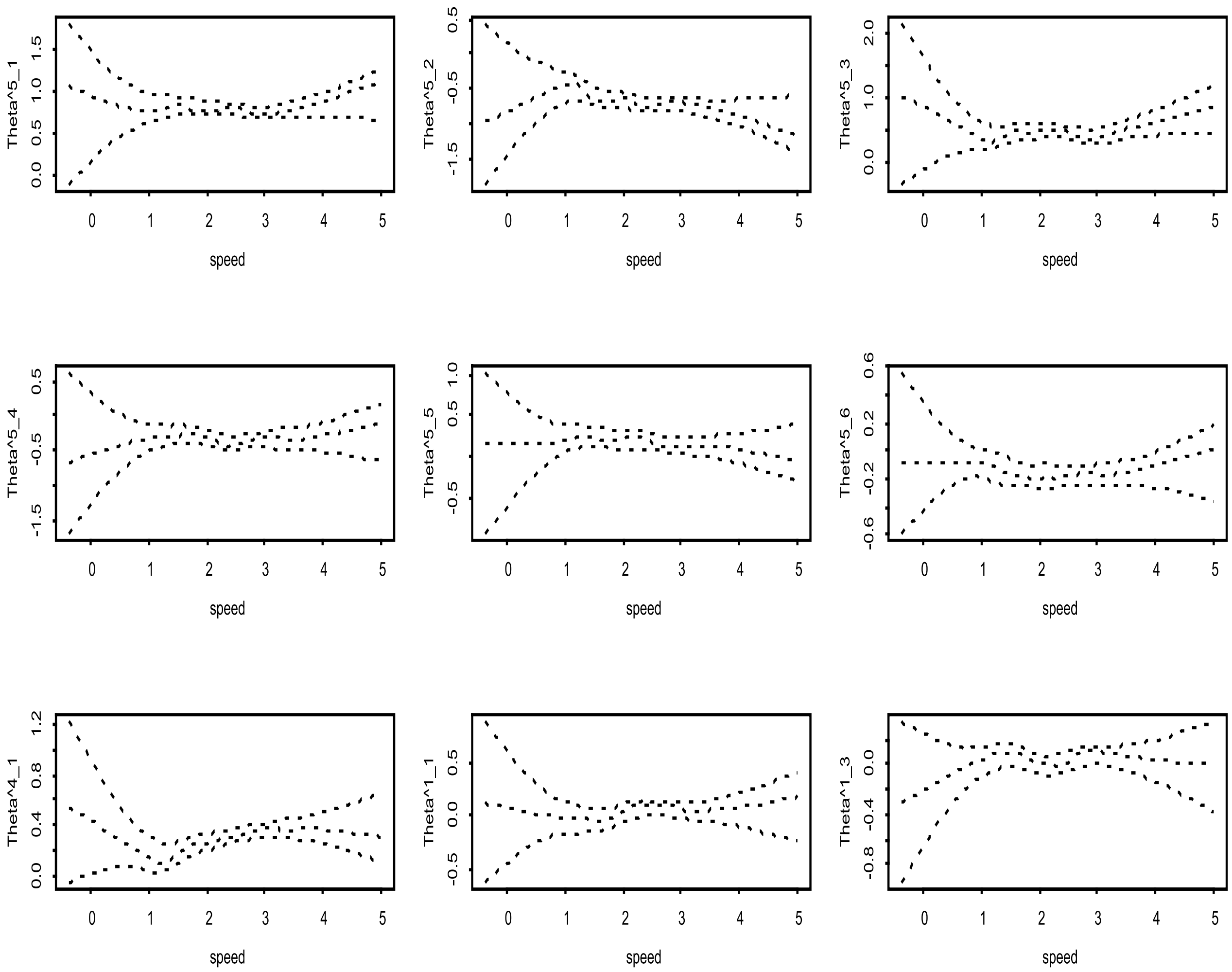

Figure 4.10: Obtained coefficients and 95% standard normal intervals based on 200 bootstrap replicates of regime 4 in model Cond 1. Here Theta^k_j stands for θ_j^k and speed level for s

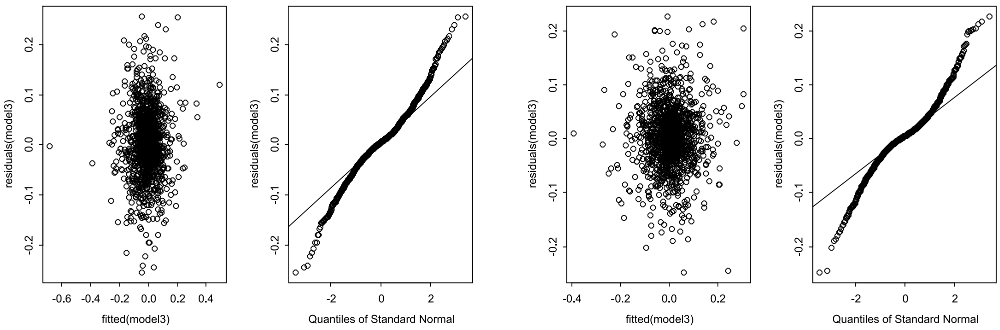

Figure 4.11: Simple diagnostics for the fit corresponding to model Cond 2 regimes 3 (left) and 4 (right)

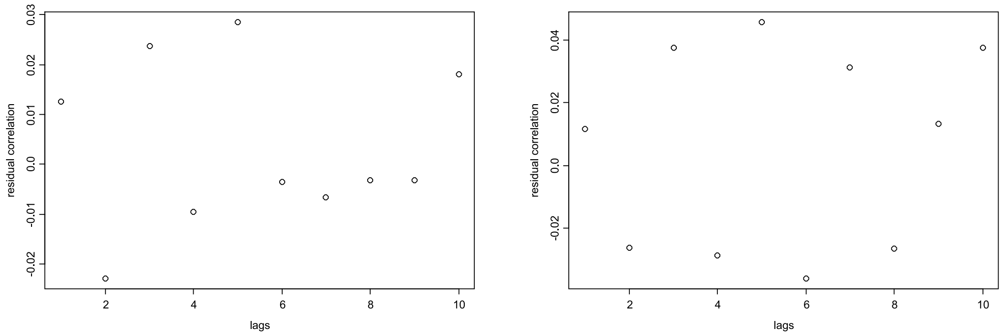

Figure 4.12: Correlation of model Cond 2 regimes 3 (left) and 4 (right) residuals in lags

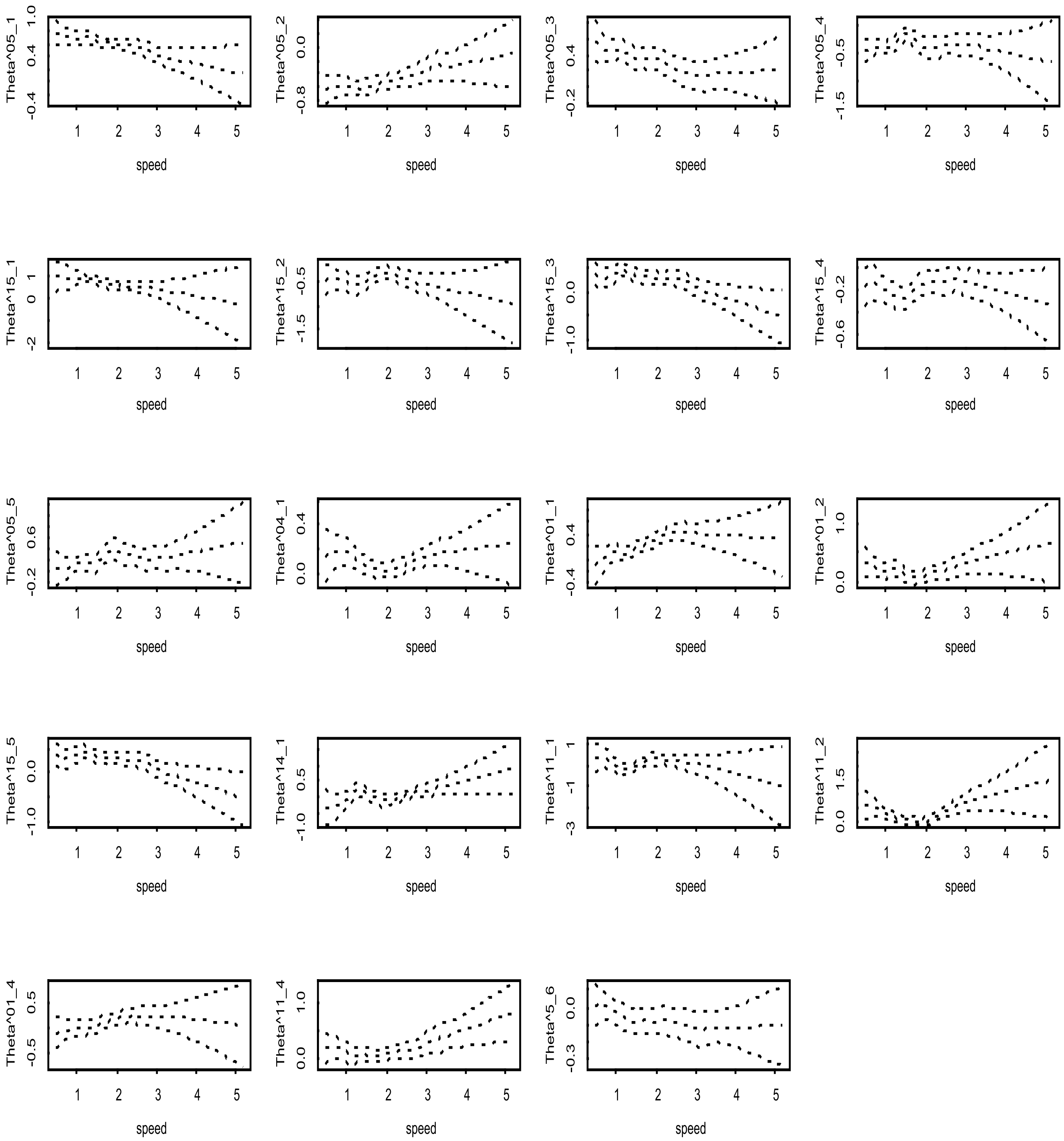

Figure 4.13: Obtained coefficients and 95% standard normal intervals based on 200 bootstrap replicates of regime 3 in model Cond 2. Here Theta^0k_j and Theta^1k_j stand for θ_j^k when $I(t, t-j) = 0$ and $I(t, t-j) = 1$,respectively. s is a wind speed level

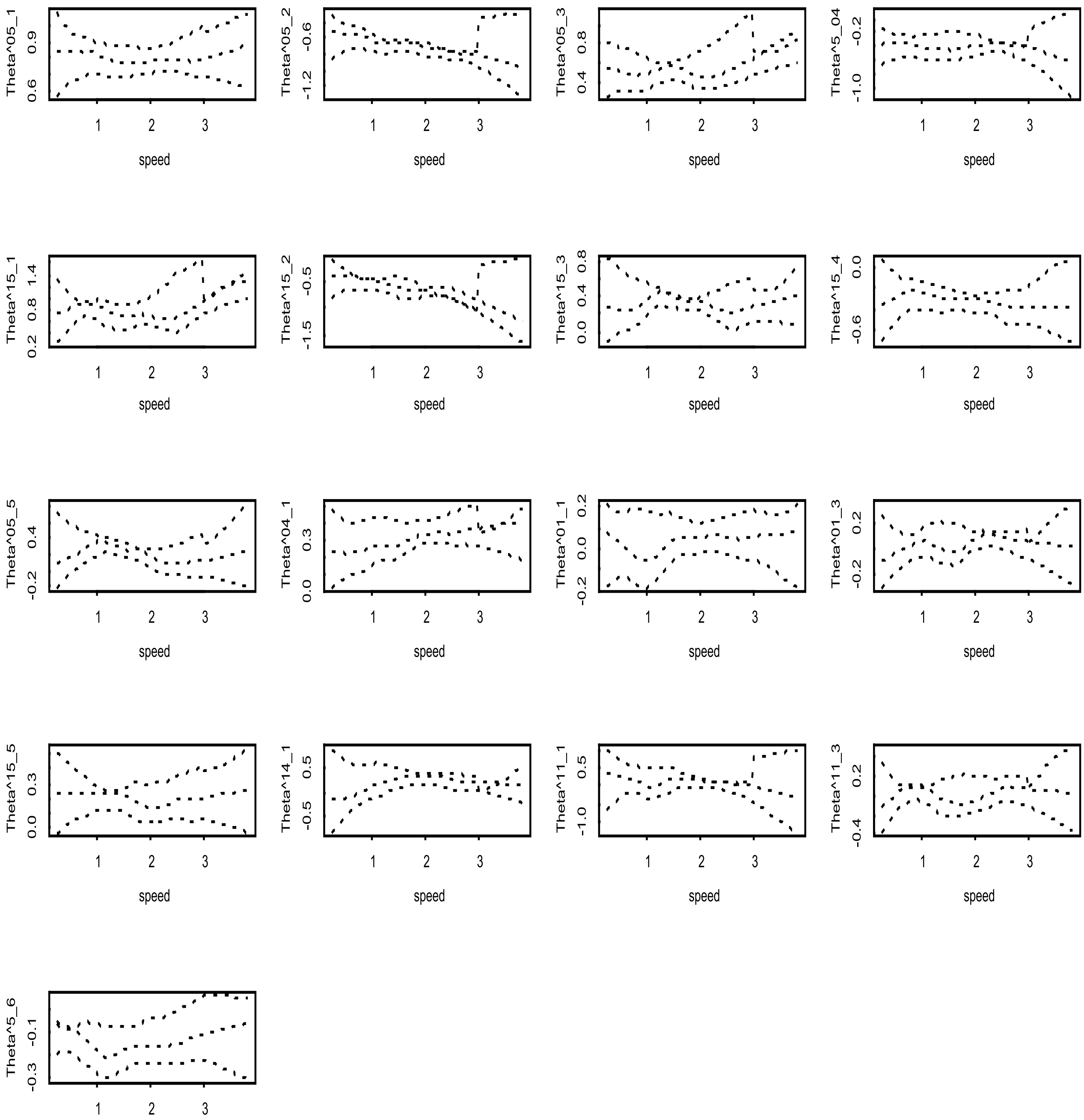

Figure 4.14: Obtained coefficients and 95% standard normal intervals based on 200 bootstrap replicates of regime 4 in model Cond 2. Here Theta^0k_j and Theta^1k_j stand for θ_j^k when $I(t, t-j) = 0$ and $I(t, t-j) = 1$,respectively. s is a wind speed level

Chapter 5

Validation

After having determined the model structure and having estimated parameters, the next step is the diagnosis of the model. During this step one can examine whether the estimations can be described by the model adequately. In the following chapter, methods used in this work for checking the models, will be discussed.

5.1 Residual analysis

One of the important stages of model checking is finding whether the residuals are white noise, that is, showing no serial correlation, homoscedastic, etc. For checking this graphical tools can be used. The first assumption is that the random errors ε_i are normally distributed. Since the random errors can be regarded as a random sample from a $N(0, \sigma^2)$ distribution, we can check whether the standardized residuals might have come from a normal distribution. A normal probability plot of the standardized residuals will give an indication of whether or not the assumption of normality of the random errors is appropriate. Recall that a normal probability plot is found by plotting the quantiles of the observed sample against the corresponding quantiles of a standard normal distribution $N(0, 1)$. If the normal probability plot shows a straight line, it is reasonable to assume that the observed sample comes from a normal distribution. If, on the other hand, the points deviate from a straight line, there is statistical evidence against the assumption that the random errors are an independent sample from a normal distribution. The two assumptions: the random errors ε_i have constant variation and zero mean, can be checked using a residual plot. A residual plot is a scatterplot of the standardized residuals against the fitted values. If the assumptions are satisfied we would expect the residuals to vary randomly around zero and we would expect the spread of the residuals to be about the same throughout the plot.

5.2 Cross-Validation

Cross validation is a model evaluation method that is, on some aspects, better than residual analysis. The problem with residual evaluations is that they do not give an indication of how well the learner will do when it is asked to make new predictions for data it has not already seen. One way to overcome this problem is to not use the entire data set when training a learner. Some of the data is removed before the training begins. Then when training is done, the data that was removed can be used to test the performance of the learned model on "new" data. This is the basic idea for a whole class of model evaluation methods called cross validation.

Holdout method is the simplest kind of cross validation. The data set is separated into two sets, called the training set and the testing set. The parameters of the model are being estimated using the training set only. Then the model is asked to predict the output values for the data in the testing set (it has never seen these output values before). The errors it makes are accumulated as before to give the mean absolute test set error, which is used to evaluate the model. The advantage of this method is that it is usually preferable to the residual method and takes no longer to compute. However, its evaluation can have a high variance. The evaluation may depend heavily on which data points end up in the training set and which end up in the test set, and thus the evaluation may be significantly different depending on how the division is made.

K-fold cross validation is one way to improve over the holdout method. The data set is divided into k subsets, and the holdout method is repeated k times. Each time, one of the k subsets is used as the test set and the other $k-1$ subsets are put together to form a training set. Then the average error characteristic across all k trials is computed. The advantage of this method is that it matters less how the data gets divided. Every data point gets to be in a test set exactly once, and gets to be in a training set $k-1$ times. The variance of the resulting estimate is reduced as k is increased. The disadvantage of this method is that the training algorithm has to be rerun from scratch k times, which means it takes k times as much computation to make an evaluation. A variant of this method is to randomly divide the data into a test and training set k different times. The advantage of doing this is that you can independently choose how large each test set is and how many trials you average over.

Leave-one-out cross validation is K-fold cross validation taken to its logical extreme, with K equal to N, the number of data points in the set. That means that N separate times, the function approximator is trained on all the data except for one point and a prediction is made for that point. As before the average error is computed and used to evaluate the model. The evaluation given by leave-one-out cross validation error is good, but at first pass it seems very expensive to compute.

5.2.1 Results

All the R^2 and RMSE estimates of the fitted models shown previously in this study were obtained for the same data set as was used in estimation steps. We were comparing and giving preferences to one or another model based on those characteristics without putting cross-validation results into the account. However, to draw a final inference, cross-validation test will be performed.

We start comparing the adequacy of the two models which showed the best performance in the previous sections of this study - Thr 4 and Cond 1. 3-fold cross correlation test is applied. The data in each regime is divided into 3 equal subsets. The results are presented in Table 5.1.

Model	subset1		subset2		subset3	
regime	R^2	RMSE	R^2	RMSE	R^2	RMSE
Cond 1, regime 3	0.4785	0.0589	0.4040	0.0778	0.4921	0.0803
Thr 4, regime 3	0.4811	0.0593	0.4778	0.0720	0.4749	0.0796
Cond 1, regime 4	0.5253	0.0558	0.5159	0.0581	0.5700	0.0689
Thr 4, regime 4	0.5341	0.0543	0.5292	0.0556	0.5529	0.0660

Table 5.1: 3-fold cross validation results for Cond 1

Cross-correlation results show that model Thr 4 describes the data more adequately than the Cond 1. R^2 and RMSE values are stable for the model Thr 4 and do not decrease in comparison with the results obtained for the whole data set (see Section 4.2.4). On the contrary, the results of Cond 1 show more abrupt jumps within the constructed subsets and R^2 declines after performing cross-validation test. Another remark is, even though there was significant difference between the R^2 and RMSE values of those models obtained for the entire data set (Table 4.10), the performance on a 'new' data was fairly similar. All the arguments indicate that model Thr 4 describes the data most adequately and efficient among all the models examined in this study.

Chapter 6

Conclusions and Future Work

6.1 Conclusions

In this work new models and methods for improving on-line short-term predictions of wind power were derived and examined. The study was focused on the improvement of the one-hour wind power predictions. However, the methodology used in the analysis, could be applied for a longer-term predictions in the similar manner.

The results of the work showed a great potential in improving the WPPT by modelling the spatio-temporal correlation of the errors. The captured error propagation appeared to be dependent on the forecasted weather situation (mainly wind direction) and geographical position of the wind farms. Methods applied in the project captured a non-linear behavior of error dependency. The best results were obtained while fitting threshold models with regime switching according to the forecasted wind direction. The model adequately explains more than 47% of the error variation for one-hour predictions. During the study more complex models were fitted to data. Varying coefficient models were fitted in order to capture the dependency on the forecasted wind speed level and the time since the last weather forecast was obtained. However, further diagnosis of this approach showed that the improvements it gives in comparison with the threshold model are not significant and even doubtful having in mind the results of the cross-validation tests.

6.2 Future Work

The promising results obtained in the study revealed broad perspectives for the future work. First step could be trying to apply the derived methods for the longer-term predictions.

In this case the information from a larger geographical region could be taken into account. Since a dependency on the weather situation was noticed, it would be of a high interest to include forecast information obtained from different sources. Possibly, it could improve the quality of the predictions, reducing the uncertainty level associated with the weather forecasts.

Several attempts taken in this study aiming to include wind speed level into the model building procedure did not give reliable improvements. However, investigation held in the Identification step indicated that some dependency presents. It could be another push up for the future work to investigate it. One of the possible approach to take could be Markov Regime Switching Model which assumes that the analyzed time series depends not only on the known variables, but also on some hidden processes (see Appendix A). The motivation for this can be the fact that many weather related phenomena are complex and sometimes difficult to capture. Once the Markov approach evidences the dependency on such hidden process, it would be a motivation for further examination and attempts to determine this process. Possibly, wind speed level could somehow be related to this unobserved process since, as already mentioned before, Identification step showed some dependency on it.

Acknowledgments

I wish to express my sincere gratitude to all who have contributed to this master thesis. In particular I want to thank:

- My supervisors and advisers: Algimantas Aksomaitis, Henrik Madsen and Henrik Aalborg Nielsen.
- Extremely kind and positive people from the faculty of fundamental science at Kaunas University of Technology: Vytautas Janilionis and Jonas Valantinas for the encouragement, support and consideration.
- Ewelina Kotwa for fruitful discussions and advices.

Thank you!

Bibliography

[1] Chambers J.M., Hastie T.J. eds (1991) Statistical models i S, Wadsworth, Belmont, CA.

[2] Chatfield C., The Analysis of Time Series, an introduction, fifth edition

[3] Cleveland W. S., Loader C., Smoothing by Local Regression: Principles and Methods

[4] Cleveland W. S., Devlin, S.J.,(1988) Locally Weighted Regression: An approach to regression analysis by locall fitting, Journal of American Statistical Assocination 83, 596-610.

[5] Eiliers P. H. C., Marx B.D. (2004) Splines, Knots and Penalties,

[6] Gregor Giebel The State-Of-The-Art in Short-Term Prediction of Wind Power. A literature Overview. Project ANEMOS, cont. nr ENK5-CT-2003-00665

[7] Hamilton James D. Time Series Analysis, (1994) Princeton University Press, Princeton.

[8] Hastie, T. J. and Tibshirani, R.J. Varying-coefficient models.

[9] Hastie, T. J. and Tibshirani, R.J. (1990) Generalized Additive Models, Chapman and Hall, London/ New York.

[10] Katinas V., Markevičius A., Burlakovas A. 2006. Vėjo energetika ir jos artimiausia perspektyva Lietuvoje. Energetika, Nr. 3, p.67-76

[11] Madsen H. (2006) Time Series Analysis 2nd edition, DTU Lyngby.

[12] Madsen H., Holst J. (2000), Modelling Non-Linear and non-Stationary Time Series, IMM.

[13] Mantas Marčiukaitis, Vėjo elektrinių galios prognozės modeliai ir jų pritaikymo galimybės Lietuvoje, Lietuvos energetikos institutas 2007

[14] Millais Corrin The leading role of wind energy in Europe, EWEA, EU Power - issue 2/2005

[15] Nielsen H.A. (1997) An S-PLUS/ R library for locally weighted fitting of linear models

[16] Perce D. A., "R^2 Measures for Time Series", Journal of the American Statistical Association, June 1979, Vol. 74, 901-909.

[17] Pinson P., Madsen H. eds, (2007) Regime-switching modelling of the fluctuations of offshore wind generation,

[18] Plukas K. (2001), Skaitiniai metodai ir algoritmai, Kaunas, 272-420.

[19] Sampson P.D., "Comment on Splines and Restricted Least Squares", Journal of the American Statistical Association, June 1979, Vol. 74, 303-305.

[20] Tong H. (1990), Non-Linear Time Series -A Dynamical System Approach, Oxford University Press.

Appendix A

Markov Regime Switching Model

While working on this thesis, another model has been examined. The time frames though, did not allow to apply it fully. Therefore, it will just be discussed here briefly and stated as a possible field of future research.

The last model of this paper is then Markov Switching Auto Regressive model (MSAR). A broader description of this topic can be found in the literature ([7], [12], [17]).The model assumes that the change of regime is governed by an unobservable process, which in this case is a two-state Markov Chain. This idea depicts the influence of some external signals which in the context of weather phenomena are perfectly justified. In the figure (A.1) we can observe that the time series seem to have a dual character. There are periods when variability is very low and those, for which the process experiences more abrupt jumps. In the following chapter, we only describe the theoretical part concerning MSAR models, involving model structure and estimation with the famous EM-algorithm.

A.1 Modelling

First, for sake of complexity, let's recall how the Markov Chain is defined. Let $\{s_t\}$ be a sequence of realizations of a random variable with a finite space $\{1, 2, ..., R\}$. We say that $\{s_t\}$ follows a first order Markov Chain when the following equation is fulfilled:

$$P(s_t = j | s_{t-1} = i, s_{t-2}, ..., s_0) = P(s_t = j | s_{t-1} = i), \forall i, j, t \tag{A.1}$$

It means that the transition between the states depends only on the most recent value of the process. The above probability is also called the transition probability, since it represents the probability that state i will be followed by state j, and is denoted as p_{ij}. The transition matrix

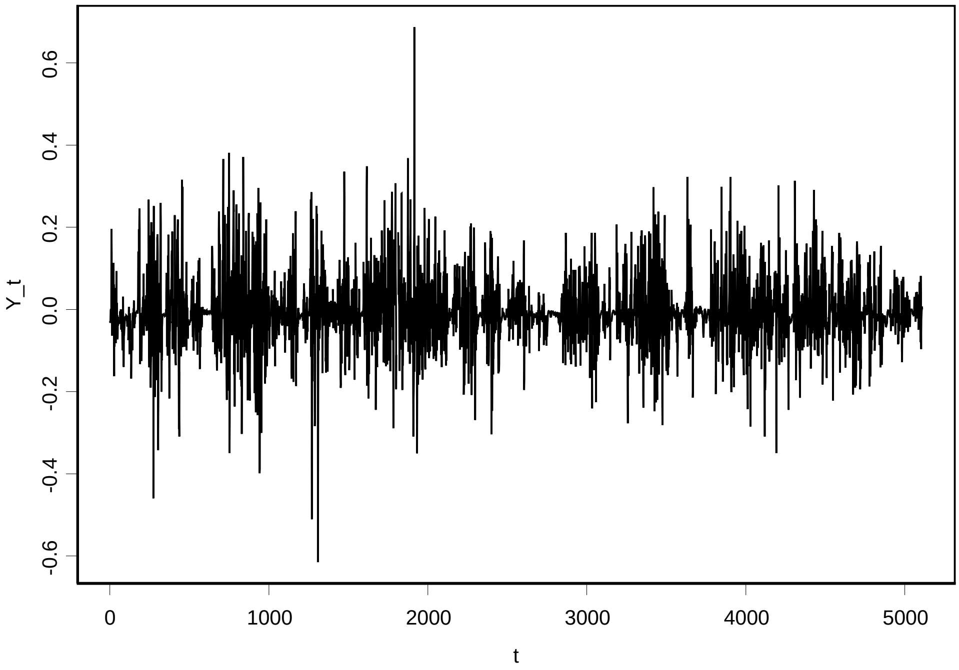

Figure A.1: Prediction errors at $Group\,5$

$\boldsymbol{P}$ gathers all transition probabilities in a following manner:

$$P = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1R} \\ p_{21} & p_{22} & \cdots & p_{2R} \\ \vdots & \vdots & \ddots & \vdots \\ p_{R1} & p_{R2} & \cdots & p_{RR} \end{pmatrix} \tag{A.2}$$

The practice says that the most effort in estimation process will be required by the transition matrix. We also notice that

$$\sum_{j=1}^{R} p_{ij} = 1, \forall i \tag{A.3}$$

Additionally, we assume that the chain is ergodic, which means that all states can be reached and none is the final one. For this reason the following condition can be formulated:

$$p_{ij} > 0, \forall i, j \tag{A.4}$$

Now, we can define our process as following:

$$y_t = \theta_0^{(s_t)} + \sum_{i=1}^{p_{s_t}} \theta_i^{(s_t)} y_{t-i} + \sigma_{s_t} \epsilon_t \tag{A.5}$$

where s_t is the Markov Chain of regime sequence mentioned above, ϵ_t is a white noise sequence, θ_i is the coefficient standing by the i^{th} AR part and σ_k is a standard deviation of the residuals in the k^{th} regime. We assume that $Var(\epsilon_t) = 1$. Moreover, we define the model parameters set as below:

$$\Theta_m = (\boldsymbol{\theta}^{(1)}, ..., \boldsymbol{\theta}^{(R)}, \sigma, \boldsymbol{P})^T \tag{A.6}$$

The matrix $\boldsymbol{P}$ is the same as in (A.2). The AR parameters are gathered in vectors equivalent for each regime

$$\boldsymbol{\theta}^j = (\theta_0^{(j)}, \theta_1^{(j)} ..., \theta_{p_j}^{(j)})^T, k = 1, ..., R \tag{A.7}$$

and finally σ contains the standard deviations of the noise in all regimes:

$$\boldsymbol{\sigma} = (\sigma_1, \sigma_2, ..., \sigma_R)^T \tag{A.8}$$

A.2 Estimation

Since the process that governs the change of the regime cannot be observed, we are not able to say for sure in which of them the time series actually is. This is the first, quite obvious problem which appears. In order to solve it, in the next paragraph the inference about which regime was more likely to give a considered observation, has to be made. After dealing with that we proceed with a proper estimation of parameters set. As in ([7]), the Maximum Likelihood Estimates and the EM Algorithm has been used. The algorithm is an iterative method of estimating the parameters of MSAR by successively repeating each of its two steps: Expectation and Maximization.

A.2.1 Inference about the Unobserved Regime

As mentioned before, we can never be sure about the origin of each observation. For this reason we introduce (after Hamilton) the idea of filtered probability, which can be interpreted as a probability that the process is in state j at time t, given all the information available until that moment and the parameters set. We copy the notation from ([17]).

$$\hat{\xi}_{t|t}^{(j)} = P(s_t = j|\Omega_t, \Theta_m) \tag{A.9}$$

From the definition of conditional probability, we know that

$$\hat{\xi}_{t|t}^{(j)} = \frac{f(y_t, s_t = j|\Omega_{t-1}, \Theta_m)}{f(y_t|\Omega_{t-1}, \Theta_m)} \tag{A.10}$$

While proceeding with computation we arrive at the following expressions of the nominator and denominator, respectively:

$$f(y_t, s_t = j|\Omega_{t-1}, \Theta_m) = \sum_{i=1}^{R} \hat{\xi}_{t|t-1}^{(i)} f(y_t|s_t = j, \Omega_{t-1}, \Theta_m) \tag{A.11}$$

and

$$f(y_t|\Omega_{t-1},\Theta_m)=\sum_{j=1}^{R} f(y_t,s_t=j|\Omega_{t-1},\Theta_m) \tag{A.12}$$

where we recognize $\hat{\xi}^{(i)}_{t|t-1}$ as a forecast of the probability of the Markov Chain being in state i, and moreover, we know, that assuming Gaussianity of the noise sequence, the conditional density of y_t given $s_t = j$, Ω_{t-1} and Θ_m can be presented as follows:

$$f(y_t|s_t=j,\Omega_{t-1},\Theta_m)=\frac{1}{\sqrt{2\pi}\sigma_j}exp\left(-\frac{(y_t-\boldsymbol{x}^T_{t,j}\boldsymbol{\theta}^{(j)})^2}{2\sigma_j^2}\right) \tag{A.13}$$

with $\boldsymbol{x}^T_{t,j}=(1,y_{t-1},...,y_{t-p_j})$. In order to make the forthcoming formulation of the filter more readable, we will use the vectorial notation. Following this idea, the filtered probability for every regime can be presented in the following way:

$$\hat{\boldsymbol{\xi}}_{t|t}=\left(\hat{\xi}^{(1)}_{t|t},\hat{\xi}^{(2)}_{t|t},...,\hat{\xi}^{(R)}_{t|t}\right)^T \tag{A.14}$$

In addition, let η_t contain the conditional density of y_t for every state of the Markov Chain.

$$\eta_t=(f(y_t|s_t=1,\Omega_{t-1},\Theta_m),...,f(y_t|s_t=R,\Omega_{t-1},\Theta_m))^T \tag{A.15}$$

For this notation, we can finally state the Hamilton Filter, consisting of 2 equations, by which recursive application we obtain the vector of the filtered probabilities. The equations are given by:

$$\hat{\boldsymbol{\xi}}_{t|t}=\frac{\hat{\boldsymbol{\xi}}_{t|t-1}\otimes\eta_t}{1^T_R(\hat{\boldsymbol{\xi}}_{t|t-1}\otimes\eta_t)} \tag{A.16}$$

$$\hat{\boldsymbol{\xi}}_{t|t-1}=\boldsymbol{P}^T\hat{\boldsymbol{\xi}}_{t-1|t-1} \tag{A.17}$$

where $\otimes$ denotes element wise multiplication.

A.2.2 EM Algorithm - estimation of model parameters

As a first step of EM Algorithm we will use the vector of the filtered probabilities obtained in the previous section. This is so called Expectation step. The second part, consists of applying three update equations, for $\boldsymbol{\theta}^{(j)}$, $\boldsymbol{\sigma}$ and the matrix $\boldsymbol{P}$. The asymptotical results show that each iteration of EM Algorithm increases the value of log-likelihood function for the parameters. In order to begin the algorithm we need to set the initial values of all parameters as well as the firs value of the filtered probability $\hat{\boldsymbol{\xi}}_{1|0}$.

Before proceeding with updating our parameters set, it is required to calculate the vector of smoothed probability for which we use the fact that in the moment of estimation, the whole data set (all observations) are known. It can be formulated as below:

$$\hat{\boldsymbol{\xi}}_{t|T} = \hat{\boldsymbol{\xi}}_{t|t} \otimes \left(\boldsymbol{P} \left(\hat{\boldsymbol{\xi}}_{t+1|T} \oslash \boldsymbol{P}^T \hat{\boldsymbol{\xi}}_{t|t} \right) \right) \tag{A.18}$$

with $\otimes$ and $\oslash$ being the element wise multiplication and division respectively. It's a straightforward conclusion that the calculation should be made recursively but backwards.

After obtaining the sequence of smoothed probabilities, we can re-estimate transition matrix by applying for each of its elements

$$\hat{p}_{ij} = \frac{\sum_{p_{max}+1}^{T} \hat{\xi}^j_{t|T} \hat{\xi}^i_{t-1|T}}{\sum_{p_{max}+1}^{T} \hat{\xi}^i_{t-1|T}} \tag{A.19}$$

where p_{max} is the maximum lag in AR models. Now, only the coefficients of the AR part and the variance of each regime should be up-dated. In order to do that, we will use Weighted Least Squares method, with the smoothed probabilities of the observations being in a particular regime as weighting vector. Without getting into details and avoiding derivation, the ready-to-apply formula will be given below:

$$\hat{\boldsymbol{\theta}}^j = (\tilde{\boldsymbol{x}}_j^T \Sigma_j \tilde{\boldsymbol{x}}_j)^{-1} \tilde{\boldsymbol{x}}_j^T \Sigma_j \boldsymbol{y}_j \tag{A.20}$$

where

$$\Sigma_j = \begin{pmatrix} \hat{\xi}^j_{1|T} & & 0 \\ & \ddots & \\ 0 & & \hat{\xi}^j_{T|T} \end{pmatrix} \tag{A.21}$$

$$\tilde{\boldsymbol{x}}_j = \begin{pmatrix} \boldsymbol{x}^T_{1,j} \\ \vdots \\ \boldsymbol{x}^T_{T,j} \end{pmatrix} = \begin{pmatrix} 1 & y_{p_j} & \dots & y_1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & y_{T-1} & \dots & y_{T-p_j} \end{pmatrix} \tag{A.22}$$

and

$$\boldsymbol{y}_j = \begin{pmatrix} y_{p_j+1} \\ \vdots \\ y_T \end{pmatrix} \tag{A.23}$$

The last part is to recalculate the variance of freshly established noise values. The following equation deals with that:

$$\hat{\sigma}_j^2 = \frac{1}{\sum_{t=1+p_{max}}^{T} \hat{\xi}^j_{t|T}} \sum_{t=1+p_{max}}^{T} \left(y_t - \boldsymbol{x}^T_{t,j} \hat{\boldsymbol{\theta}}^{(j)} \right)^2 \hat{\xi}^j_{t|T} \tag{A.24}$$

which is just a standard way of calculating the variance form the model fit summing the squared residuals in j^{th} regime corrected by the smooth probabilities.