Delft University of Technology
Faculty of Electrical Engineering, Mathematics and Computer Science
Delft Institute of Applied Mathematics

Master of Science Thesis

Wind speed modeling

by

Alicja Lojowska

Delft, the Netherlands
August 2009

Members of the committee

Chairperson of Graduate Committee:

Prof. dr. R. M. Cooke
Delft University of Technology, the Netherlands
Resources for Future, United States

Committee:

Prof. dr. R. M. Cooke
Delft University of Technology, the Netherlands
Resources for Future, United States

Prof. ir. L. van der Sluis (supervisor)
Delft University of Technology, the Netherlands

Dr. ir. D. Kurowicka (daily supervisor)
Delft University of Technology, the Netherlands

Dr. ir. G. Papaefthymiou (daily supervisor)
Delft University of Technology, the Netherlands

Dr. ir. E. Cator
Delft University of Technology, the Netherlands

Contents

List of Figures

List of Tables

Chapter 1

Purpose of the project

Nowadays, the renewable energy is gaining more and more interest. Mainly due to the fact that renewable generation offers an advantage of being environmentally friendly and its source is available practically everywhere. In order to allow for high stochastic renewable generation, power systems have to follow structural changes in the existing distribution and transmission network. The power systems with a few large power stations supplying the transmission network will be replaced by systems characterized by decentralized structure with many small-scale integrated generators (e.g. windmills) [16]. However, the sources such as wind and sun are in their nature uncontrollable and consequently the electrical power output cannot be controlled as well. Large scale implementation of this type of generation may cause strong power fluctuation in the grid. The idea of energy storages seems to provide a good solution to this problem. Such a storage would prevent wasting of energy in cases of surplus and would supply it back to the system when a deficiency occurs.

The implementation of energy buffers in future power systems will face many significant issues like e.g. the optimal distribution of energy storages, their capacity, operational strategies and their impacts in a system. The project "Role of Energy Storages in Future Power Systems", which will start at the beginning of September 2009, is going to develop methodologies for the planning and operation of power systems with large quantities of small-scale distributed storage devices. For this purpose, power system models, energy storage models and time series models will be developed and joined together in a one energy management model which will be further used in simulation studies.

The mentioned time series models will represent a stochastic renewable

resources (e.g. wind) and system energy consumption (load). The time series model for wind speed will have to be able to simulate wind speeds in all wind farms of interest. However, it is not feasible to build so highly dimensional structure since it would be too complex and financially challenged. Therefore a certain assumption should be made regarding the number and distribution of sites taken into account. It will certainly lead to greater simplicity and usefulness. In such a model, each considered site would be a representation of a certain area. It is very reasonable assumption since the Netherlands are a small country and places which are relatively close are characterized by almost identical wind speeds.

The goal of the thesis is to develop a univariate and multivariate model for wind speed by means of statistical tools. Although the made assumption regarding the number of sites, their amount will be still quite significant from modeling point of view. Therefore the developed model cannot be too much elaborate but should captures the most important characteristics of the data. We will propose a methodology which will help in building a model satisfying aforementioned properties. However, we will present our results for at most three sites which provide a good representation of inland, onshore and offshore wind speeds. We believe that the generalization to higher dimension will not raise problems thanks to paying attention to simplicity.

The artificial time series of wind speed will be further transformed to obtain respective time series of wind power. In this way, if wind speed/power and load are simulated simultaneously, their difference will be nothing but the input to an energy storage. Specifically, denoting a generation at time t X_t and load at time t L_t, the input(or output) to(from) energy storage is $D_t = X_t - L_t$ and will play a crucial role in the assessing an optimal dimension of energy storage device. It also shows the importance of time series analysis which gives insight into the evolution of the processes in time in contrast to the frequency distributions which would be not enough for the analysis of storage systems.

Chapter 2

Wind speed data

The research was conducted on the basis of data provided by Royal Netherlands Meteorological Institute (KNMI). This high quality dataset comprising 53 long station records of near-surface wind in the Netherlands was developed within the framework of the HYDRA (hydraulic conditions) project [1]. It is, however, not measured wind speed at the anemometer's height but the so-called potential wind speed which has a few advantages over the former. The Figure 2.1 presents the KNMI measuring network. The potential wind speed is defined as corrected hourly averaged records so that they correspond with wind speed at 10 m height over short grassland free of obstacles. The correction involves also weakening of the influence of inhomogeneities that may have occurred in data due to changes in observational routines, station relocations or measuring techniques. Therefore, the correction gives us the possibility to compare the data recorded at different station which is very important in multivariate modeling of wind speed. The detailed information on the corrections made to the data can be found in [2].

The analysis of the data has revealed a few main features of the wind which are explainable in meteorological and physical terms and are discussed in the following sections. For the purpose of presenting wind speed characteristics, data from two stations were chosen: Schiphol-inland station and Europlatform- offshore station.

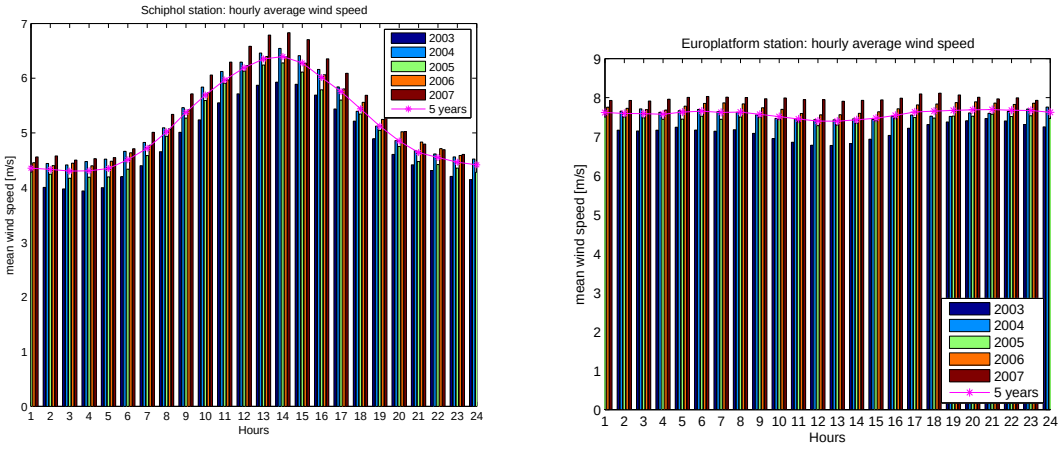
2.1 Diurnal and annual variations of the wind

The diurnal variations of wind speed have been inferred from the observations collected at Schiphol station. Figure 2.2a shows the diurnal variations

Figure 2.1: The KNMI measurement network

of mean wind speed (averaged over several years). The wind speed increases after sunrise, attains a maximum at 1-2 pm and then gradually decreases. This variation appears because the temperature differences between the sea surface and the land surface are larger during the day than during night. This meteorological fact seems to be an advantage since the peak values of the wind happen at similar hours as the peak values of electricity consumption (load). Furthermore, it can be noticed that the diurnal wave is very similar for each year considered (2003-2007 years) but the hourly averages differ between years indicating that e.g. year 2007 was exposed to stronger wind than year 2003. When one considers hourly averages at offshore station, no such diurnal variations are visible, see Figure 2.2b.

As the temperatures during winter are lower than in the other seasons,

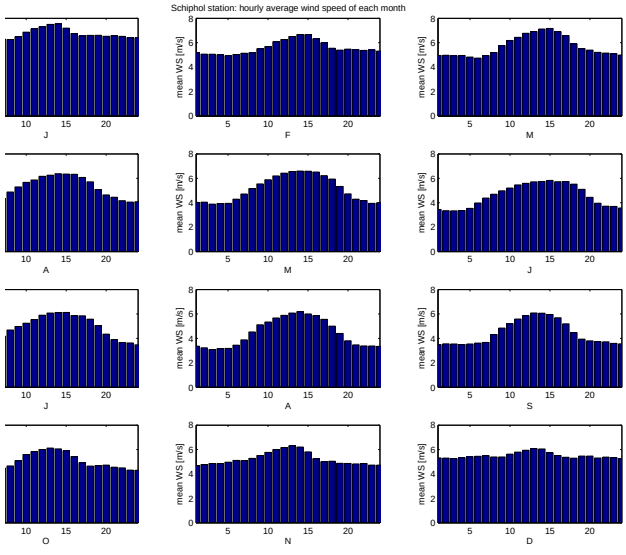


(a) Schiphol

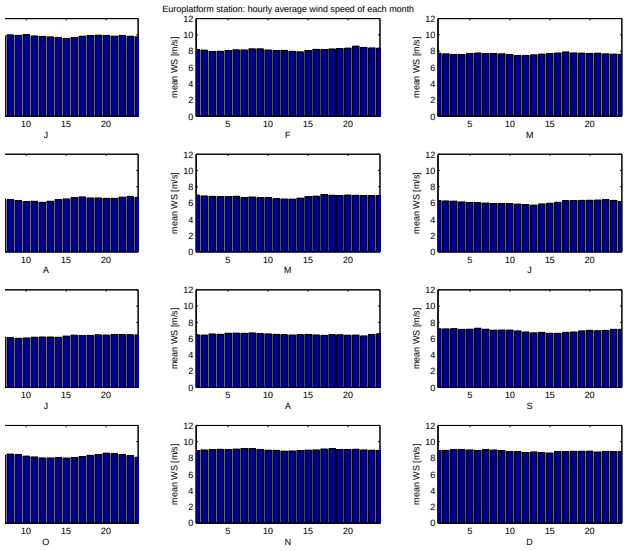
(b) Europlatform

Figure 2.2: Diurnal periodicity

one would expect that diurnal variations of inland wind may be less visible in winter. Indeed, the Figure 2.3a confirms that December and January enjoys approximately the same level of hourly averaged values which are the same as peak values in summer months. Taking into consideration the data from Europlatform, the diurnal variation does not appear but there are differences in mean values between months. The wind speed on offshore stations blows stronger than on land due to lower roughness of a surface and lack of obstacles which contribute to wind turbulence. However, it is more expensive to build wind farm on the sea where turbines are exposed to tough conditions due to salt, seawater, waves and strong wind.



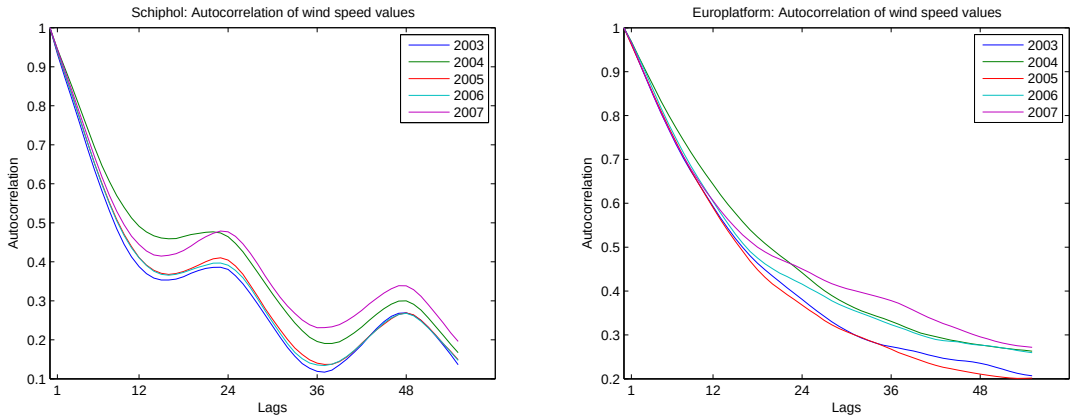
(a) Schiphol



(b) Europlatform

Figure 2.3: Diurnal periodicity in each month

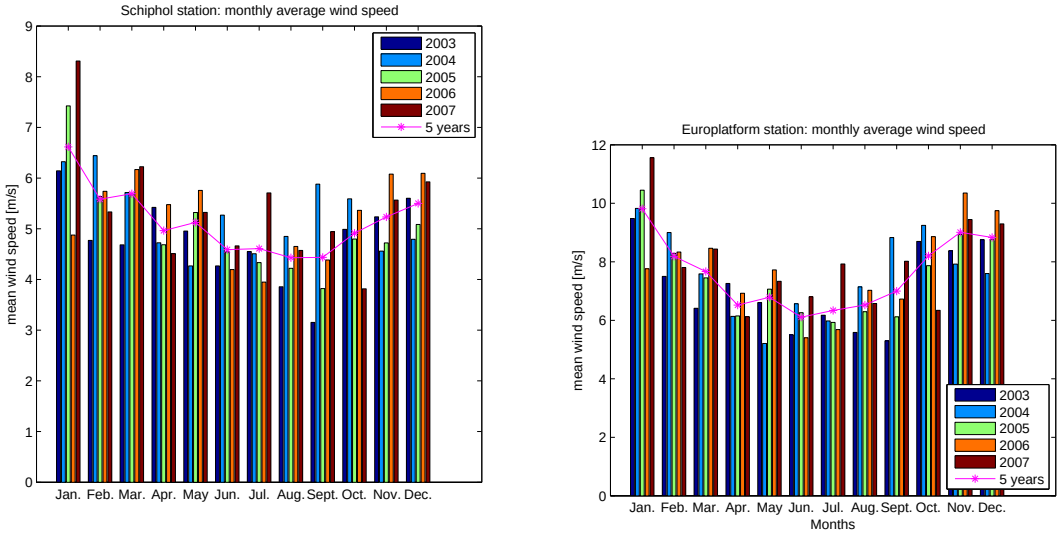
So far the diurnal variation was presented using mean values for each hour. The pattern of the daily periodicity can be also noticed in the autocorrelation of the data sequence in each considered year, see Figure 2.4a. The autocorrelation function reaches its highest values every 24 hours as expected. Moreover, the autocorrelation function obtained from 2007 year gets usually the highest values among all years, whereas the function for 2003 year, the lowest. When considering hourly averages, year 2003 turned out to have lower means than year 2007 which, together with the latter observation, would suggest that the degree of similarity between observations separated by a certain time span is higher when the wind speed observations are (in mean) higher. The data from Europlatform still does not reveal any diurnal variations according to Figure 2.4b. As it was already noticed, the strength of wind speed depends on the season of the year and is described below.



(a) Schiphol (b) Europlatform

Figure 2.4: Autocorrelations

The Figures 2.5a-2.5b present monthly averages for several years and for five years all together. One can see that summer winds are generally weaker than winter winds for inland and offshore stations. Furthermore, just like the diurnal variation matches the diurnal electricity consumption, the annual pattern in wind speed matches the pattern in electricity demand. Because of the periodicity in further part of the project we will consider months separately.



(a) Schiphol (b) Europlatform

Figure 2.5: Annual variations

2.2 Wind speed distribution

There are several probability distribution proposed for modeling wind speed e.g. Weibull, Lognormal, Gamma. However, usually Weibull distribution is chosen for modeling wind speed [21]. The Figure 2.6 shows the histogram of wind speed recorded at Schiphol in January 2003. The Kolmogorov-Smirnow test checking the adequacy of the Weibull model has indicated that Weibull distribution does not describe the data well.

The Figure presents a very important issue. Namely we can notice that the accuracy of the measurements increases with the wind speed. It may indicate that a meteorological station uses rather simple anemometer to obtain the measurements and therefore a relatively light wind speed posses high measurement error.

2.3 Persistence

The notion of wind persistence appears frequently in the literature concerning wind engineering, climatology or energy. However, the definition of persistence differs from author to author. We define the persistence as in [26], [27] in the context of the duration of surface wind speeds within

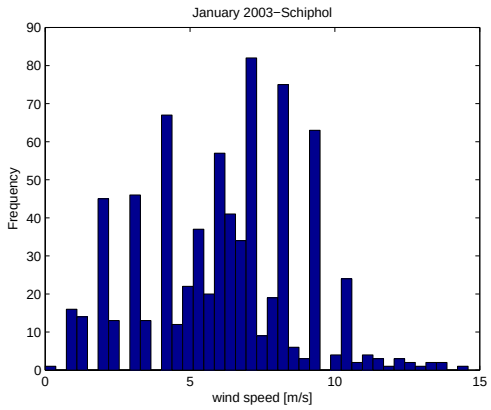


Figure 2.6: Histogram of wind speeds recorded at Schiphol in January 2003

specified wind speed classes. Thus, the analysis of persistence is a way of exploring the dynamic features of the wind speed.

The length of the periods of no power or of full power generated from wind correspond to the length of the periods of low and high wind speed respectively. Therefore, from power systems standpoint, it is of high importance to analyze the distribution of the run durations below or above a certain threshold. However, we should take a note that the number of excursions from a given threshold will depend very much on this threshold and consequently, the computed distributions will be in some cases of little reliability. The analysis of persistence usually is performed as a part of assessing the wind resources at the sites of interest since the reliability and predictability of generated power has implications for network design and meshing of technologies to meet electricity demand. It gives opportunity to compare power quality at sites characterized by different weather conditions or/and different topography [26], [28], [29], [30]. Usually zones with better attributes from power perspective require higher costs which have to be involved in building a wind farm, connecting it to the grid and maintenance. Therefore the analysis of persistence contributes to deciding whether the expensive wind farm will in future give enough big profits to cover at least the investment costs.

In the case of this project, the persistence is considered due to playing important role in the issue of energy storages in power systems. Therefore it will be entertained for the purpose of the comparison between the synthetic and real data sequence but also between the months: January and July. The Figure 2.7 presents the above-threshold excursions (red dots). In the

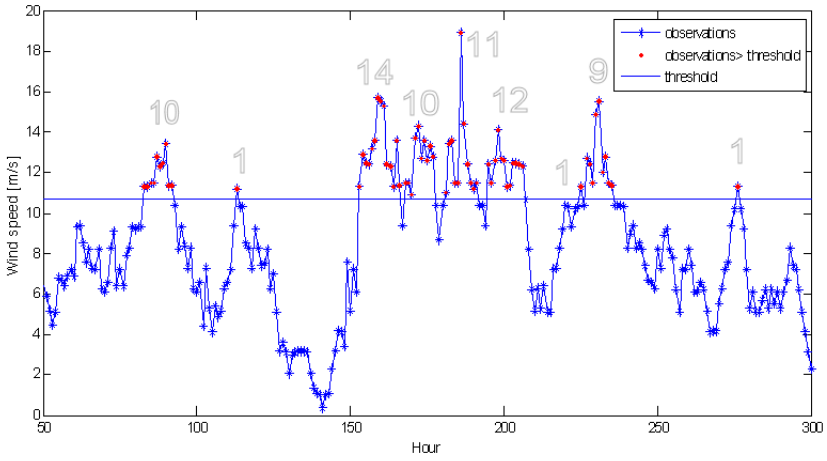


Figure 2.7: Persistence: The excursion lengths

persistence analysis we will be interested mainly in the excursion lengths (big font numbers) but also in the number of excursions. The high interest in wind energy has led to some developments of the models for persistence e.g. [27] proposes a composite distribution for modeling the distribution of excursions lengths which consists of a power function distribution and scaled exponential distribution. Another way of modeling persistence is presented in [31] where the Markov theory is applied.

Chapter 3

Univariate case

From a statistical standpoint a given wind speed time series $\{Y_t\}$ is a realization of a stochastic process. The goal is to construct stochastic models which could define a mechanism responsible for producing such sequences of values. Stochastic, as opposed to deterministic, means that if it were used to generate several sets of observations over the same time period, each set of observations would be different, but they would all obey the same probabilistic laws. Usually, when we deal with recorded measurements, we have only one realization of the process. However, given a set of data covering a few years we can assume that wind speeds corresponding to a certain month e.g. July are realizations of one unknown process. Hence, taking into analysis measurements from period 2003-2005 will result in five realizations of the process governing in July. The Figure 3.1 presents wind speed observations in July in years 2003-2005. Although the mentioned assumption of multiple realizations lacks verification, it is important to make use of information that is available to us. This idea will be entertained throughout the thesis and will lead to development of an averaged model representing a process in given month. More details regarding this issue will be presented further.

The first stage in modeling the sequence of observations called time series is specification of the class of models which could appropriately capture the most important characteristics of the data. When we deal with wind speed measurements, we intuitively know that the strength of wind speed at present is somehow dependent on the wind speed in the preceding period. The theory of time series analysis was established mainly by Box and Jenkins who described in detail the approach to building model [3]. The following sections will present the methodology used to model wind speed

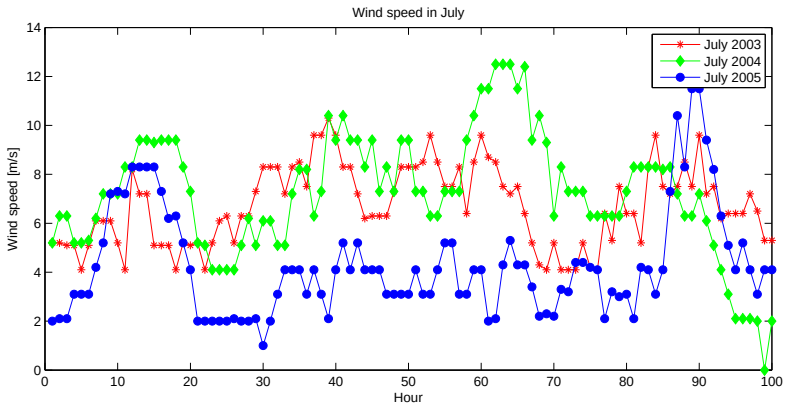


Figure 3.1: Wind speed observations in July in 2003-2005

and statistical tools for building, estimating and verifying the model. The theoretical information was written on the basis of the time series theory described in several books: [3], [9]-[13], [17]. For more detailed description and references the reader is referred to them.

3.1 Methodology

Recently wind speed modeling has gained much interest especially due to growing demand on wind power. The researchers usually propose simple ARMA models [14], [19]-[23] but more complex models start being common [5]-[6], [24]-[25]. A thoroughful analysis of the wind speed time series has led us to proposing an ARMA-GARCH model as a model capturing the wind speed characteristics appropriately. ARMA-GARCH was developed for modeling financial time series and finds application mainly in that field. The applications to other time series types are rarely seen, especially concerning wind speed (only [5] and [6]).

Since the present level of wind speed (Y_t) depends on its immediate past we describe this univariate time series Y_t by the process

$$Y_t = E(Y_t|\Omega_{t-1}) + \varepsilon_t$$

where $E(\cdot|\cdot)$ denotes the conditional expectation operator, Ω_{t-1} the information set at time $t-1$, and ε_t the innovations or residuals of the time series which are uncorrelated, have mean zero and play the role of the unpredictable part of the time series. We should here take a note that we

consider only wind speed observations; wind speed direction, temperature etc. will not appear in the mean equation. In the case of ARMA-GARCH model, the mean equation is the ARMA process and the innovations are generated from the GARCH model. The ARMA(p,q) mean equation has the following (mean adjusted) form [1]:

$$Y_t - \mu = \sum_{i=1}^{p} \phi_i (Y_{t-i} - \mu) + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} + \varepsilon_t \tag{3.1}$$

$$= \Phi(B)(Y_t - \mu) + \Theta(B)\varepsilon_t$$

where μ is mean of time series, ϕ is autoregressive coefficient and θ is moving average coefficient. Moroever, the equation is also expressed using backshift operator B i.e. $BY_t = Y_{t-1}$. The functions $\Phi(B)$ and $\Theta(B)$ are polynomials of degree p and q respectively in the backward shift operator B. When $q = 0$ we have a pure autoregressive process and when $p = 0$ pure moving average process.

It may happen that squared residuals exhibit significant serial correlation. It indicates that errors are not independent although they are serially uncorrelated. Residuals are called then conditionally heteroscedastic and GARCH (Generalized Autoregressive Conditional Heteroscedasticity) models have been proved to be very successful at modeling the serial correlation in the second moment of the underlying time series. The formulation of the GARCH model for errors ε_t is:

$$\varepsilon_t = z_t \sigma_t$$

$$z_t \sim \mathcal{D}(0, 1)$$

$$Var(\varepsilon_t | \Omega_{t-1}) = E(\varepsilon_t^2 | \Omega_{t-1}) = \sigma_t^2 = K + \sum_{j=1}^{a} \alpha_j \sigma_{t-j}^2 + \sum_{i=1}^{b} \beta_i \varepsilon_{t-1}^2 \tag{3.2}$$

$$= K + \alpha(B)\sigma_{t-1}^2 + \beta(B)\varepsilon_{t-1}^2$$

where z_t are iid random variables with zero mean and unit variance and $\mathcal{D}$ is their probability density function. Common choices for density function $\mathcal{D}$ are normal distribution, student-t distribution and generalized error distribution. Thus, under the GARCH(a,b) model, the conditional variance of

[1] An another frequently used form in the literature is $Y_t = \mu + \Phi(B)Y_t + \Theta(B)\varepsilon_t$, but in this case μ does not stand for the mean of time series

ε_t, σ_t^2 (3.2), depends on the squared residuals in the previous b periods, and the conditional variance in the previous a periods. Usually a GARCH(1,1) model with only three parameters in the conditional variance equation is adequate to obtain a good model fit. The GARCH model was proposed by Bollerslev [8] as a generalization of ARCH model which was initially proposed by Engle [7] and corresponds to GARCH(0,b). The ARCH model described by the conditional variance equation:

$$\sigma_t^2 = K + \beta_1 \varepsilon_{t-1}^2 + \beta_2 \varepsilon_{t-2}^2 + \ldots + \beta_b \varepsilon_{t-b}^2$$

can be rewritten in the following form:

$$\varepsilon_t^2 = K + \beta_1 \varepsilon_{t-1}^2 + \beta_2 \varepsilon_{t-2}^2 + \ldots + \beta_b \varepsilon_{t-b}^2 + u_t$$

where $u_t = \varepsilon_t^2 - E(\varepsilon_t^2|\Omega_{t-1})$ is a mean zero white noise process. It can be noticed that it is AR(b) process for squared residuals ε_t^2 which explains the reason for calling the model "Autoregressive". Analogously, the GARCH model can be expressed as ARMA model for squared residuals (for simplicity GARCH(1,1) is considered):

$$\varepsilon_t^2 = K + (\alpha_1 + \beta_1)\varepsilon_{t-1}^2 - \alpha_1 u_{t-1} + u_t$$

where again $u_t = \varepsilon_t^2 - E(\varepsilon_t^2|\Omega_{t-1})$ stands for white noise disturbance term. The above equation represents the ARMA(1,1) model for squared residuals. Due to mentioned relation with autoregressive processes, GARCH models inherit many properties from the theory of ARMA models. Some of them are : volatility clustering, fat tails and mean reversion. Considering the simplest GARCH(1,1) model

$$\sigma_t^2 = K + \alpha_1 \sigma_{t-1}^2 + \beta_1 \varepsilon_{t-1}^2$$

we can notice that a large ε_{t-1}^2 will result in a large σ_t^2. In other words, the large ε_{t-1}^2 will tend to be followed by another large ε_t^2 creating the behavior known as volatility clustering. Moreover, the distribution of errors modeled using GARCH is fat tailed even if the assumed conditional distribution is normal, a formal explaination of this fact can be found e.g. in [7], [17],[4]. Moreover, assuming that $\sum_{i=1}^{a} \alpha_i + \sum_{j=1}^{b} \beta_j < 1$ it may be shown that $Var(\varepsilon_t) = K/(1 - \sum_{i=1}^{a} \alpha_i + \sum_{j=1}^{b} \beta_j)$ which is the formula for computing unconditional variance of errors. If this assumption holds, then the volatility of the underlying time series will be always pulled towards the long run level expressed by the above formula even if a time series experiences large volatility (see [9] for more details).

Modeling simultaneously conditional mean and variance using seemingly different models ARMA GARCH, is actually working with two ARMA models: one which is applied to the levels of wind speed and second, to the squared residuals. Generally speaking, the building an ARMA GARCH model is based on removing any serial correlation in the data using ARMA and then finding appropriate GARCH model for the residuals if a conditional heteroscedasticity has been detected (following sections will describe it in detail).

When a model for wind speeds has been built and satisfies most of statistical criteria of being adequate, its estimated coefficients and structure need some interpretation in terms of the wind physics. Interpreting such a model is, however, quite difficult especially due to the fact that wind is a part of large atmospheric mechanism which includes many phenomena interacting with each other (e.g. temperature, pressure) which are not considered in model building.

Whereas the Autoregressive moving average model provides quite intuitive description of the serial correlation in wind speeds, the conditional heteroscedasticity modeled by GARCH seems to be quite peculiar. Ewing *et al* [6] apply a similar methodology to modeling 15 minutes averages of wind speed and propose interpreting conditional variance σ_t^2 as turbulence. Moreover, they think of shocks to wind speed ε_t as of wind gusts or lulls corresponding to positive and negative errors respectively. Therefore according to GARCH model, turbulence is influenced by past wind shocks and past turbulence levels. The importance of turbulence issue is crucial from the standpoint of generated power since turbulence decreases the possibility of using the energy in the wind effectively for a wind turbine. It also imposes more wear and tear on the wind turbine.

The usefulness of such interpretation is less reasonable for our purposes since we deal with hourly values of wind speed and concerning turbulence then is questionable. However, we will sometimes refer to the proposed interpretation in order to gain more intuition and understanding of the process.

3.2 Stationarity

Stationarity has been always an indispensable issue in a time series modeling because stationary time series can be well described in terms of its first and second order moments. Formally, we call a process $\{Y_t\}$ covariance

stationary [2] if

$$E(Y_t) = \mu \tag{3.3}$$

$$cov(Y_t, Y_{t-j}) = E((Y_t - \mu)(Y_{t-j} - \mu)) = \gamma_j$$

We will call the covariance stationary time series simply a stationary time series. We can notice that the first and second moments are time invariant. The parameter γ_j is called the jth order or lag j autocovariance of $\{Y_t\}$. The autocorrelations of $\{Y_t\}$ are defined by :

$$\rho_j = \frac{cov(Y_t, Y_{t-j})}{\sqrt{var(Y_t)var(Y_{t-j})}} = \frac{\gamma_j}{\gamma_0} \tag{3.4}$$

and a plot of ρ_j against j is called the autocorrelation function (ACF). If a time series is stationary, the estimate of ACF may suggest which of the many possible stationary time series models is a suitable candidate for representing the dependence in the data. The sample autocovariance and lag j sample autocorrelation are defined as follows:

$$\hat{\gamma_j} = \frac{1}{T} \sum_{t=j+1}^{T} (Y_t - \bar{Y})(Y_{t-j} - \bar{Y}) \tag{3.5}$$

$$\hat{\rho}_j = \frac{\hat{\gamma}_j}{\hat{\gamma}_0} \tag{3.6}$$

where $\bar{Y}$ is the sample mean and T is the number of data values. The sample ACF is a plot of $\hat{\rho}_j$ against j.

In the case of ARMA model described in the previous section,

$$Y_t - \mu = \Phi(B)(Y_t - \mu) + \Theta(B)\varepsilon_t$$

where $\Phi(B) = 1 - \phi_1 B - \ldots - \phi_p B^p$, the stationarity is satisfied if the roots of the characteristic equation

$$\Phi(z) = 1 - \phi_1 z - \phi_2 z^2 - \ldots - \phi_p z^p = 0 \tag{3.7}$$

lie outside the complex unit circle (have modulus greater than one). A necessary condition for stationarity that is useful in practice is that $|\phi_1 + \ldots + \phi_p| < 1$.

[2] it is also called second order stationary or weakly stationary.

In the case that there is one root lying on a complex unit circle then we deal with a process called an integrated process of order 1 and denote it I(1). I(1) process has the following form:

$$Y_t = Y_{t-1} + u_t$$

where u_t is a stationary time series. The first difference of Y_t leads to stationary time series i.e.

$$\Delta Y_t = u_t$$

Because of the above property I(1) processes are called difference stationary. The stationary process is sometimes denoted as I(0).

Sometimes it may be hard to distinguish between the stationary and difference stationary processes especially if we are given time series covering short period of time. Therefore several tests were developed in order to deal with this issue. The tests which consider the null hypothesis of the existence of unit root are called unit root tests i.e if we consider a simple AR(1) model

$$Y_t = \phi Y_{t-1} + \varepsilon_t$$

the hypothesis is:

$$H_0 : \phi = 1 (\text{unit root in } \Phi(z) = 0) \Rightarrow Y_t \sim I(1)$$

$$H_1 : |\phi| < 1 \Rightarrow Y_t \sim I(0)$$

The most famous test are the augmented Dickey-Fuller test and Phillips-Perron test. However, they posses a drawback that they cannot distinguish highly persistent stationary processes from nonstationary processes very well. Therefore the so-called efficient unit root test were proposed and include: point optimal test, DF-GLS and modified efficient Phillips-Perron tests. In [9] more information can be found together with references.

According to the literature, if the sample autocorrelation function decays very slowly it may be an indication of nonstationarity. The usual advice is taking the first difference and then fitting a time series model. However, if the time series is highly persistent but not nonstationary then it may lead to very bad results e.g. simulated sample from the model differs significantly from the original time series. Therefore it is of importance to verify, using efficient unit root tests, if the considered time series come from a nonstationary process.

3.3 Model selection

According to Box and Jenkins [3] the process of building model has an iterative structure. After postulating general class of models we identify model to be tentatively entertained, we estimate its parameters and we check if the model is adequate. If the model turns out to be inadequate we go back to the step of identification. When the model satisfy all requirements stated in diagnostic checking then we can use it for simulating or forcasting purposes. This section aims to present the most important tools used in each iteration stage and the next one serves their application to wind speed data.

3.3.1 Transformation

Before going through all iterative stages, the data have to be prepared. The preparation involves transformations that adjust for undesirable features in dataset like nongaussian distribution, nonstationarity or seasonality. Usually, we assume in ARMA model representation 3.1 that innovations come from normal distribution and this implies that modeled data should be at least approximately Gaussian. Since we know that wind speed does not exhibit required normality but is rather Weibull distributed we have to apply suitable methods for dealing with it. Some ideas have already appeared in literature including e.g. Brown et al. [14] who proposes power transformation choosing the power using the fact that Weibull distribution with shape parameter close to 3.6 resembles Normal distribution. However, instead of trying to get approximately Gaussian distribution, we can just transform our data to uniformity using their CDF and next transforming them to standard normal distribution. Beside the latter approach seems to be mathematically more correct, it is also more general and can be used for data arbitrarily distributed. On the other hand, it may be memory consuming in the case of using empirical CDFs which have to be saved for back-transformation purpose.

Another important issue in data preparation is checking for seasonality. Very often data exhibit some regular patterns and the one of approaches for dealing with it is differencing the data which is taking difference between consecutive observations or separated by a certain period. The differenced data is then easier to model. We propose here another method for removing periodicity which is based on transforming each data point using its respective CDF. More details are presented further in this section.

The last undesired feature to check is nonstationarity. In general, if

there is no trend in data and the autocorrelation function decays quickly and time series exhibit mean-reverting character then we have nothing to worry about. If, however, some of the mentioned properties do not hold then suitable tests (e.g. unit root tests) have to be performed in order to gain more evidence regarding nonstationarity. In the case of wind speed we have already presented its diurnal seasonality concerning inland stations. In order to deal with both the nongaussianity and diurnal periodicity in wind speed, instead of using two separate transformations we use one concise method summarized below.

1. determination of CDF for each hour of the day
2. transforming each data value through its hourly CDF to uniformity
3. transforming the uniformly distributed time series to normality

In order to determine CDF (1) we can obviously try to fit parametric distributions for every 24 hours but more simple approach is to derive empirical CDF or estimate cumulative distribution function using kernel estimator. The latter method was applied in wind data transformation.

3.3.2 Identification

The purpose of model identification is specifying certain tentative models which are worth careful investigation. The methods for model specification can be divided into two groups. The first group focus on analyzing a several sorts of correlations by mainly visual inspection, whereas the second one is based on applying the criterion functions. The latter methods require, however, full estimation of a considered model and preferable models are those which minimize criterion function. There is no hierarchy of methods e.i. no methods are better than others but all are intended to indicate reasonable candidates for models.

Correlation methods

- **SACF**

  The purpose of analyzing patterns in sample autocorrelation function (SACF) is finding resemblance to known autocorrelation function of common ARMA models. Moreover, SACF can be also used as indicator of nonstationarity when it is persistent ($\hat{\rho}_k$ decreases linearly

instead of exponentially). According to theory, for stationary ARMA models

$$\hat{\rho}_k \rightarrow_p \rho_k, \qquad \text{as } n \rightarrow \infty$$

where $\rightarrow_p$ denotes convergence in probability and n the time series length. In addition, $\hat{\rho}_k$ is asymptotically normal with mean ρ_k and variance being function of the ρ_i's. In the case of white noise one obtains the following

$$var(\hat{\rho}_k) \approx \frac{1}{n} \qquad \text{for all } k > 0$$

The MA(q) process is characterized by the autocorrelation function which cuts off after lag q e.i $\rho_k \neq 0$ for $k \leq q$ and $\rho_k = 0$ for $k > q$. Therefore if the sample ACF $\hat{\rho}_h$ is significantly different from zero for $0 \leq h \leq q$ and negligible for $h > q$ then as a tentative model MA(q) can be considered. It is of course important what "negligible" means in that context. If we deal with MA(q) process, the asymptotic variance of $\hat{\rho}_k$ for $k > q$ is

$$var(\hat{\rho}_k) = \frac{1 + 2(\rho_1^2 + \cdots + \rho_q^2)}{n}$$

Hence, it is advisable to check if $\hat{\rho}_k$ falls between the bounds $\pm 1.96\sqrt{var(\hat{\rho}_k)}$.

- **SPACF**

  Since for MA(q) models the autocorrelation function is zero for lags beyond q, the sample autocorrelation is a good indicator of the order of the process. However, the autocorrelation of an AR(p) process do not become zero after a certain number of lags- they die off rather than cut off. So a different function is needed to help to determine the order of autoregressive model. For computing the AR coefficient, the Yule-Walker equations can be used. It is obvious that coefficients of order higher that p are zero. Hence, one can identify the order of the AR process having its coefficients estimates. The p-th order Yule-Walker equation is

$$\begin{bmatrix} \rho_1 \\ \rho_2 \\ \vdots \\ \rho_p \end{bmatrix} = \begin{bmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{p-2} & \rho_{p-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{p-3} & \rho_{p-2} \\ \vdots & & & & & \vdots \\ \rho_{p-1} & \rho_{p-2} & \rho_{p-3} & \cdots & \rho_1 & 1 \end{bmatrix} \begin{bmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_p \end{bmatrix}$$

If we replace autocorrelations ρ_i by their sample counterparts $\hat{\rho}_i$ then estimate $\hat{\phi_{p,p}}$ of ϕ_p in above matrix equation is called the partial autocorrelation at lag p. In the light of previous remark, for an AR(p) process the following holds

$$\hat{\phi}_{p,p} \neq 0, \text{ but } \hat{\phi}_{k,k} = 0 \text{ for } k > p$$

Foregoing relations formulate the cutting off property for AR(p) model. For an observed time series, one needs to estimate the partial autocorrelation function at several lags. According to Yule-Walker equation, one has only to estimate autocorrelation function $\hat{\rho}_k$ for $k = 1, 2, \ldots$ and solve linear equations. It was shown that, under the hypothesis that an AR(p) model is correct, the SPACF at lags greater than p are approximately normally distributed with zero means and variances $\frac{1}{n}$. Thus, for $k > p$, $\pm 1.96/\sqrt{n}$ can be used as critical limits on $\hat{\phi}_{k,k}$ to test the null hypothesis that an AR(p) model is correct.

- **SEACF**

  Autocorrelation and partial autocorrelation functions are useful in indicating candidates for MA and AR models respectively. However, in the case of specifying mixed autoregressive moving average models they fail. The reason is the fact that both autocorrelation and partial autocorrelation do not posses cutting off property. That is why different tools were developed for ARMA model specification and they include among others EACF-Extended Autocorrelation Function, SCAN- the Smallest Canonical Correlation.

  The EACF was created on the basis of the idea that if the AR part of a mixed ARMA model is known, "filtering out" the autoregression from the observed time series results in a pure MA process that enjoys the cutoff property in its ACF. The AR coefficients may be estimated by a finite sequence of autoregressions. In order to illustrate this idea, the procedure will be discussed on the basis of ARMA(1,1) model:

$$Y_t = \phi Y_{t-1} + \varepsilon_t - \theta \varepsilon_{t-1}$$

  In the first step, the simple linear regression of Y_t on Y_{t-1} is performed. Although, the estimated coefficient ϕ is not correct, the obtained residuals contain information about the MA process. Therefore, a second regression of Y_t is performed, this time, on Y_{t-1} and on lag 1 of the residuals from the first regression. The estimated coefficient

$\tilde{\phi}$ turns out to be a consistent estimate. Consequently, the process defined by $W_t = Y_t - \tilde{\phi}Y_{t-1}$ is approximately MA(1) process. If the considered model is ARMA(1,2) then the consistent coefficient of Y_{t-1} is obtained by the regression of Y_t on Y_{t-1} the lag 1 of the residuals from second regression and the lag 2 of the residuals from first regression. The same procedure holds for general ARMA(p,q) but requires larger number of regression, namely q. Usually orders of AR and MA are not known and iterative procedure has to be performed. Thus, let us define autoregressive residuals using AR coefficients estimated iteratively assuming that AR order is k and MA order is j:

$$W_{t,k,j} = Y_t - \tilde{\phi}_1 Y_{t-1} - \cdots - \tilde{\phi}_k Y_{t-k}$$

The sample autocorrelations of $W_{t,k,j}$ are called the extended sample

AR/MA	0	1	2	3	4	5	6	7	8	9	10	11	12	13
0	x	x	x	x	x	x	x	x	x	x	x	x	x	x
1	x	0*	0	0	0	0	0	0	0	0	0	0	0	0
2	x	x	0	0	0	0	0	0	0	0	0	0	0	0
3	x	x	x	0	0	0	0	0	0	0	0	0	0	0
4	x	x	x	x	0	0	0	0	0	0	0	0	0	0
5	x	x	x	x	x	0	0	0	0	0	0	0	0	0
6	x	x	x	x	x	x	0	0	0	0	0	0	0	0
7	x	x	x	x	x	x	x	0	0	0	0	0	0	0

Figure 3.2: Theoretical eacf of ARMA(1,1) model

autocorrelations. If the true model is ARMA(p,g) then for $k = p$ and $j \geq q$ the process $\{W_{t,k,j}\}$ is approximately an MA(q) model. The information in the sample EACF can be presented in the table such as in Figure 3.2, where the element in the kth row and jth column is X if the lag $j + 1$ sample autocorrelation of $\{W_{t,k,j}\}$ is significantly different from 0 indicating that the model ARMA(k,j) is not worth further consideration and the element is 0 otherwise. Figure 3.2 presents the theoretical EACF ; the triangle of zeros can be noticed where 0^* which corresponds to ARMA(1,1) and indicates the best and most parsimonious candidate. When one deals with real data, this kind of triangle is hardly visible because many of correlations may be statistically not significant by chance.

Information criteria

An information criteria is a measure of the goodness of fit of an estimated statistical model but is not a test in the sense of hypothesis testing, rather it is a tool for model selection. Given a data set, several competing models may be ranked according to their value of the information function, with the one having the lowest value being the best. From these values one may infer that e.g the top three models are in a tie and the rest are far worse, but one should not assign a value above which a given model is "rejected". All formulas for information criteria contain a "penalty term" which is the function of the number of parameters in a model and helps to ensure selection of parsimonious models and to avoid choosing models with too many parameters. There are several information criteria available in literature and some of them are presented below.

Let us denote $L_n(k)$ the maximum likelihood of a model with k parameters based on a sample of size n.

- **AIC**

  The Akaike's information criteria is defined as

  $$AIC(model) = -2ln(L_n(k))/n + 2k/n \tag{3.8}$$

  If the considered model is ARMA(p,q) with Gaussian errors i.e.

  $$Y_t - \mu = \sum_{i=1}^{p}(Y_{t-i} - \mu) + \sum_{j=1}^{q}\varepsilon_{t-j} + \varepsilon_t,\ \varepsilon_t \sim i.i.d.N(0, \sigma^2) \tag{3.9}$$

  then the formula 3.8 can be rewritten in the following form [3]

  $$AIC(ARMA(p,q)) = ln(\hat{\sigma}^2) + 2(1+p+q)/n$$

  where $\hat{\sigma}^2$ is the maximum likelihood estimator of the variance of the innovation noises.

  Using this criteria we select the model which minimizes the AIC. The AIC is also an estimator of the average Kullback-Leibler divergence of the estimated model from the true model.

- **BIC**

[3] We use the fact that for ARMA model with normally distributed errors we have that $ln(L_n(k)) = -\frac{1}{2}n - \frac{1}{2}nln(2\pi) - \frac{1}{2}nln(\hat{\sigma}^2)$

Another method for determining the model's orders is choosing model that minimizes the Schwarz Bayesian Information Criteria (BIC) defined as

$$BIC(model) = -2ln(L_n(k))/n + kln(n)/n \tag{3.10}$$

Analogously to AIC, in the case of ARMA model (3.9), the formula 3.10 takes the following form:

$$BIC(ARMA(p,q)) = ln(\hat{\sigma}^2) + (1+p+q)ln(n)/n$$

Monte Carlo studies regarding fitting autoregressive models has shown that AIC has tendency to overestimate p. The BIC criterion was created to correct the overfitting nature of AIC. It has a valuable property of consistence in the sense that if the data are in fact observation from ARMA(p,q) model and $\hat{p}$ and $\hat{q}$ are estimated orders found by minimizing BIC, then $\hat{p} \to p$ and $\hat{q} \to q$ as $n \to \infty$. This property is not possessed by AIC statistics.

Apart from order determination, an equally important issue is finding the subset of nonzero coefficients of an ARMA model especially in the case of high order ARMA models. For example, the model $Y_t = 0.7Y_{t-10} + \varepsilon_t + 0.8\varepsilon_{t-10}$ is a subset of ARMA(10,10) model. For this reason the method of Hannan and Rissanen for estimating the ARMA orders was extended to solving the problem of finding an optimal subset ARMA model. Without going into theoretical details, one can examine a few best subset ARMA models in terms of chosen criterion function and obtain information about tentative models for further study. The algorithm of this method is implemented in R in package TSA. The Figure 3.8b shows the output which is a table with pattern indicating which lag of the observed time series and which of the error process enter into best subset models. The models are sorted according to their criterion function value, with better models placed in higher rows and with darker shades.

3.3.3 Estimation

After the orders p and q are specified, one still needs the estimation of models parameters. One of the methods for estimation coefficient is maximum likelihood estimation described below.

Maximum likelihood estimator

The likelihood function L for given time series $Y_1, Y_2, \ldots, Y_n$ is defined to be the joint probability density of obtaining the data actually observed.

However, it is considered as a function of the unknown parameters in the model with the observed data held fixed. For ARMA models L is a function of $\phi, \theta, \mu, \sigma^2$ given the observations $Y_1, Y_2, \ldots, Y_n$. The maximum likelihood estimators are then defined as those values of the parameters for which the data actually observed are most likely, that is, the values that maximize likelihood function. The unconditional log-likelihood is given by

$$l(\phi, \theta, \sigma) = f(\phi, \theta) - nln(\sigma) - \frac{S(\phi, \theta)}{2\sigma^2}$$

where $S(\phi, \theta)$ denotes unconditional sum of squares function.

3.3.4 Diagnostic checking

After a model has been identified and estimated, there is still a question whether the model is adequate. Several methods for testing model adequacy are available and are discussed in this section. These tools have to be sensitive enough to capture all significant deviations from adequacy. Unfortunately no such comprehensive methods exist and only "part" of non-suitability may be indicated. However, no model form ever represents the truth absolutely. Therefore it is normal that the well fitted models will always enjoy some discrepancies. In general, the model is regarded as a model describing the data adequately if when applied to long data does not show serious discrepancies.

Residuals may be defined in a very general way as

$$residual = actual - predicted$$

If the considered model is well specified and estimated so that its coefficients are reasonably close to the true values, then the residuals should behave in a similar manner to white noise. Namely, they are expected to be serially uncorrelated and identically normally distributed [4]. In the case that at least one of the latter properties does not hold for estimated residuals, then the model may be not a right one. Moreover, deviations from these properties can sometimes suggest the way of finding the more appropriate model. Suppose the following model has been fitted with maximum likelihood estimates $(\hat{\phi}, \hat{\theta})$

$$\phi(B)Y_t = \theta(B)\varepsilon_t$$

[4] Different distributions are also allowed. It depends on the assumptions made before estimation. Different likelihoods functions will be built for different distributions of innovations

then the quantities $\hat{\varepsilon}_t = \hat{\theta}^{-1}(B)\hat{\phi}(B)Y_t$ are called residuals. It can be shown that the following relation holds:

$$\hat{\varepsilon}_t = \varepsilon_t + O(\frac{1}{\sqrt{n}})$$

where n is sample size. Hence, as the series length increases, the $\hat{\varepsilon}_t$'s becomes close to the white noise ε_t's. Let us now consider the AR(2) model $Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \varepsilon_t$. In the estimation procedure one obtains estimated $\hat{\phi_1}, \hat{\phi_2}$. The residuals in this case are defined as:

$$\hat{\varepsilon}_t = Y_t - \hat{\phi_1} Y_{t-1} - \hat{\phi_2} Y_{t-2}$$

- **Residual graph**

  The first step in Residual analysis is a visual inspection. As it was already mentioned, residuals have to resemble the white noise process. Therefore one may check if there is no trends, cyclic components, strong deviations from zero or nonconstancy of the variance. Very often standardization of residuals is applied by dividing their values by the estimate of white noise standard deviation $\hat{\sigma}$. Then it is easier to notice unusual size of residuals. The visual inspection of the residuals plot aims to inform one about probable shortcoming of the model and has to be followed by a few next steps of diagnostic checking. Figures presenting residuals are given in the next section.

- **Residual ACF**

  In order to verify if the residuals are serially uncorrelated, the sample autocorrelation function is plotted. Analogously to white noise process, the sample ACF of resiudals $\hat{\rho}_k$ should have mean zero and standard deviation equal to $n^{-1/2}$. However, it was shown (see [3] for references) that $n^{-1/2}$ constitute an upper bound for standard error. The residuals for small lags can be substantially less than $1/n$ and highly correlated. The latter effect gradually disappear over time. That is why it is important to examine more carefully the sample ACF at starting lags. As an example of these results, let us consider the case of AR(1) which is assumed to fit well a large set of data. Then one has:

$$Var(\hat{\rho}_1) \approx \frac{\phi^2}{n} \text{ for } k = 1$$

$$Var(\hat{\rho}_k) \approx \frac{1 - (1 - \phi^2)\phi^{k-2}}{n}, \text{ for } k > 1$$

- **Tests**

  Instead of examining values of $\hat{\rho}_k$ individually, one may consider if a group of them does not indicate inadequacy of the model. On the basis of this idea the so called Portmanteau lack of fit test was created. Originally it was developed by Box and Pierce and its statistic has the following form:

  $$Q = n(\hat{\rho}_1^2 + \hat{\rho}_s^2 + \cdots + \hat{\rho}_k^2)$$

  where k is the number of estimated autocorrelations taken into account (from ARMA(p,q) model) and n is data size. The statistic is approximately distributed as $\chi^2(k-p-q)$ and if the fitted model is inappropriate then the value of the statistic tends to inflate. Later it was discovered by Ljung and Box that the chi-squared distribution is not good enough estimate for the distribution of Q and consequently they proposed a modified form of statistic which does not posses the mentioned drawback. The improved statistic is given below:

  $$Q_* = n(n+2)(\frac{\hat{\rho}_1^2}{n-1} + \frac{\hat{\rho}_2^2}{n-2} + \cdots + \frac{\hat{\rho}_k^2}{n-k})$$

- **Normality of the residuals**

  In order to check if the residuals come from normal distribution, one may plot the so called QQ plot which a plot of the quantiles of the data set against (in this case) the quantiles of normal distribution. If the points follow closely the straight line than it is an indicator of normality of residuals. Apart from QQ graphs, one can also perform normality tests like the Shapiro-Wilk or Jarque-Bera test.

3.3.5 Checking heteroskedasticity

So far we have considered the matters regarding fitting an adequate ARMA model to the time series. The residuals obtained from fitting an ARMA model are supposed to be uncorrelated and distributed according to the distribution superimposed in maximum likelihood estimation. It may happen, however, that residuals are not characterized by constant conditional variance and exhibit e.g. alternating periods of low and high variability. Then, it is indication of the so-called ARCH effect or equivalently heteroscedasticity. Apart from the visual inspection of the graph of residuals there are more formal techniques for checking the presence of ARCH effect and are described in this section. According to the Engle's strategy, when the conditional

variance is not constant, it is possible to model the conditional variance using ARCH(q) (GARCH(0,b) presented in the section 3.1) process using the square of the estimated residuals obtained from the application of ARMA model to the time series at hand. Let us consider ε_t^2 modeled using AR(q) process as shown below:

$$\varepsilon_t^2 = K + \alpha_1 \varepsilon_{t-1}^2 + \alpha_2 \varepsilon_{t-2}^2 + \ldots + \alpha_q \varepsilon_{t-q}^2 + \nu_t \tag{3.11}$$

We can notice that for a given sample, ε_t^2 is an unbiased estimate of σ_t^2. Therefore, we can examine ε_t^2 to check for conditional heteroscedasticity. The following tools are supposed to help us detecting the ARCH effect:

Autocorrelation plot

Calculate and plot the sample autocorrelation of the squared residuals. If the autocorrelation is significantly high at some lags it indicates that ARCH effect appears in residuals.

Tests

Tests the hypothesis: H_0: No ARCH effect H_1: ARCH effect present

- Perform the Ljung Box test on the squared residuals ε_t^2 taking into consideration several numbers of correlation coefficients. If the null hypothesis is rejected then we reject the hypothesis of no ARCH errors.
- The next test is the Lagrange multiplier test applied to the linear regression 3.11. The usual R^2 statistic is computed and assuming that we deal with a sample of T residuals, the test statistic $LM = TR^2$ converges to $\chi^2(q)$ distribution. If LM is sufficiently large, rejection of the null hypothesis that $\alpha_1 \ldots \alpha_q$ are jointly equal to zero is equivalent to rejecting the null hypothesis of no ARCH errors. On the other hand, if LM is sufficiently low, it is possible to conclude that there are no ARCH effects.

3.3.6 Backtransformation

As soon as the model is found it can be used for simulation purpose. Since the model was based on the transformed data, the simulated sequence of values have to be transformed back to wind speed domain. It is possible by reversing the transformation process described in section 3.3 in the following way:

1. transforming simulated time series to uniformity

2. transforming each data value of the uniform time series through the inverse of the hourly CDF to the wind speed domain

3.4 Example

After discussing all needed tools in finding a good model it is time to show how they work in practice. We will present the Box-Jenkins iterative procedure applied to wind speed time series recorded at Schiphol station in July 2004. The considered wind speeds are shown in Figure 3.3.

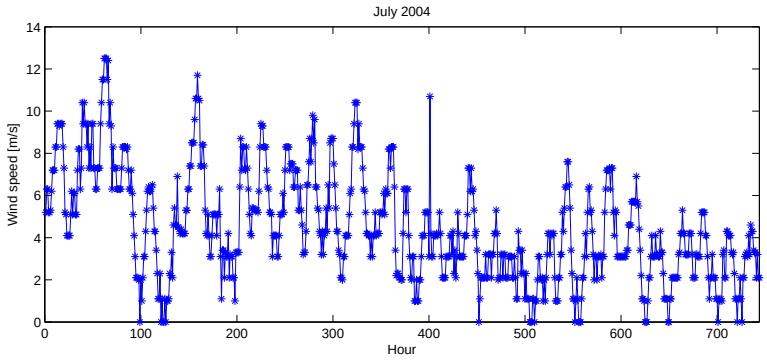


Figure 3.3: Wind speed in July 2004

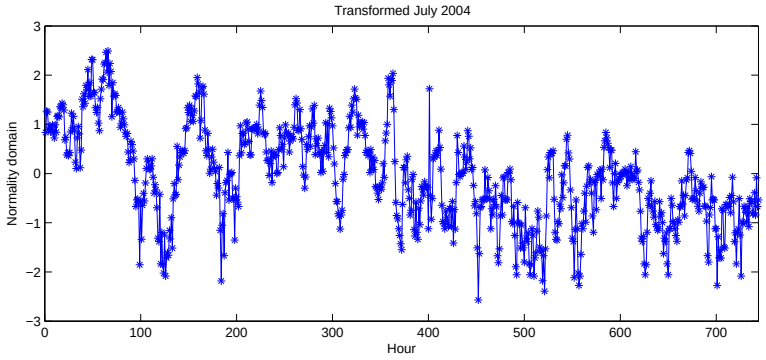


Figure 3.4: Time series of transformed July observations

3.4.1 Transformation

In order to transform July data we need to compute CDF for each hour of the day i.e. for each random variable $H_1, H_2, \ldots, H_{24}$. If we assume that the Julys from years $2003 - 2007$ are realizations of the same process [5] then we have access to more realization of variables $H_1, H_2, \ldots, H_{24}$ than considering July 2004 alone. Thus, we transform data from July 2004 using all available values for $H_1, H_2, \ldots, H_{24}$ in all Julys $2003 - 2007$. The Figure 3.5 shows kernel CDF estimator for several hours.

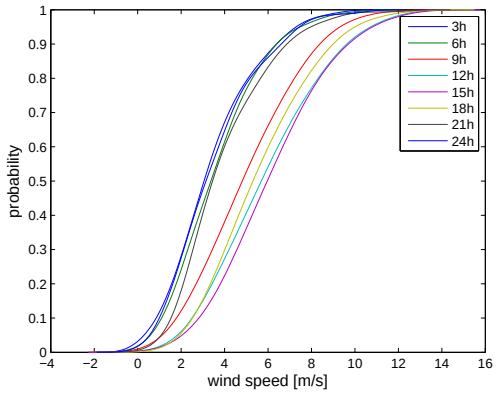


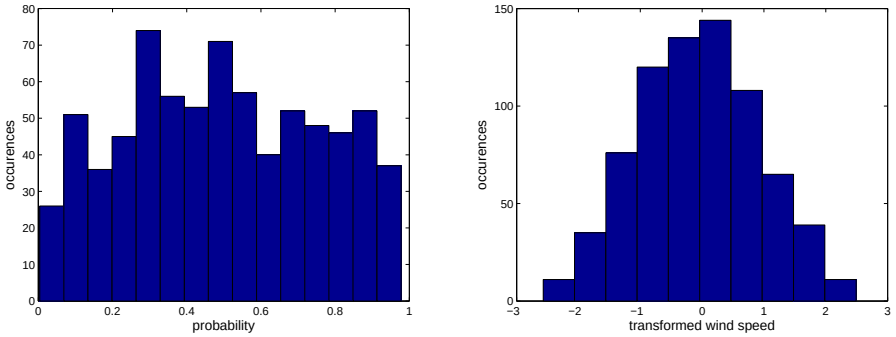
Figure 3.5: Hourly CDF

After transformation of data using hourly CDFs we deal with observations uniformly distributed (see Figure 3.6a) which are further transformed to normality using inverse standard normal distribution. The Figure 3.6b presents all data values transformed using their respective hourly CDF and using inverse standard normal cumulative distribution function.

In order to check if the transormations led us to normally distributed observations we perform four normality tests: Jarque Bera test, Lilliefors, Kolmogorov Smirnov and Shapiro Wilk test. At the 1% level of significance two of these tests judge a departure from normality (Lilliefors and Kolmogorov Smirnov tests), whereas the other two fail to reject the null hypothesis of normality (see Table 3.1). The wind speed time series from July 2004 transformed according to this procedure is presented in Figure 3.4

At this stage of modeling we should consider the stationarity of the time series. Analyzing the plot of transformed wind speeds we can notice that there is no trend. Moreover, the series tend to return over time to zero

[5] see section 3.5.1 for more details



(a) Histogram of data transformed through their hourly CDFs (b) Histogram of transformed July data

Figure 3.6

Normality test	p-value
Jarque Bera	0.0169
Lilliefors	< 0.001
Kolmogorov Smirnov	0.0042
Shapiro Wilk	0.0122

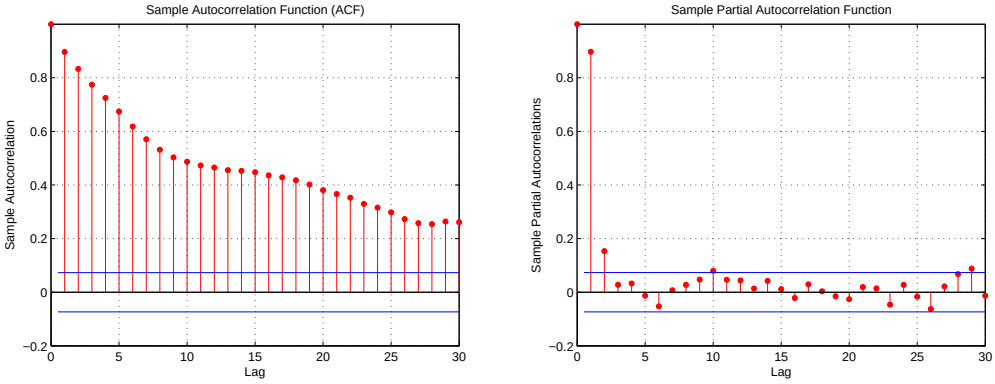
Table 3.1: Normality tests applied to July transformed data

i.e. exhibit mean reversion behaviour. This is not a surprising observation since we know from experience that wind speed behaves in a "stable" way reverting to its usual mean. However, if we are given wind speed time series over very short period of time e.g. one week it can be the case that they look very like random walk attaining zero e.g. only once. Therefore it is important to know the behaviour of the process over long horizons. In order to support our conjecture, we performed unit root tests (see section 3.2) for the considered time series. According to three efficient unit root tests, we reject the null hypothesis of present unit root at 5% level, see Appendix Table 6.1.

3.4.2 Identification of ARMA

At the beginning we focus on finding as good as possible model for the conditional mean. When this is done we try to find a suitable model for the conditional variance if a heteroscedasticity effect occurs. Before we start the identification process, we carry out mean adjustment using the overall sample mean. Therefore no intercept will be estimated.

The Figures 3.7a,3.7b, and 3.8 present the tools intended to help choosing tentative models. The autocorrelation function seems to decay quite steeply but attains high values for quite significant number of lags. According to the theory it could be an indication of nonstationarity. However, according to the observations made in the previous section and unit root tests we do not have any doubts regarding the stationarity of the time series.



(a) Autocorrelation (b) Partial autocorrelation

Figure 3.7: Identification tools

The values of ACF are high for all presented lags indicating that MA model is certainly not appropriate. The PACF shown in Figure 3.7b cuts off after lag 2 hinting AR(2) process. On the other hand the edge of triangle with zeros in Figure 3.8a corresponds to ARMA(2,1) and ARMA(1,1). Best subset ARMA selection supports the latter model. Summarizing, we got the following tentative models to consider in further analysis: AR(2), ARMA(2,1) and ARMA(1,1).

The next step involves the analysis of tentative models and their augmentations with respect to their AIC and BIC. We choose the model which minimizes both AIC and BIC. However, it may happen that one model have higher BIC but smaller AIC than other. Then we use likelihood ratio test to

Tentative model	AIC	BIC
AR(1)	875.86	885.08
AR(2)	860.95	874.78
AR(3)	862.44	880.89
ARMA(1,1)	860.27	874.10
ARMA(2,1)	862.23	880.67
ARMA(1,2)	862.23	880.68
ARMA(3,1)	864.20	887.26

Table 3.2: AIC and BIC of several tentative models

solve this problem. The Table 3.2 presents the values of information criteria for several tentative models. The models were fitted using maximum likelihood estimation and assuming that errors are normally distributed. We can notice that ARMA(1,1) performs best but very similarly to AR(2) model. This happens quite often and both competitive models could be used to model the data [17]. In further analysis we will consider ARMA(1,1). The estimated parameters for ARMA(1,1) are shown in Table 3.3

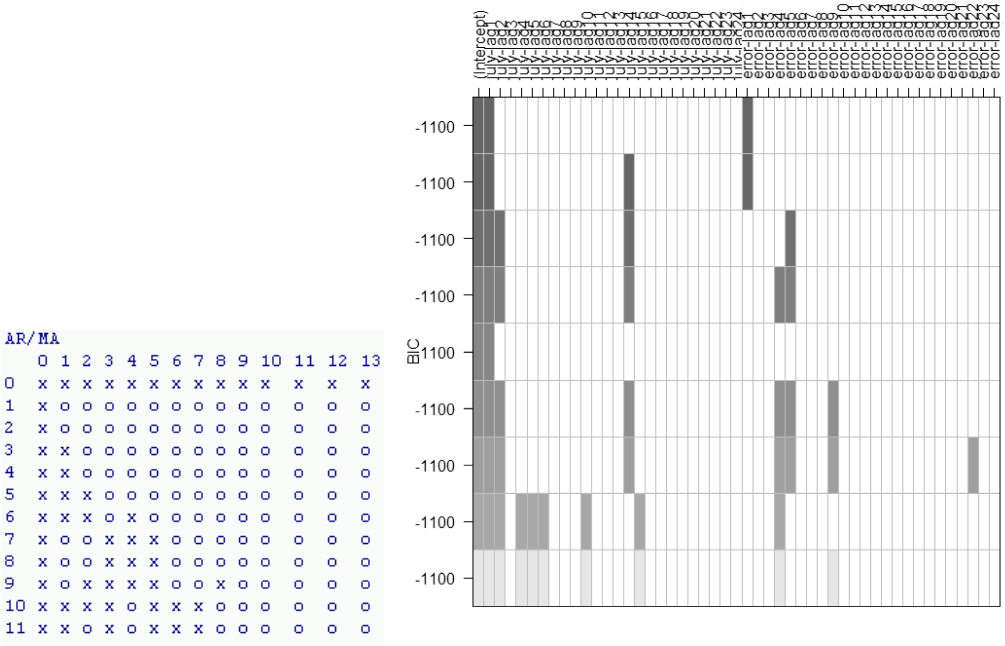
Parameter	Value	Standard Error	T Statistic
AR(1)	0.93022	0.01416	65.6952
MA(1)	-0.17338	0.029596	-5.8583
K	0.18458	0.0070488	26.1864

Table 3.3: Parameters, standard errors and T statistics of ARMA(1,1) model

3.4.3 Diagnostic checking

In order to check the adequacy of the ARMA(1,1) model we have to analyze the residuals. They should posses the desired properties. Namely, they should be serially uncorrelated and they should be normally distributed.

In order to check if the distribution of residuals is Gaussian we perform several normality test. The results are shown in Table 3.4. We can notice that only Kolmogorov Smirnov test does not rejects the null hypothesis. The other tests reject it strongly. In order to find out if the distribution of the residuals possess fatter tails than normal distribution we analyze the normal quantile-quantile plot, see Figure 3.9a. We can notice that the



(a) Extended autocorrelation

(b) Best subset ARMA selection based on BIC

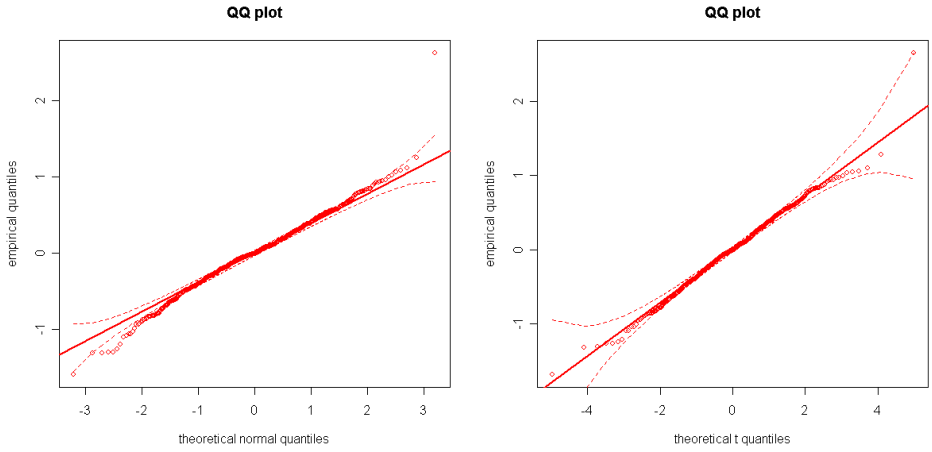
Figure 3.8: Identification tools

distribution of residuals are characterized by fatter tails than the normal distribution since there are strong departures from the straight line and consequently from 95% confidence envelop. Therefore we have also compared the residuals distribution with the Student-t distribution, which has fatter tails than Gaussian, using again qq plot, see Figure 3.9b. In contrast to normal distribution, the Student-t qqplot does not raise any incomptability suspicions because all points lie comfortably in the envelop. The leptokurtic character of the distribution of errors may be an indication that residuals are conditionally heteroscedastic.

The next desired feature of the residuals is serial uncorrelation. We check the presence of this property by inspecting the sample autocorrelation of the residuals and performing Ljung Box test for several numbers of lags. The Figure 3.10 shows the sample autocorrelation function of residuals. We can see that all sample autocorrelation coefficients lie in the confidence bounds computed under the null hypothesis of white noise. Therefore the residuals seem to be serially uncorrelated. The performed Ljung Box gives high

Normality test	p-value
Jarque Bera	< 0.001
Lilliefors	< 0.001
Kolmogorov Smirnov	0.1583
Shapiro Wilk	< 0.001

Table 3.4: Normality tests for residuals with superimposed normal distribution



(a) Normal QQ plot (b) Student-T qq plot

Figure 3.9: Diagnostic tools-qq plots

p-values for all considered numbers of lags, i.e. $5, 10, 15, 20, 25, 30$ giving evidence of residuals' serial uncorrelation, see Table 3.5. We can conclude that ARMA(1,1) describes well the dependence structure in the considered time series. However, the assumption of normally distributed errors is not appropriate. In the next section we will examine whether the squared residuals are correlated. If this is the case it would explain the fatter tailed character of the residuals distribution.

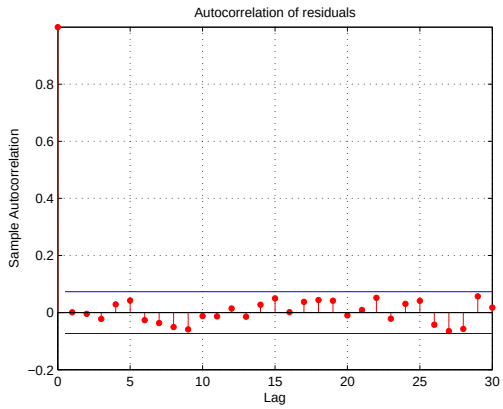


Figure 3.10: The sample autocorrelation of the residuals from ARMA(1,1) model

lags	p-value	Qstat	Critical value
5	0.7977	2.3584	11.0705
10	0.5738	8.5652	18.3070
15	0.7185	11.4721	24.9958
20	0.7507	15.4408	31.4104
25	0.7494	19.9508	37.6525
30	0.4725	29.8673	43.7730

Table 3.5: Ljung Box test performed on residuals

3.4.4 Checking for heteroscedasticity

At the beginning of checking for the conditional heteroscedasticity we examine the plot of residuals. The residuals are shown in Figure 3.11a. It is not straightforward to judge about presence of an ARCH effect on the basis of this figure. The periods of low and high variability are not clearly visible and this suggests that even if errors are heteroscedastic, the coefficients in the variance equation will be not very high. In order to formally check the presence of ARCH effect we plot the autocorrelation of squared residuals-see Figure 3.11b. We see that first lag autocorrelation coefficient is very significant indicating the presence of ARCH effect. This result is in

lags	p-value	Qstat	Critical value
5	0.0000	78.1909	11.0705
10	0.0000	81.2580	18.3070
15	0.0000	83.4714	24.9958
20	0.0000	86.1113	31.4104
25	0.0000	88.6498	37.6525
30	0.0000	100.0150	43.7730

Table 3.6: Ljung Box test for squared residuals

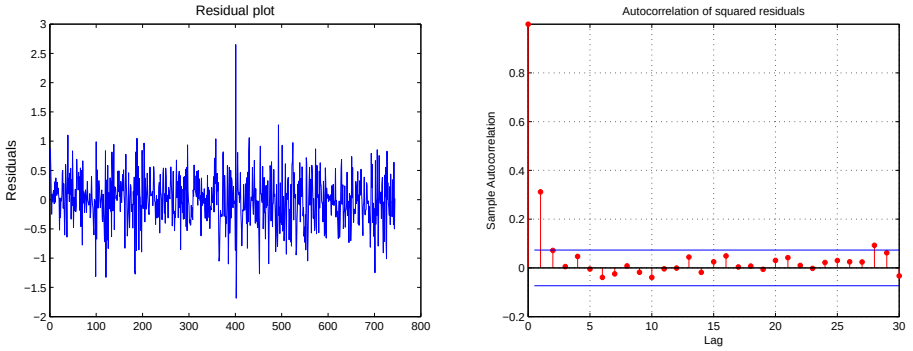
accordance with Ljung Box test performed on squared residuals. The null hypothesis of uncorrelated squared residuals is strongly rejected for all considered number of lags (see Table 3.6). The Engle's hypothesis test also gives evidence for the presence of heteroscedasticity in errors-all p-values are extremely low, see Table 3.7, rejecting the null hypothesis of no ARCH effects. Therefore we can conclude that the residuals from ARMA(1,1) model are conditionally heteroscedastic and consequently an adequate GARCH model have to be chosen.

3.4.5 Identification of GARCH model

Similarly to finding appropriate model for mean, we try to find an adequate model for variance. In order to find out the order of GARCH model appropriate for the residuals from ARMA(1,1) we plot the sample partial autocorrelation function of squared residuals, see Figure 3.12. We can notice that only first lag is significant indicating that a simple ARCH(1) model i.e. GARCH(0,1) would be a good candidate for our data. The GARCH model, apart from order, requires specifying the conditional distribution.

lags	p-value	Qstat	Critical value
1	0	72.5442	3.8415
2	0.0000	72.7723	5.9915
3	0.0000	72.8766	7.8147
4	0.0000	74.7756	9.4877
5	0.0000	75.6929	11.0705
6	0.0000	76.1948	12.5916

Table 3.7: ARCH test for the residuals from ARMA(1,1)



(a) Residuals (b) Autocorrelation of squared residuals

Figure 3.11

In Matlab there are two options for the conditional distribution: normal and Student-T. For the purpose of choosing an appropriate model we compare AIC and BIC computed for GARCH models with different orders and conditional distributions. The Table 3.8 presents the results. We can see that the GARCH models with Student-T distribution superimposed perform much better than their gaussian counterparts. Furthermore, two models obtained the best scores: the GARCH(0,1), which was also suggested by sample partial autocorrelation, and GARCH(1,1). However, the GARCH(1,1) has lower AIC value, whereas the GARCH(0,1) has lower BIC. Therefore we need to use likelihood ratio test in order to decide which model should be entertained. The p-value obtained from the likelihood ratio test is equal to 0.01 indicating that the GARCH(0,1) model is rejected at 0.05 significance level.

The Table 3.9 presents the estimated parameters for ARMA-GARCH-T

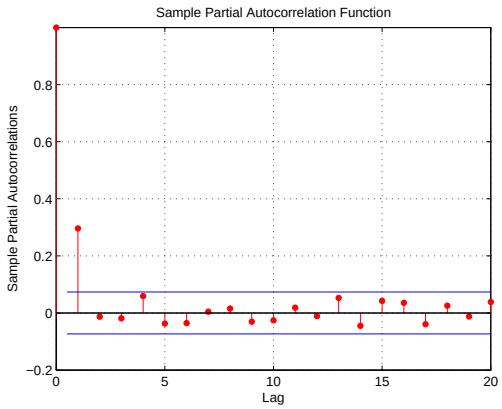


Figure 3.12: The sample partial autocorrelation of the squared residuals from fitted ARMA(1,1) model

model. All estimated coefficients are significant at the standard 0.05 significance level. In order to find out if the ARMA-GARCH model is adequate for our data we need to perform diagnostic checking described in the next section.

Tentative model	AIC-T	BIC-T	AIC	BIC
GARCH(1,1)	800.66	828.34	809.52	832.58
GARCH(0,1)	804.50	827.66	1368.80	1387.30
GARCH(0,2)	801.72	829.4	1034.80	1057.9
GARCH(1,2)	802.65	834.94	811.52	839.19
GARCH(2,1)	802.66	834.95	811.52	839.19

Table 3.8: AIC and BIC of several tentative GARCH models. AIC-T and BIC-T refers to GARCH models with Student-T distribution superimposed

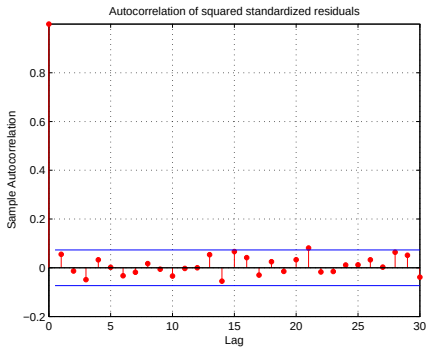
3.4.6 Diagnostic checking

The standardized residuals from the ARMA-GARCH model should be distributed according to the conditional distribution superimposed. Moreover they should be uncorrelated and should not be conditionally heteroscedastic. We have already checked that residuals from ARMA model are uncorrelated. In order to find out if the standardized residuals are not conditionally heteroscedastic we perform Ljung Box and ARCH test on squared standardized

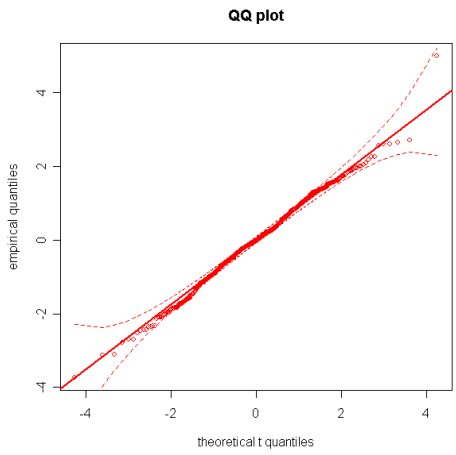
Mean: ARMA(1,1); Variance: GARCH(1,1)-T			
Parameter	Value	Standard Error	T Statistic
AR(1)	0.93418	0.013341	70.0207
MA(1)	-0.13168	0.044335	-2.9702
K	0.069015	0.023952	2.8814
GARCH(1)	0.39992	0.15892	2.5166
ARCH(1)	0.22521	0.060134	3.7451
DoF	11.025	3.9278	2.8069
AIC: 800.6639 BIC: 828.3361			

Table 3.9: ARMA(1,1)-GARCH(1,1)-T estimated parameters

residuals, see Table 3.10. We can see that for all considered lags the p-values are high and the null hypothesis is not rejected. Therefore we may conclude that the squared standardized residuals are uncorrelated and consequently the standardized residuals are not conditionally heteroscedastic. The sample autocorrelation plot supports the latter statement, see Figure 3.13a. Moreover, according to Student-T qq plot (Figure 3.13b) the distribution of the standardized residuals is well captured by the Student -T distribution since there are no departures from 95% envelope.



(a) Squared standardized residuals



(b) Student-T QQ plot for standardized residuals

Figure 3.13

Ljung Box test		ARCH test	
lags	p-value	lag	p-value
5	0.4133	1	0.1310
10	0.7059	2	0.2901
15	0.4476	3	0.2484
20	0.5512	4	0.2751
25	0.5074	5	0.3937
30	0.3954	6	0.4165

Table 3.10: Ljung Box and ARCH tests applied to squared standardized residuals

3.5 The analysis of simulations

3.5.1 Averaged model

Under the assumption that each month of July is a realization of an unknown model, fitting a model to each given time series of July's measurements gives us more information about how the general model may look like. On the other hand we could proceed with several wind speed time series obtained during certain month like with one long time series by connecting the observations from the end of one month with the observations from the beginning of the same month but a year later. However, the standard statistical tools for time series modeling cannot be applied to such time series. Consequently, some counterparts to these tools would have to be developed and implemented. For example, an adequate sample autocorrelation function would not take into account the products of observations from two different months. This method necessitates a complex implementation which gets even more complicated when multivariate modeling is considered. Therefore, due to time constrains, it was not possible to follow this approach.

If the method of fitting a model to each given month is followed, one obtains a number of possibly different models. It would be desired to find a method to aggregate all obtained models in the way that the resulted model captures all important characteristics of wind speed in a specified month. Thus, such a model is expected to generate different types of a month- month with low, high, moderate wind speed.

Having at hand the sequence of observations from July 2003-2007 we fit the model to each year following the Box-Jenkins procedure. The idea is to aggregate all these models by building an "averaged model" by averaging the respective parameters in the obtained models. The Table 3.11-3.12 presents the parameters of all models fitted to July and January 2003-2007 together with the coefficients of the averaged model.

The models fitted differ from each other by the order and the magnitude of parameters. However, they have a common form: the first autoregressive lag coefficient ϕ_1 is always very significant and high and quite often appears alone i.e. AR order is 1. Surprisingly, in many cases after lag 1 there is a "gap" of parameter values fixed to zero and is followed by a significant coefficient which may be an indication of a sort of seasonality in data. The latter fact stands in contradiction with our intuition. We would not expect that present value of wind speed Y_t would depend to higher degree on Y_{t-k} than on Y_{t-k+1} where $k < 24$. Therefore it is not straightforward to interpret the suggested periodicity in terms of wind physics. It may be caused, however,

not directly by some unknown meteorological aspects of wind but rather by errors in measurements. Although we cannot avoid parameters at high lags due to their significance, we can notice that their value is considerably smaller than the parameter ϕ_1. Hence, the seasonality in the transformed data is very weak.

AR	MA	K	GARCH	ARCH	DoF
2003: ARMA(1;1)-GARCH(1;1)-T					
0.95	-0.07	0.05	0.43	0.2	9.83
2004: ARMA(1;1)-GARCH(1;1)-T					
0.93	-0.13	0.07	0.4	0.22	11.02
2005: ARMA(1,2,18;only 14)-GARCH(1;1)-T					
0.81;0.09;0.05	0.08	0.05	0.54	0.14	6.85
2006: ARMA(1,2;only 5 & 6)-GARCH(1;1)-T					
0.76;0.11	0.09;0.08	0.11	0	0.19	11.31
2007:ARMA(1;1)-T					
0.93	-0.06	0.17	0	0	4.21
the averaged model:ARMA(1,2,18;1,5,6,14)-GARCH(1;1)-T					
0.87;0.04;0.01	-0.05;0.02;0.01;0.02	0.09	0.27	0.15	8.64

Table 3.11: The estimated parameters of models fitted to July's time series. The averaged model.

It is worth noticing that the autoregressive parameters sum to a value which is close to one e.g. 0.95. The effective unit root tests have rejected the null hypothesis of nonstationarity (see Table 6.1 in the Appendix) pointing out a highly persistent character of considered processes. The moving average part of the process is also frequently seasonal and besides, its parameters are very small.

Comparing respective ARMA parameters estimated for July and January we do not find any striking differences in magnitude. On the other hand, the GARCH model gives us higher parameters in the case of January than in the case of July. Furthermore, the heteroscedasticity in errors is even sometimes not apparent in some years of July.

Section 3.1 mentioned the interpretation of the conditional volatility as the wind turbulence and shocks to the mean as wind gusts or lulls. Thus, the GARCH model describes the relation between the current turbulence level, past turbulence levels and square of the past shocks to the mean. Consequently, a strong wind gust makes the turbulence increase and will keep influencing it for the period of time depending on GARCH and ARCH

coefficients α_i, β_i. Specifically, the volatility persistence can be estimated by the sum of coefficients from GARCH(a,b) model :

$$\sum_{j=1}^{a} \alpha_j + \sum_{i=1}^{b} \beta_i \tag{3.12}$$

AR	MA	K	GARCH	ARCH	DoF
2003: ARMA(1;1,4)-GARCH(2;1)-T					
0.95	-0.05;0.05	0.001	0.26;0.64	0.08	5.33
2004: ARMA(1;only 3)-GARCH(1;1)-T					
0.96	0.09	0.02	0.6	0.16	9.87
2005: ARMA(1;1)-GARCH(3;1)-T					
0.97	-0.04	0.01	0;0.16;0.55	0.22	9.86
2006: ARMA(2;only 5)-GARCH(1;1)-T					
0.79;0.15	0.07	0.003	0.85	0.08	14.11
2007:AR(1,7, 11)-GARCH(1;1)-T					
0.98;-0.06;0.04	-	0.02	0.6	0.16	12.59
the averaged model:ARMA(2,7, 11;1,3,4,5)-GARCH(3;1)-T					
0.93;0.03;-0.01;0.008	-0.02;0.02;0.01;0.01	0.01	0.46;0.16;0.12	0.14	10.35

Table 3.12: The estimated parameters of models fitted to January's time series. The averaged model.

The closer the sum gets to 1 the more volatility is persistent. Moreover, the sum equal to 1 would indicate that the shock will permanently affect the conditional variance (turbulence). This kind of GARCH model is a nonstationary process and is called integrated GARCH (IGARCH). All estimated models for July and January are covariance stationary i.e. $\sum_{j=1}^{a} \alpha_j + \sum_{i=1}^{b} \beta_i < 1$. However, there is a significant difference between July and January regarding the magnitude of 3.12. The July is characterized by models with the sum 3.12 varying in the range of $(0, 0.63)$ whereas the sum for January models in $(0.76, 0.98)$. This indicates that gusts occurring in January affect the turbulence for relatively longer time than gusts in July. The latter finding may lead to the general conclusion that higher wind speed exhibit higher persistence of volatility since we know that wind blows stronger in January than in July. To similar conclusions came Bradley T.Ewing et al [6].

The differences in volatility persistence are well depicted in Figure 3.14 which presents the conditional volatility simulated from the averaged model

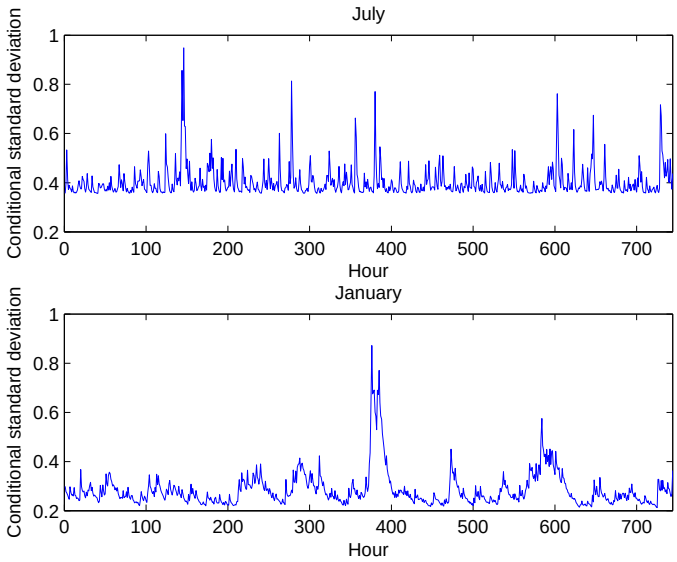


Figure 3.14: Simulated conditional standard deviation process from the averaged model for July (upper panel) and January (lower panel)

for July and January. We can notice that conditional variance in the case of July gets higher values than in the case of January. However, from physical law perspective we cannot judge it as a general rule. We know only that higher wind speed causes higher turbulence on the sea due to increase in the roughness of the sea surface and that wind speed at high vertical distance from the ground is characterized by smaller turbulence because of being separated from the frictional source- the surface of the earth.

3.5.2 Methods for verification

So far it has been shown how to choose an appropriate model for the time series at hand. However, it is now important to verify if the model that we obtained is able to generate wind speed time series with statistical features of the real wind speed. In order to perform such statistical analysis we need first to consider which (statistical) properties are prominent from the point of view of future application of the simulations. The developed models are going to be used in extensive simulation studies where different operational strategies and implementation scenarios for energy storages will be tested for their impact on power system. Therefore, since the energy storage issue will

play the leading role, it is important to check not only the correspondence of wind speed values frequency but also to get insight into dynamic temporal features like persistence.

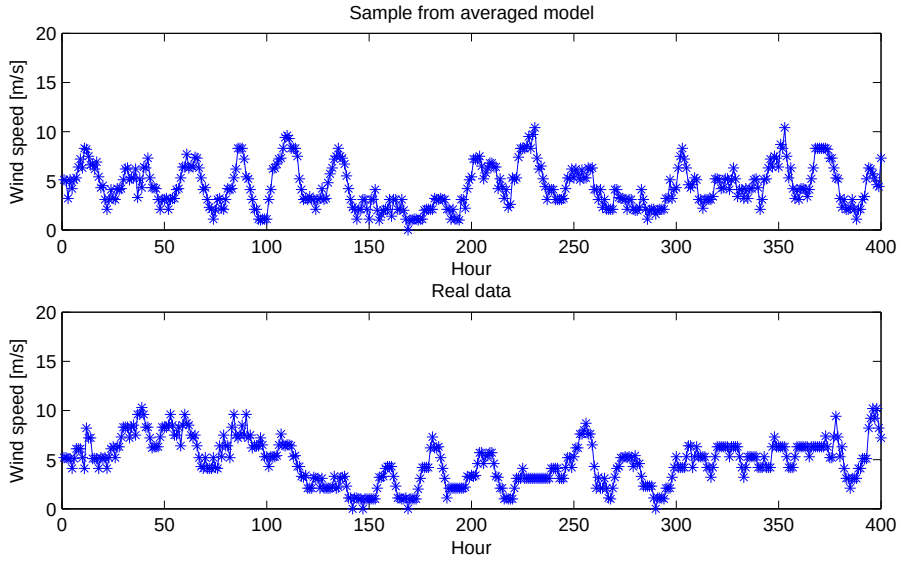


Figure 3.15: July-The comparison of simulated and observed data.

In the first stage of verification we should generate some number of samples from the model, plot them against time and compare them with the real time series plot. Obviously we do not expect from samples to resemble perfectly the real observations but we want only to roughly compare their dynamic behaviour. Significant differences would be an indication that chosen class of models is too small or the model was fitted erroneously. The Figures 3.15-3.16 present arbitrary samples from the averaged model for July and January plotted against time. There are no striking differences between the time series of observations and samples. Therefore we can step further in the statistical validation by performing the analysis of sample distributions and sample statistics.

Distribution and autocorrelation

The purpose of the averaged model is to "explain" the random process which produced the time series in July (or January). In other words we expect that the observed time series can be regarded as realizations of this model. If

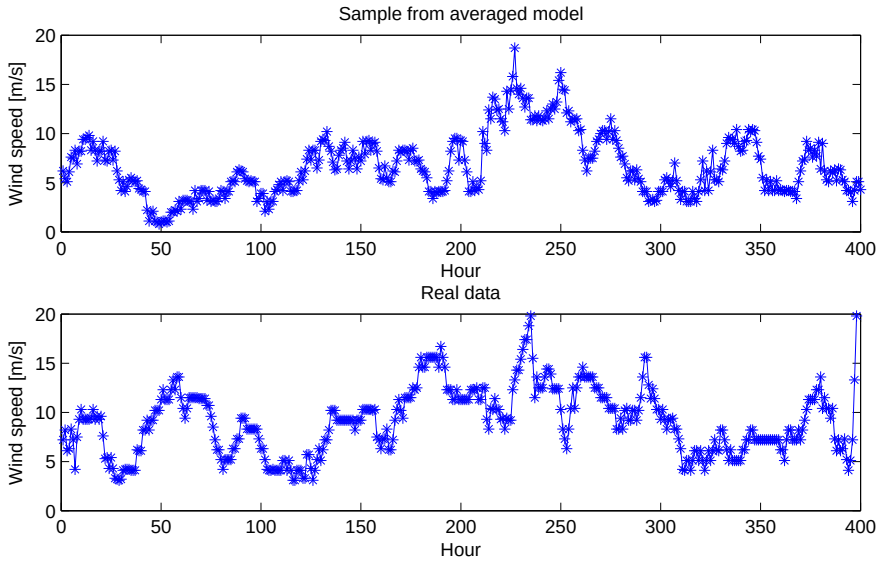


Figure 3.16: January-The comparison of simulated and observed data.

the averaged model captures well the behaviour of the unknown process, then all statistics that we derive from the observed time series should lie in the uncertainty bounds created by simulated statistics. Therefore, the model of interest is expected to produce broad enough uncertainty to be able to capture the characteristics of the data but, on the other hand, the uncertainty cannot be too big in order to make the model reliable.

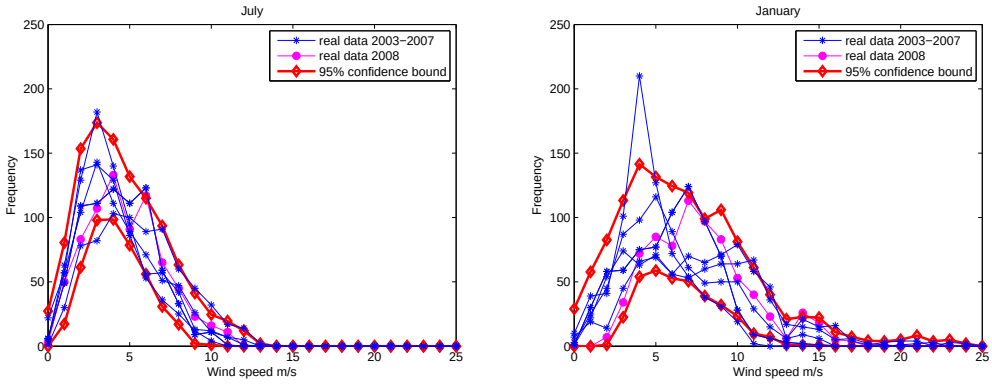
We proceed as follows: we simulate 50 samples of size 744 (31 days times 24 hours) and for each replication we evaluate the value of the statistic of interest. It is presented schematically below:

$$\begin{array}{c} y_1^{(1)} \dots y_{744}^{(1)} \to u^{(1)} \\ y_1^{(2)} \dots y_{744}^{(2)} \to u^{(2)} \\ \vdots \qquad\qquad \vdots \\ y_1^{(50)} \dots y_{744}^{(50)} \to u^{(50)} \end{array}$$

where $y_i^{(j)}$ is the j^{th} replication of the i^{th} sample member and $u^{(j)}$ is the j^{th} replication of the statistic of interest. The collection of values $u^{(1)} \dots u^{(50)}$ generated by simulation gives information about the distribution of the statistic of interest U. Therefore, it is straightforward to verify if the value of the statistic for the observed time series, e.g. u(july 2003), comes from the distribution of statistic U by checking if it lies between

approximated confidence bounds obtained from simulation.

The averaged model was built on the basis of the data from five years: 2003-2007 of a certain month. Hence, in order to verify how well our model represents the stochastic process generating the wind speeds in the chosen month it is not enough to take into account only the years used in model building but to use also an "external" time series, e.g. from year 2008.



(a) (b)

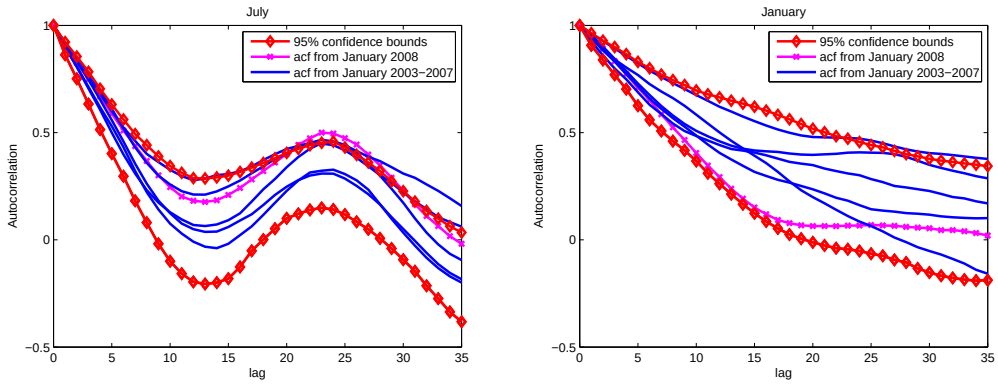
Figure 3.17: July and January-The comparison of simulated and real data with respect to distribution. The confidence bounds were computed on the basis of 50 samples

We start our analysis with considering the histogram of synthetic and real time series. Similar comparison can be found in many articles regarding modelling wind speeds [15], [18]-[21]. However, they compare the frequencies (or relative frequencies) of observed time series only with frequencies of one synthetic time series without taking into consideration the uncertainty in the model. The Figures 3.17a-3.17b present the histograms. We can notice that the distributions of observed time series lie quite confidently in the uncertainty bounds created by sample distributions. Moreover, the time series from 2008 which were not used in building the averaged model also do not raise any doubts about the models validity for both July and January. The note should be made regarding the differences between the uncertainty in the two averaged models. Namely, the averaged model for July seems to produce lower uncertainty than the averaged model for January. The higher uncertainty of January's model is caused by the higher sum of autoregressive coefficients (see section 3.2). Therefore the wind speeds in January are less mean-reverting than the wind speeds in July.

The sample autocorrelation function provides an assessment of the degree of dependence in the data. Therefore the model representing the considered time series should be characterized by the autocorrelation function which is a good approximation to the one derived from the data. Moreover, if the data are realized values of a stationary time series then the sample autocorrelation will provide us with an estimate of the ACF.

According to performed unit root tests, the considered time series are stationary and hence the analysis on the basis of sample autocorrelations is reasonable. Analogously to the analysis of histograms, we compute sample autocorrelation function for each simulated sample, construct confidence bounds and see whether they surround the sample autocorrelations derived from the real data. The Figures 3.18a-3.18b present the sample autocorrelations for the July(left panel) and January case (right panel).

The sample autocorrelations derived from time series are situated within the respective simulated confidence bounds in both cases. Therefore we can conclude that the averaged models capture well the dependence structure in the data.



(a) (b)

Figure 3.18: July and January-The comparison of simulated and original data with respect to autocorrelation. The confidence bounds were computed on the basis of 50 samples

Persistence

The next step in the validation of our models turns our attention to the dynamic features of wind speeds, both artificial and real. Recalling from

section 2.3, we define the persistence as a duration of wind speed within a specified wind class e.g. above/below a certain threshold. Hence, the persistence analysis depends considerably on the choice of wind classes. Therefore, at this point we should ask ourselves what kind of wind speeds are favorable from power generation perspective. Wind turbines are not able to produce power from arbitrarily high or low wind speed due to technical constraints. Specifically, the wind turbine starts running at wind speed around 4 m/s (it is called the cut in wind speed) whereas when wind speed exceed 25 m/s the wind turbine has to be shut down in order to avoid damaging the turbine or its surroundings. In the latter case we call the stop wind speed the cut out wind speed. As a consequence, the desired wind speed should be dominated by values falling between cut in and cut off which naturally suggests the wind speed class to consider. However, in order to make persistence analysis effective we have to gain insight into the relationship between wind speed and generated power and possibly extract more wind speed classes worth consideration. The mentioned relationship is known in wind engineering as

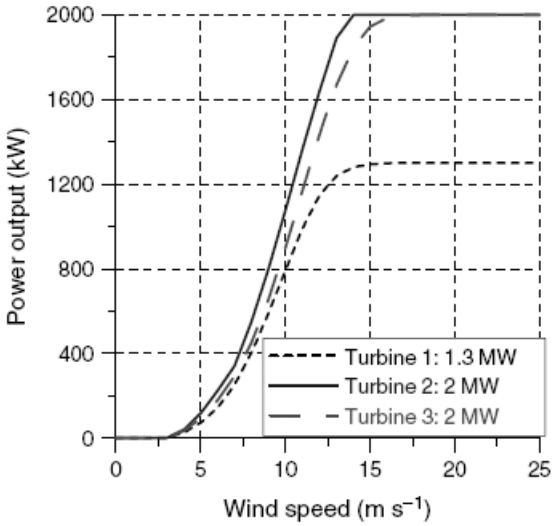


Figure 3.19: Power curves for three different turbines

the power curve. The Figure 3.19 presents power curves for three different turbines. The power curves remain at zero until the wind speed reaches the cut in and then the power starts increasing with the cube of the velocity up to the the so-called rated wind speed (it is around 15 m/s) i.e. the point at which the conversion efficiency is near its maximum. For velocities higher

than the rated power the maximal power output is obtained until the cut off wind speed is reached and wind energy converter has to be stopped so no power is then produced.

Therefore we would be interested, for example, in the persistence in the interval $(15, 25)$ m/s which corresponds to maximal power output. On the other hand, the persistence below the cut in is also worth consideration since it gives information about the distribution and amount of the run durations of no power. However, we should keep in mind that values for the cut in, cut off and rated power regard wind speed occurring at hub height. Since the strength of the wind increases with the distance from the ground level we cannot use mentioned values as thresholds for KNMI wind speed data. Therefore, they have to be transformed in order to make them comparable with our observations measured at 10 m height. The formula [6] used for computing wind speed at a certain height above ground level is presented below:

$$Y = Y_{ref} \frac{ln(z/z_0)}{ln(z_{ref}/z_0)} \tag{3.13}$$

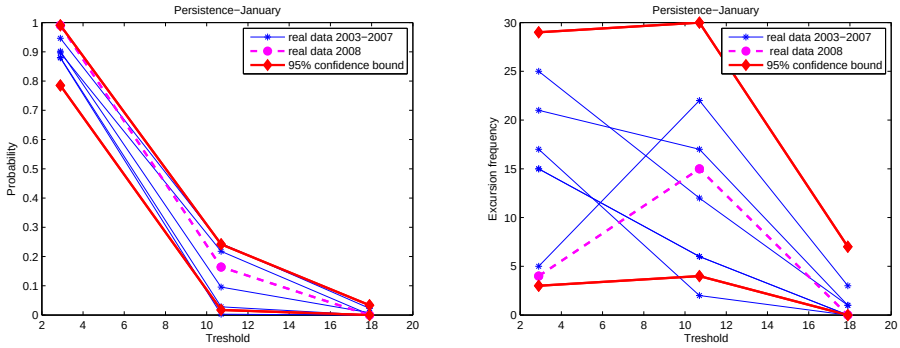
where Y is the wind speed at height of interest z above ground level,
Y_{ref} is the reference speed, i.e. wind speed we already know at height z_{ref},
z_0 is the roughness length in the current wind direction.

We assume that the cut in, cut off and rated wind speed values correspond to the hub of the standard height equal to 100 m. Thus, $Y_{ref} = 4, 15$ or 25 m/s, $z_{ref} = 100$ m and the counterparts Y will be computed for $z = 10$m and the roughness length of 0.03 which corresponds to the roughness of grass. The formula 3.13 is called the wind shear formula or power law. The application of the formula gives the following counterparts of cut in, rated and cut off wind speeds for 10 m height: 2.9, 10.7 and 17.9 m/s respectively.

The persistence analysis will be performed on the basis of the method described below which gives us statistics which are supposed to describe the persistence features. However, we need to establish first a representation of the earlier mentioned run durations using random variable notation. Assuming a certain threshold we denote E_i the ith excursion length; for a given data we obtain a set of values $\{e_1, \ldots, e_N\}$ which is a realization of a set of random variables $\{E_1, \ldots, E_N\}$, where N is the number of excursions from the assumed threshold. Therefore, analyzing the persistence we deal

[6] The formula assumes so-called neutral atmospheric stability conditions, i.e. that the ground surface is neither heated nor cooled compared to the air temperature

simultaneously with two random variables : E (the random variable which has the same distribution as all the E_is) and N.



(a) Total excursion length (b) Excursion frequency

Figure 3.20: Persistence analysis: January, 100 samples considered.

We consider three wind speed classes: above cut in, above rated and above cut out wind speed. For each of these thresholds we are interested in the total amount of time that wind speed was exceeding an assumed value. In other words we are interested in the variable $E_1 + \ldots + E_N$. Moreover, considering $(E_1 + \ldots + E_N)/744$ as a function of threshold is a survival curve of the wind speed. It provides information regarding the percentage of time that wind speed was above a certain threshold. However it does not explain the persistence entirely since two different data sets with the same survival curve may differ very much in the number of exceedances from a considered wind speed level. Therefore we need to simultaneously examine the distribution of the random variable N. If for the considered thresholds, the total excursion length and the excursion frequency computed for simulated data resemble statistically the persistence characteristics of real data then the model captures well the dynamic persistence properties. The Figures 3.20a-3.20b present the method applied to the January case. The plots take into account only wind speed levels of interest.

For the threshold equal to cut in we can notice that there is a quite visible bias in the survival plot. It indicates that synthetic sequences tend to spend relatively less time above the threshold than real data sequences. Considering the other wind speed levels, the confidence bounds reflect well the uncertainty in the total excursion length of real data.

Although simulated wind speeds seem to stay above 10.7 m/s and 17.9

threshold	μ_0	$\bar{X}$	$\bar{\sigma}/\sqrt{n}$	Z stat	p value
July					
cut in	579.17	591.54	4.05	3.05	0.003
cut in (N)	33.5	36.44	0.28	10.34	0
rated	10.33	8.3	0.8	-2.55	0.01
rated (N)	3.17	3.54	0.11	3.45	0.01
cut out	0	0	0	-	-
cut out (N)	0	0	0	-	-
January					
cut in	681.33	674.11	3.85	-1.87	0.06
cut in (N)	14.5	14.17	0.25	-1.38	0.17
rated	93.33	77.84	3.74	-4.14	0.001
rated (N)	12.33	15.47	0.25	12.62	0
cut out	4	6.3	0.77	2.9	0.004
cut out (N)	0.83	1.7	0.08	10.54	0

Table 3.13: The results from performing T-test. (N) refers to the result for random variable N

m/s for statistically similar amount of time as real wind speed, they tend to cross these levels relatively more often which can be noticed in the Figure 3.20b.

The above analysis is based only on the visual inspection of the graphs. It would be desired to make such analysis on the basis of strictly statistical tool like e.g. hypothesis testing. For a given threshold we have 6 realizations of statistics $E_1 + \ldots + E_N$ and N from real data. Let us compute averages of these realizations and proceed analogously with simulated samples i.e. let simulate 6 samples and compute the averages for the given threshold and let repeat it 100 times. Thus, we obtain 100 realizations of the averaged statistics $(x_1, \ldots, x_{100})$ and 1 realization from the data (μ_0). The idea is to apply one-sample location test i.e. t-test in order to compare the mean of the set $x_1, \ldots, x_{100}$ to a given constant μ_0. Therefore we test the null hypothesis : $H_0 : \bar{X} = \mu_0$ using the statistic $Z = (\bar{X} - \mu_0)/\bar{\sigma}/\sqrt{n}$ which has Student-T distribution with $n - 1$[7] degrees of freedom. The p-values from the test applied to July and January for each threshold of interest are shown in Table 3.13.

We can compare the results obtained from visual inspection with the

[7] in our case $n - 1$=99

output from the test. For January, although there was a noted bias in the percentage of time spent above cut in wind speed, the persistence according to test seems to be in order, the p-values are equal to 0.06 and 0.17 which means that the null hypothesis is not rejected at the 0.05 significance level. However, for the rated and cut out wind speed, there is a strong rejection of the null hypothesis when the mean excursion frequencies are compared. This is also reflected in the biases in Figure 3.20b. Concerning the mean total excursion length, the null hypothesis is not rejected at the 0.1% level.

In the case of July, there is no wind speed exceeding the cut out level in both artificial and real time series. For cut in, the mean excursion frequency of artificial time series seems to be significantly different from the mean computed from the data. On the other hand, the persistence above rated wind speed is quite well captured by the model since the p-values for mean number of exceedances and mean total time of excursions are both equal to 0.01.

We can conclude that artificial time series exhibit statistically similar total excursion length. However there are some overestimation regarding the excursion frequency.

Chapter 4

Multivariate case

The univariate time series models are useful when one considers wind power generation in a certain relatively small area. However, in order to have information about the wind power generated in larger region e.g. a country, multivariate analysis, which takes into account the spatial correlation, has to be performed. Multivariate modeling uncovers dynamic relationships among time series recorded at different places. Such an analysis is feasible since wind speed is recorder at measurement stations in equally spaced time intervals and is adjusted to the standard height of 10 meters. Moreover, modeling time series jointly may result in improving accuracy of forecasts and consequently in statistically better simulations if there is historical information on one series contained in the historical data of another. On the other hand, finding a model capturing the characteristics of each time series together with their co-movements can turn out to be not straightforward.

In the multivariate analysis we consider time series of measurements recorded at three meteorological stations: Schiphol, K13 and Cadzand which are chosen as representatives for three different types of location: inland, offshore and onshore respectively. Their positions can be found in Figure 2.1. The distance from the K13 station to Cadzand is 205 km, whereas the distances from Cadzand to Schiphol and from Schiphol to K13 are: 139 km and 146 km respectively. From the analysis in section 2.1 we know about the impact of terrain roughness on the wind speed. The table 4.1 contains averages of wind speed measured at three stations. We can see that the mean values increases with the decrease in roughness terrain. Thus, wind speeds are the highest (in average) at K13 station, whereas at Schiphol the lowest.

Apart from the opportunity of simulating wind speed at several places

	january 03	january 04	january 05	january 06	january 07
Schiphol	6.14	6.32	7.42	4.88	8.30
Cadzand	6.97	7.20	7.73	5.36	8.57
K13	9.67	9.92	11.73	7.86	12.11

Table 4.1: Mean wind speed in January evaluated from data at three stations [m/s].

simultaneously, the multivariate model will also provide us with the description of the dynamic dependence among wind measurements. There exist many factors that influence the dynamic nature of wind in space. Some of them are : the distance between stations, prevailing wind direction, common type of location and so on. We will try to recover the most important characteristics of the relationship among wind speeds from the three stations and suggest which factors are the most influential on them.

We will proceed analogously to univariate case. First, the generalized methodology will be presented, then the process of model selection with some examples will be described and at the end we will validate the simulations. Specification, estimation and diagnostic checking for multivariate ARMA-GARCH model were performed using S-plus with the Finmetrics module.

4.1 Methodology

The introduced methodology in section 3.1 is going to be generalized to handle the multivariate time series. A natural extension of the autoregressive moving average model to dynamic multivariate time series is the vector ARMA model (VARMA)[11],[13],[17],[35]. Assuming that we consider n time series variables $\{Y_{1t}\}, \ldots, \{Y_{nt}\}$ the VARMA(p,q) is expressed in the following way:

$$\mathbf{Y}_t - \mathbf{M} = \Pi_1(\mathbf{Y}_{t-1} - \mathbf{M}) + \ldots + \Pi_p(\mathbf{Y}_{t-p} - \mathbf{M}) + \tag{4.1}$$

$$+ \Psi_1\mathbf{E}_{t-1} + \ldots + \Psi_q\mathbf{E}_{t-q} + \mathbf{E}_t$$

where $\mathbf{Y}_t = (Y_{1t}, \ldots, Y_{nt})'$, Π_i, Ψ_j are $n \times n$ parameter matrices, $\mathbf{M}$ is a vector of means and $\mathbf{E}_t$ is a multivariate white noise process with zero mean which is modeled by multivariate GARCH model described below. The text concerning multivariate GARCH models was prepared on the basis of [9],[33]-[35]. The detailed information and references can be found there.

Let us denote Σ_t the conditional covariance matrix of multivariate error vector $\mathbf{E}_t$ such that:

$$\mathbf{E}_t = \Sigma_t^{1/2}\mathbf{Z}_t, \qquad \mathbf{Z}_t \sim \mathcal{MD}(0, I_n) \tag{4.2}$$

where $\mathbf{Z}_t$ is n-dimensional i.i.d. white noise with mean zero and covariance matrix I_n (identity matrix of order n). $\mathcal{MD}$ stands for multivariate density function which is usually chosen to be either the multivariate normal distribution or the multivariate Student-t distribution.

We express the conditional variance equation as:

$$Var(\mathbf{Y}_t|\Omega_{t-1}) = Var(\mathbf{E}_t|\Omega_{t-1}) = \Sigma_t = D_t R D_t = (\rho_{ij}\sigma_{iit}\sigma_{jjt})$$

where

$$D_t = diag(\sigma_{11t} \dots \sigma_{nnt})$$

σ_{iit}^2 can be defined as any univariate GARCH model, $R = (\rho_{ij})$ is constant conditional correlation matrix. Assuming that GARCH(1,1) is specified for each conditional variance in D_t we have:

$$\sigma_{iit}^2 = K_i + \alpha_i \varepsilon_{i,t-1}^2 + \beta_i \sigma_{ii,t-1}^2 \qquad i = 1, \dots, n$$

The above specified model inherited the name from the assumption regarding correlation R and is called the Constant Conditional Correlation model (CCC). When a GARCH model is used for description of conditional variances, we refer to CCC-GARCH model.

Recently multivariate GARCH models have received a lot of attention, especially due to high interest in modeling volatility of returns, resulting in the development of many various parametric formulations. The way of parametric specification is governed by the trade off between flexibility and parsimony. The models which can capture sophisticated dynamics of the conditional variances and covariances well enough, are represented by many parameters and consequently lead to estimation problems in highly dimensional time series. The high complexity of the model can be reduced by reducing the number of parameters or/and imposing constraints on the matrix structure.

The multivariate GARCH models can be divided into several categories. One category includes models of conditional variances and correlations. These models are based on the decomposition of the conditional covariance matrix Σ_t into conditional standard deviations and correlations. The simplest such a model is Constant Conditional Correlation already presented in this section. It was proposed by Borllerslev [32]. The advantages of

multivariate modeling using CCC are working with relatively small number of parameters: $n(n-1)/2$ and computational attractiveness since during estimation, one has to invert the conditional correlation matrix only once per iteration. Because we are going to work with highly dimensional time series during the project, the aforementioned properties make the model suitable for our purposes. On the other hand, the simplicity in estimation has to be paid by lower flexibility of the model. Precisely, the conditional variances are modeled separately and therefore there is no interactions between them. Moreover, since the conditional correlation is assumed to be constant, the conditional covariances are proportional to the product of the corresponding standard deviations and as a result the conditional variances and covariances vary in a restricted way so that the conditional correlations are time-invariant. Therefore, it is important to verify if the assumption of constant conditional correlations may seem to be realistic for the multivariate time series of interest. We will try to answer this question by considering much more flexible models and compare their results in diagnostic checking. If the assumption of constant correlation is not appropriate for our data, the CCC model will provide much worse fit then more complex models.

The models allowing for more dynamics are contained in a class where conditional covariance is modeled directly. They are generalizations of univariate GARCH models and thus, every conditional variance and covariance may be a function of lagged conditional variances and covariances, as well as lagged squared residuals and cross-products of residuals. As a first such a model VEC was proposed by Bollerslev and is defined as follows (presented in VEC(1,1) form):

$$vech(\Sigma_t) = \mathbf{K} + \mathbf{A}vech(\mathbf{E}_t\mathbf{E}_t') + \mathbf{G}vech(\Sigma_{t-1})$$

where vech(.) denotes the operator that stacks the lower triangular portion of n×n as a $n(n+1)/2 \times 1$ vector. $\mathbf{A}$ and $\mathbf{G}$ are square parameter matrices of order $n(n+1)/2$ and $\mathbf{K}$ is a $n(n+1)/2 \times 1$ vector. The model was further simplified to the form where matrices $\mathbf{A}$ and $\mathbf{G}$ are diagonal and called Diagonal VEC model (DVEC). In this way, the conditional variances follow a GARCH process, whereas the conditional covariances can be treated as a GARCH model in terms of the cross-moment of the errors. The number of estimated parameters in DVEC(1,1) model is $3n(n+1)/2$.

Since the VEC model does not produce positive definite conditional variances, the need was to construct a model which would not posses this drawback. Therefore BEKK model was proposed by Baba, Engle, Kraft and

Kroner. BEKK(1,1) model is defined as:

$$\Sigma_t = \mathbf{K}\mathbf{K}' + \mathbf{A}'\mathbf{E}_{t-1}\mathbf{E}_{t-1}\mathbf{A} + \mathbf{G}'\Sigma_{t-1}\mathbf{G}$$

where K is a lower triangular matrix, but $\mathbf{A}$ and $\mathbf{G}$ are unrestricted square matrices. The dynamics allowed by the BEKK model are richer than the DVEC model. Unlike in the DVEC model, the conditional variance of one series can influence the conditional variance of another series. It is possible thanks to $n(n-1)$ more parameters than DVEC model requires.

The estimation of parameters is more difficult in VEC and BEKK models since they need inverting the conditional covariance matrix for each time t in every iteration of the numerical optimization. If the number of time series considered in multivariate analysis is high and additionally they cover long time span, the numerical procedure is time consuming and numerically unstable.

Although the complex models do not seem to be appropriate for highly dimensional wind speed time series (due to mentioned numerical problems), we would like to take them into consideration in order to see whether the assumption of constant correlation is realistic. We will perform the analysis only for three time series so the computational problems should not occur.

4.2 Model selection

Specification of the ARMA order starts being not an easy task in multivariate analysis. VARMA models suffer from nonuniqueness of representation.

For example let us consider bivariate VAR(1) process (VARMA(1,0)):

$$Y_t = \Phi Y_{t-1} + E_t \qquad \{E_t\} \sim WN(0, \Sigma)$$

with

$$\Phi = \begin{bmatrix} 0 & 0.5 \\ 0 & 0 \end{bmatrix}$$

it may be shown that Y_t has an alternative representation as an MA(1) (VARMA(0,1)) process

$$Y_t = E_t + \Phi E_{t-1}$$

see [11],[13] and [35] for more details. This example shows that it may be the case that we cannot distinguish between multivariate ARMA models of different orders. Therefore, in order to avoid such problems, we will restrict our attention to multivariate autoregressive models (VAR) only.

The lag length for the VAR(p) model may be determined using model selection criteria 3.8 and 3.10. The AIC and BIC for multivariate Gaussian VAR(p) can be rewritten as follows:

$$AIC(p) = ln|\bar{\Sigma}(p)| + \frac{2}{T}pn^2$$

$$BIC(p) = ln|\bar{\Sigma}(p)| + \frac{lnT}{T}pn^2$$

where $\bar{\Sigma}(p) = T^{-1}\sum_{t=1}^{T}\hat{E}_t\hat{E}_t'$ is the residual covariance matrix and T is the sample size. The last information criteria, which did not appear in univariate analysis, is the Hannan-Quinn information criterion defined as:

$$HQ(model) = -2ln(L_T(k))/T + 2kln(ln(T))/T \tag{4.3}$$

In the case of Gaussian VAR(p) model HQ takes the following form

$$HQ(p) = ln|\bar{\Sigma}(p)| + \frac{2lnlnT}{T}pn^2.$$

The information criteria very often indicate different lag lengths. The solution to this problem could be applying likelihood ratio test. The decision may be also supported by the diagnostic checking results. If we notice that the difference in diagnostic results between models with significantly different orders is negligible then it can be decided to choose a simpler model.

Before we describe the diagnostic checking tools we need to define cross covariance and correlation matrices. For a univariate time series the autocovariances and autocorrelations summarize the linear time dependence in the data. In the case of multivariate time series $\mathbf{Y_t}$ each component has autocovariances and autocorrelations but there are also lead-lag covariances and correlations between all possible pairs of components. For a vector time series $\{\mathbf{Y}_t\}$ with mean vector $\mathbf{0}$ let,

$$\Gamma(l) = E(\mathbf{Y}_t\mathbf{Y}_{t-l}) = \{\gamma_{ij}(l)\}, \qquad i,j = 1\ldots n, l = 0, \pm1, \pm2, \ldots$$

be the lag l-cross-covariance matrix and $\rho(l) = \{\rho_{ij}(l)\}$ be the corresponding cross-correlation matrix.

The univariate Ljung Box test may be extended to multivariate case. The multivariate portmanteau test is designed for testing:

$$H_0 : R_h = (\rho(1), \ldots, \rho(h)) = 0 \text{ against } H_1 : R_h \neq 0$$

The test statistic was proposed by Hosking:

$$Q_h := T^2 \sum_{i=1}^{h} (T-i)^{-1} tr(\hat{\Gamma}(i)' \hat{\Gamma}(0)^{-1} \hat{\Gamma}(i) \hat{\Gamma}(0)^{-1})$$

Under the null hypothesis the Q_h is distributed asymptotically as $\chi^2(n^2 h)$. We will call this test a multivariate Ljung Box test.

The model selection procedure should be performed analogously to the univariate case. The difference is that we have to take into consideration the cross correlations. First we find an appropriate AR model using information criteria. Then we perform univariate and multivariate Ljung Box tests on residuals. We plot the sample autocorrelation and crosscorrelations in order to see if there are significant departures from 95% confidence bounds. Next we check for heteroscedasticity by plotting the auto and cross correlations of squared residuals. We can perform univariate and multivariate Ljung Box tests as well. If there is conditional heteroscedasticity we try to find an appriopriate multivariate GARCH. The next section will give evidence that CCC-GARCH model should be an adequate for modeling wind speed volatility. Therefore we need to choose order of the CCC-GARCH model using information criteria. In multivariate case usually CCC-GARCH(1,1) performs best. In the final step we need to check if there are no significant auto and cross correlations coefficients in the squared standardized residuals. Again, by making a plot and performing Ljung Box tests. Univariate ARCH tests may be used as well.

4.3 CCC assumption

Since the assumption of constant correlation in a CCC-GARCH model may turn out to be unrealistic in many empirical applications, we would like to perform some tests to find out if this is the case. The section 4.1 presented models DVEC and BEKK which allow for non-constant correlation. In order to verify if the constant correlation assumption is reasonable for modeling wind speed dynamics, we will compare the diagnostic results of all three models : CCC-GARCH, DVEC and BEKK. For this purpose we will use univariate and multivariate Ljung Box test for testing the serial correlation in standardized squared residuals. Specifically, we will compare the p-values evaluated for different models.

If the constant correlation constrains the model's ability to capture the volatility dynamics substantially, it will be seen in low p-values. Moreover these p-values will be significantly lower than the p-values obtained for

DVEC and BEKK models. On the other hand, comparable p-values may indicate that the difference in capturing the dynamical behavior is relatively small. Consequently, it would show that higly sophisticated model is not compulsory.

We will consider models CCC-GARCH(1,1), DVEC(1,1) and BEKK(1,1) since they were found to perform best in their own classes according to AIC and BIC. The results are presented for lag length equal to 20, see Table 4.2. Other lag lengths were also studied, e.g. 10 and 30, and similar conclusions were derived.

model	LB Sch	LB Cad	LB K13	MLB
	January 2003			
CCC	0.36	0.84	6.9e-006	1e-008
DVEC	0.44	0.80	5.4e-006	5e-009
BEKK	0.02	0.34	0	0
	January 2004			
CCC	0.41	0.87	0.01	3e-008
DVEC	0.46	0.85	0.03	8e-007
BEKK	0.01	0.23	0.03	0.01
	January 2005			
CCC	0.33	0.61	0.87	0.03
DVEC	0.30	0.71	0.74	0.03
BEKK	3e-008	0.38	0.57	9e-006
	January 2006			
CCC	0.28	0.32	0.67	0.50
DVEC	0.82	0.79	0.68	0.98
BEKK	0.41	0.07	0.18	0.32

Table 4.2: P-values from performed univariate and multivariate Ljung Box tests on standardized squared residuals from multivariate GARCH models. The chosen lag length is 20. LB stands for Ljung Box test and MLB for Multivariate Ljung Box test.

We can notice from the table that the univariate Ljung Box test provides very similar p-values for CCC-GARCH and DVEC models. It can be justified since in both models conditional variances follow a univariate GARCH(1,1) model. However, they model conditional covariances differently. In CCC-GARCH model the conditional covariances are proportional to the product of conditional variances and respective conditional

correlation, whereas the DVEC models conditional covariance as univariate GARCH in terms of cross-products of errors. The mentioned difference has its result in the p-values of multivariate Ljung Box test. The p-values for DVEC are higher than for CCC-GARCH but certainly not significantly higher. The BEKK is a quite complex model, its performance, according to the table, seems to be worse than the others. Quite often the p-values are substantially lower. It suggests than the BEKK model is too elaborate. Indeed, a large number of estimated parameters was not significant.

From the analysis of p-values we can conclude that the assumption of constant correlation seems to be reasonable. The CCC-GARCH model outperformed the BEKK model and gave a comparable results to the DVEC model. Therefore, even if the conditional correlation is not constant, the CCC-GARCH model should capture the volatility dynamics satisfactorily enough. As a result it is a very attractive candidate for modeling highly dimensional multivariate time series of wind speed.

4.4 The analysis of simulations

4.4.1 Averaged model

Considering the univariate case we showed that the constructed averaged models perform well enough for simulation purposes. In this section we will extend this approach to the higher dimension. We analyze January time series from the three KNMI stations: Schiphol, Cadzand and K13 in the usual time span 2003-2007. The tools discussed in section 4.2 has helped in the specification and diagnostic process of choosing a model. Consequently we obtained five trivariate time series models describing the wind speed in January from five consecutive years.

This section aims to discuss the models obtained for Januaries as well as the averaged model constructed on their basis. In order to validate if the averaged model captures well the dynamic mechanism of wind speed, the next section will provide the analysis of artificial time series. If the result is satisfactory then it would be the next indication, apart from section 4.3, that the constant conditional correlation is a reasonable assumption.

Just like in univariate case, the fitted univariate models differ from each other by the VAR order. On the other hand, the multivariate GARCH models are of the same order - CCC-GARCH(1,1) was found to perform best in all five cases according to information criteria. The estimated parameters for VAR-CCC-GARCH models are given in the appendix.

Since the multivariate analysis emphasize on the structure of relationship

among time series, we start our consideration with analyzing the feedback or/and unidirectional relations. For this purpose we present autoregressive parameters in the form of indicator symbols. The sign plus means that the t-ratio of the respective coefficient is higher than 1.5. The minus appears otherwise. Each horizontal sequence of matrices corresponds to the VAR parametric matrices of a model fitted to January. This way of presenting the parametric matrices clearly depicts the significance patterns.

$$\begin{bmatrix} + & - & + \\ + & + & + \\ - & - & + \end{bmatrix} \begin{bmatrix} - & + & - \\ - & - & - \\ - & - & - \end{bmatrix} \begin{bmatrix} - & + & + \\ - & - & - \\ - & - & + \end{bmatrix}$$

$$\begin{bmatrix} + & + & + \\ + & + & + \\ + & + & + \end{bmatrix} \begin{bmatrix} - & - & - \\ - & - & + \\ - & - & - \end{bmatrix} \begin{bmatrix} - & + & + \\ + & - & - \\ - & - & + \end{bmatrix}$$

$$\begin{bmatrix} + & + & + \\ + & + & + \\ - & - & + \end{bmatrix}$$

$$\begin{bmatrix} + & + & + \\ + & + & + \\ - & - & + \end{bmatrix}$$

$$\begin{bmatrix} + & + & + \\ + & + & + \\ + & + & + \end{bmatrix} \begin{bmatrix} - & - & - \\ - & - & - \\ - & - & - \end{bmatrix} \begin{bmatrix} - & + & + \\ + & - & - \\ - & - & + \end{bmatrix}$$

We can notice the common features in the VAR structures: the lag one matrix contains the highest amount of the significant parameters. It means that the wind speed at present is related with wind speeds at different stations recorded one hour ago. This type of relationship is, however, not always symmetric. For example in January 2005 and 2006 there is an unidirectional relationship between wind speed at K13 and the others - an increase in wind speed at K13 will be followed by an increase [1] in wind speeds at the other sites a one hour later. It may be explained by the prevailing westerly wind direction

[1] the coefficients in the lag one matrix are positive

The lag two coefficient matrix either is zero or contains only one significant parameter as in January 2003 and 2004. Therefore we can conclude that wind speeds separated by two hours tend to not interact with each other. We should take a note that the significance patterns of the nonzero lag-three parameters matrices are very similar. Usually, the wind speed at Schiphol relates with wind speeds at K13 and Cadzand recorded 3 hours ago. Moreover, the K13 wind speed is significantly correlated with its lag of order 3. However, most of significant coefficients in lag three matrix are negative which is quite surprising.

The CCC-GARCH models the conditional variances individually i.e. like in the univariate case. Therefore we have compared the ARCH and GARCH coefficients estimated for univariate GARCH in the previous chapter and estimated for the Schiphol component in multivariate GARCH. We do not notice any global increase/decrease in the coefficients.

In order to obtain the averaged model we took the average of the respective parameters. The obtained coefficient matrices with the model specification are given below.

$$Y_t = \Pi_1 Y_{t-1} + \Pi_2 Y_{t-2} + \Pi_3 Y_{t-3} + E_t$$
$$E_t = \Sigma_t^{1/2} Z_t$$
$$\Sigma_t = D_t R D_t$$
$$D_t^2 = K + A E_t^2 + G D_{t-1}^2$$

where

$$\Pi_1 = \begin{bmatrix} 0.77 & 0.11 & 0.17 \\ 0.10 & 0.85 & 0.09 \\ 0.03 & 0.02 & 0.99 \end{bmatrix} \Pi_2 = \begin{bmatrix} 0.01 & 0.01 & 0.03 \\ 0 & -0.01 & -0.01 \\ -0.01 & -0.01 & 0.04 \end{bmatrix}$$

$$\Pi_3 = \begin{bmatrix} 0 & -0.05 & -0.05 \\ -0.03 & 0 & -0.02 \\ -0.01 & -0.01 & -0.08 \end{bmatrix}$$

$$K = \begin{bmatrix} 0.02 \\ 0.02 \\ 0.02 \end{bmatrix} A = \begin{bmatrix} 0.14 & 0 & 0 \\ 0 & 0.15 & 0 \\ 0 & 0 & 0.18 \end{bmatrix} G = \begin{bmatrix} 0.52 & 0 & 0 \\ 0 & 0.59 & 0 \\ 0 & 0 & 0.44 \end{bmatrix}$$

$$R = \begin{bmatrix} 1 & 0.12 & 0.13 \\ 0.12 & 1 & 0.04 \\ 0.13 & 0.04 & 1 \end{bmatrix}$$

It is striking that the lag two and three AR parameter matrices are full of negative values. However, in comparison with Π_1 coefficients are very low. The persistence in volatility corresponding to Schiphol is lower by 2.2 than in the univariate case. In order to find out if the multivariate averaged model is suitable for applications, it has to go through a process of validation described in the next section. However, we first plot the simulated multivariate time series and compare it with the original e.g. from January 2004, see Figure 4.1. We can see that the artificial multiple time series exhibit very similar behaviour to the original time series.

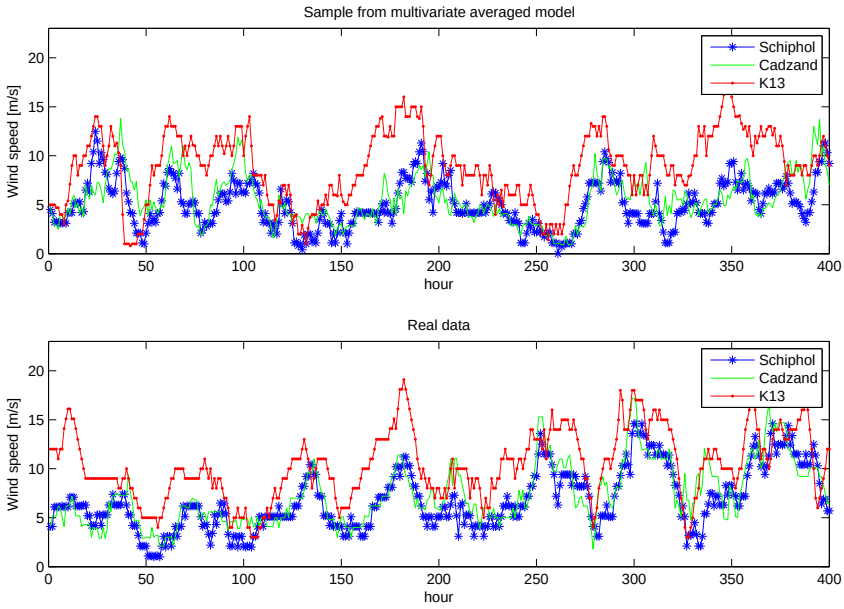
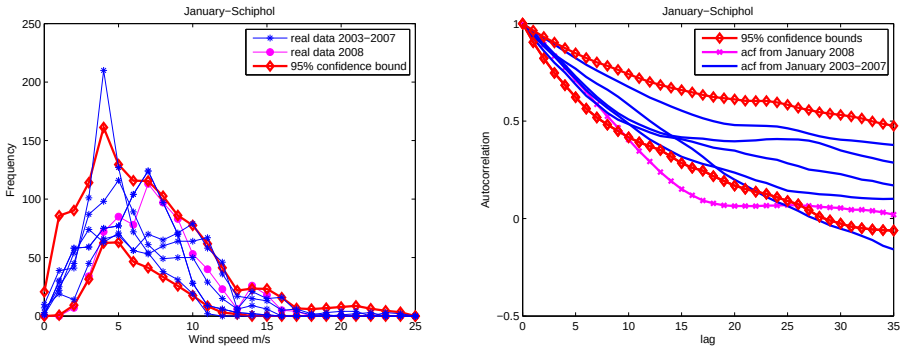


Figure 4.1: The comparison of simulated and observed time series.

4.4.2 Autocorrelations and distributions

Analogously to the univariate case, we perform the analysis of the histograms and sample autocorrelations derived from real and artificial time series. However, this time we have to work with wind speeds from three different stations. The analysis of histograms has revealed that in all three cases the results were satisfactory i.e. the model is able to produce the uncertainty of the similar degree to the real data. The Figure 4.2a presents the histograms of wind speed recorded at Schiphol station together with simulated confidence bounds. Comparing it to the histograms from univariate case (see Figure 3.17b), there are no present differences, i.e. the both models: univariate and multivariate describe uncertainty in a similar way.

In the case of sample autocorrelations, the multivariate model captures the dependence in the data in each multiple time series component adequately. The Figure 4.2b shows the sample autocorrelations from the wind speeds at Schiphol together with confidence bounds computed from simulations. The sample autocorrelations from real data lie confidently within the bounds, except from SACF in January 2008 which slightly leaves out the bounds. This result is comparable with the one from univariate case. However, in the univariate case the dependence in Schiphol time series seems to be captured a bit better (see Figure 3.18b).

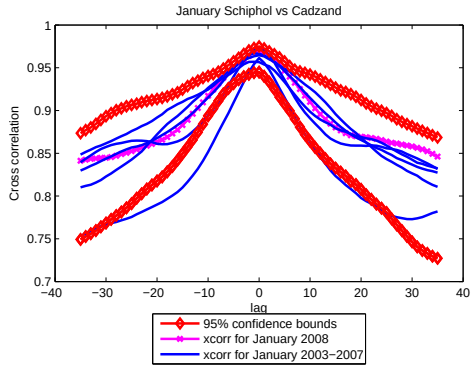


(a) (b)

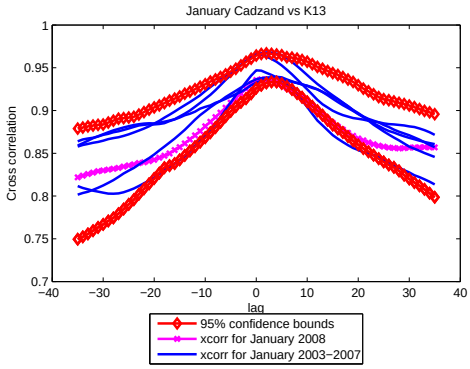
Figure 4.2: The comparison of histograms and sample autocorrelations of real time series recorded at Schiphol and artificial time series generated from the averaged multivariate model

Since the multivariate model is aimed to describe the dynamic relationships among time series, it is of importance to analyze also the sample cross-

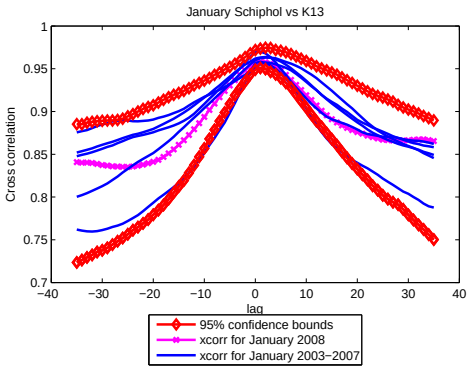
correlations in simulated vector time series. The Figures 4.3a-4.3c present the sample cross-correlations. We can notice that in all cases the sample cross-correlations lie within their respective confidence bounds computed from simulations.



(a) Schiphol-Cadzand



(b) Cadzand-K13



(c) Schiphol-K13

Figure 4.3: The comparison of sample cross-correlations from real and simulated time series

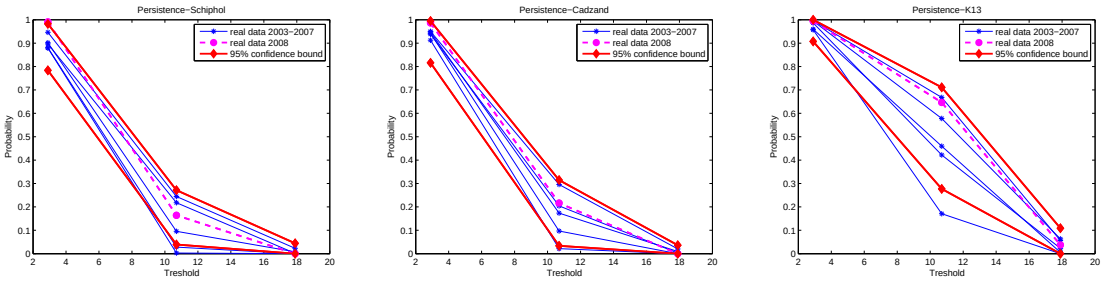
In the case of Figures 4.3b and 4.3c there is a noticeable asymmetry. The present observations recorded at Cadzand and Schiphol are more correlated with past observations recorded at K13 station than present K13 observations with past Cadzand/Schiphol observations. On the other hand the Cadzand and Schiphol time series exhibit rather a feedback relationship. These findings are in accordance with the results from the previous section.

It is caused due to, inter alia, geographical position of stations station and the prevailing wind directions.

4.4.3 Persistence

In the previous section we have shown that the averaged multivariate model captures satisfactorily the dynamic relationships among the series as well us the temporal dependence in the univariate components. This section aims to verify if it is able to generate the time series with appropriate persistence characteristics.

The Figure 4.4 presents the total excursion length for each station in the form of survival curve. We can notice that artificial wind speeds spend similar amount of time above rated and cut out thresholds to the real wind speed at all considered stations. In the case of cut in wind level, the simulations seem to underestimate slightly the total excursion length. This result is very similar to the one obtained for univariate case. Moreover according to the performed tests, in the case of all thresholds at all considered stations the null hypothesis that total excursion length obtained from simulations is not different statistically from excursion length from data is not rejected at the 0.1% significance level, see Table 4.3.

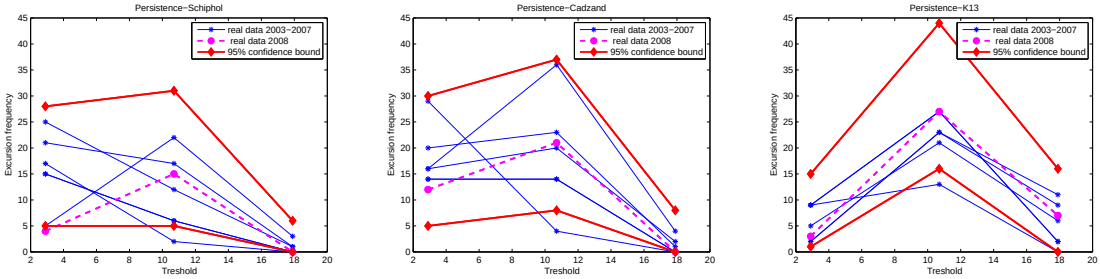


(a) Schiphol (b) Cadzand (c) K13

Figure 4.4: Total excursion length

Considering the excursion frequencies depicted in Figure 4.5, we can notice a significant bias for the rated and cut out wind speeds especially in the case of Schiphol and K13 station. The performed tests confirm it by rejecting the null strongly. Although according to Figures the excursion length for cut in is quite well reflected in simulations, the test rejects the null for this threshold as well.

Analyzing the Figures 4.4 and 4.5 we can draw conclusions regarding the character of persistence of wind speeds at considered stations. In Figure 4.4 we see that the percentage of time, that wind speed spend above thresholds, increases with the mean characteristic for each station (see Table 4.1). On the other hand, in 4.5 we notice that the number of excursions tend to stay constant with the increase of the mean. Therefore, the higher wind speeds (in average) at certain location the more persistent they are. In the case of our three stations, the highest persistence occurs at K13.



(a) Schiphol (b) Cadzand (c) K13

Figure 4.5: Excursion frequency

Although the artificial wind speeds simulated from the averaged multivariate model exhibit satisfactorily the total excursion length, the excursion frequency needs some improvements. It may be the case that finding a time series model capturing the persistence nature may be very difficult and it should be a topic for future research.

Threshold	Schiphol			Cadzand			K13		
	μ	X	p-value	μ	X	p-value	μ	X	p-value
cut in	681.33	674.77	0.08	704	699.31	0.15	730.17	726.72	0.025
cut in (N)	14.5	16.7	2e-011	17.8	15.2	2e-015	5	5.23	0.12
rated	93.33	86.69	0.14	125	109.27	0.001	365.17	349.3	0.05
rated (N)	12.33	16.99	0	19.67	22.33	9e-014	22.33	30.28	0
cut out	4	5.7	0.02	4	6.1	0.008	23	21.12	0.35
cut out (N)	0.8	2.05	0	1.17	2.1	2e-016	5.8	4.96	0

Table 4.3: Persistence test

Chapter 5

Conclusions and future research

We have presented univariate and multivariate modeling of wind speed. The time series were analyzed using strictly statistical tools in order to provide model capturing the important features of the real time series i.e. time dependence structure, spatial dependence, marginal distributions, periodic behaviour and persistence. Since wind speeds exhibit strong serial correlation and turned out to be conditionally heteroscedastic we proposed methodology based on ARMA-GARCH framework. We have shown that artificial time series simulated from univariate and multivariate models exhibited most of the aforementioned properties.

Since the models are intended to be used for simulations purposes only, we have not considered the forecasting issue. However, the analysis of predictions made by the ARMA-GARCH models would be a valuable validation method. Therefore it should be a topic for future research.

The univariate and multivariate models have shown to capture very well the dynamic dependence structure between multiple time series components as well as the time dependence in each component. However, the analysis of persistence has revealed a defect of the models. It can be the case that ARMA-GARCH model is not enough to model satisfactorily the persistence characteristics and has to be extended e.g. by additional parameters. The literature offers a broad class of asymmetric GARCH models which takes into account the sign and magnitude of innovation noise term by using additional parameter [4],[17],[9]. In this way the volatility reacts differently on positive and negative error. We have checked for this asymmetry using the news impact curve [36] which provides information about the relationship

between conditional variance and the shock term. We noticed quite significant asymmetry in the volatility. However, this issue is out of the scope of the thesis and can provide direction for the future improvements of the models.

Considering the multivariate time series, we came to conclusion that constant conditional correlation is enough to satisfactorily capture the comovements in volatility. We should make a note that CCC models were further extended by retaining the decomposition of conditional covariance but making the conditional correlation matrix time-varying. Resulting in Dynamic conditional correlation model and Extended conditional correlation model [33],[34]. It may be of interest to check the performance of the extended models and compare them with CCC. The class of conditional correlation models receives nowadays much attention in contrast to VEC and BEKK which have already matured [34].

The simulated wind speeds will be further transformed to power generated at a certain area. In section 2.3 we presented the power curve which represents the relationship between wind speed and power for one wind turbine. In order to obtain the aggregated power generation for example at wind farm having at hand only one wind speed time series we need to apply the so called multi-turbine power curve. The details regarding this approach can be found in [37].

Chapter 6

Appendix

Unit root tests performed on transformed time series recorded at Schiphol in July and January. ERS- Elliot, Rothenberg and Stock test; DF-GLS- efficient version of the Augmented Dickey Fuller (ADF) test; MPP- efficient modified Phillips-Perron (PP) test.

January	ERS	DF GLS	MPP
2003	1.76 xx	-2.65 xx	-2.67 xx
2004	2.42 x	-2.87 xx	-2.89 xx
2005	1.48 xx	-2.94 xx	-2.97 xx
2006	0.82 xx	-4.10 xx	-4.04 xx
2007	1.95 x	-2.63 xx	-2.69 xx

Table 6.1: Unit root tests- January. x significant at 5% level, xx significant at 1 % level

January	ERS	DF GLS	MPP
2003	2.04 x	-2.51 x	-2.60 xx
2004	2.70 x	-2.16 x	-2.21 x
2005	2.64 x	-2.26 x	-2.21 x
2006	0.93 xx	-3.49 xx	-3.88 xx
2007	4.05	-1.78	-1.77

Table 6.2: Unit root tests- July. x significant at 5% level, xx significant at 1 % level

The estimated parameters of the multivariate AR-CCC-GARCH model:

```
###############
#
#  JANUARY03:
#
##############

Conditional Distribution:  t
 with estimated parameter 9.354668 and standard error 1.499523

-----------------------------------------------------------------

Estimated Coefficients:
-----------------------------------------------------------------
                 Value Std.Error  t value Pr(>|t|)
AR(1; 1, 1)  0.7777006  0.039710  19.5845 0.000000
AR(1; 2, 1)  0.0680039  0.036861   1.8449 0.065458
AR(1; 3, 1)  0.0037046  0.026839   0.1380 0.890254
AR(1; 1, 2)  0.0308451  0.031814   0.9696 0.332589
AR(1; 2, 2)  0.8836537  0.040767  21.6756 0.000000
AR(1; 3, 2)  0.0176815  0.025209   0.7014 0.483287
AR(1; 1, 3)  0.1486299  0.038199   3.8909 0.000109
AR(1; 2, 3)  0.0482727  0.041612   1.1601 0.246407
AR(1; 3, 3)  1.0229353  0.040767  25.0922 0.000000
AR(2; 1, 1)  0.0655072  0.052746   1.2419 0.214654
AR(2; 2, 1)  0.0638300  0.044396   1.4378 0.150933
AR(2; 3, 1)  0.0035969  0.033716   0.1067 0.915069
AR(2; 1, 2)  0.1076747  0.039523   2.7244 0.006596
AR(2; 2, 2) -0.0125846  0.055131  -0.2283 0.819503
AR(2; 3, 2) -0.0262344  0.035135  -0.7467 0.455498
AR(2; 1, 3)  0.0531081  0.054017   0.9832 0.325851
AR(2; 2, 3)  0.0240876  0.057929   0.4158 0.677670
AR(2; 3, 3)  0.0781710  0.058236   1.3423 0.179913
AR(3; 1, 1) -0.0276876  0.038694  -0.7155 0.474502
AR(3; 2, 1) -0.0302175  0.033887  -0.8917 0.372836
                   Value Std.Error   t value  Pr(>|t|)
   AR(3; 3, 1)  0.0009329  0.027051   0.03449 0.9724994
   AR(3; 1, 2) -0.0844407  0.027525  -3.06783 0.0022359
```

```
   AR(3; 2, 2) -0.0120984  0.037155  -0.32562 0.7448041
   AR(3; 3, 2) -0.0042157  0.026815  -0.15721 0.8751204
   AR(3; 1, 3) -0.0933161  0.040454  -2.30673 0.0213488
   AR(3; 2, 3) -0.0546515  0.040665  -1.34394 0.1793837
   AR(3; 3, 3) -0.1367913  0.038502  -3.55282 0.0004056
       A(1, 1)  0.0045524  0.001542   2.95168 0.0032619
       A(2, 2)  0.0062785  0.002178   2.88270 0.0040587
       A(3, 3)  0.0268836  0.008321   3.23085 0.0012896
 ARCH(1; 1, 1)  0.1022539  0.031217   3.27558 0.0011043
 ARCH(1; 2, 2)  0.1138733  0.030264   3.76265 0.0001816
 ARCH(1; 3, 3)  0.1763428  0.060337   2.92261 0.0035784
GARCH(1; 1, 1)  0.8205182  0.045258  18.12996 0.0000000
GARCH(1; 2, 2)  0.7932420  0.052606  15.07886 0.0000000
GARCH(1; 3, 3)  0.2465766  0.189519   1.30106 0.1936473

Estimated Conditional Constant Correlation Matrix:
------------------------------------------------------------
        1.1     1.2     1.3
1.1 1.00000 0.11475 0.06711
1.2 0.11475 1.00000 0.01097
1.3 0.06711 0.01097 1.00000

Standard Errors:
        [,1]    [,2]    [,3]
[1,]      NA 0.04448 0.04107
[2,] 0.04448      NA 0.04468
[3,] 0.04107 0.04468      NA

------------------------------------------------------------
#################
#
#  JANUARY04:
#
#################
Conditional Distribution:  t
 with estimated parameter 14.90368 and standard error 4.052508

------------------------------------------------------------

Estimated Coefficients:
```

```
------------------------------------------------------------
                Value Std.Error  t value   Pr(>|t|)
AR(1; 1, 1)  0.720524  0.040368  17.8490 0.000e+000
AR(1; 2, 1)  0.084101  0.037262   2.2570 2.430e-002
AR(1; 3, 1)  0.062108  0.030366   2.0453 4.118e-002
AR(1; 1, 2)  0.168312  0.031858   5.2832 1.677e-007
AR(1; 2, 2)  0.935483  0.041412  22.5898 0.000e+000
AR(1; 3, 2)  0.049460  0.026494   1.8668 6.233e-002
AR(1; 1, 3)  0.182512  0.033970   5.3727 1.044e-007
AR(1; 2, 3)  0.140536  0.039331   3.5732 3.758e-004
AR(1; 3, 3)  1.024580  0.042909  23.8778 0.000e+000
AR(2; 1, 1)  0.019449  0.049583   0.3922 6.950e-001
AR(2; 2, 1)  0.031534  0.047141   0.6689 5.037e-001
AR(2; 3, 1) -0.046443  0.038658  -1.2014 2.300e-001
AR(2; 1, 2) -0.024696  0.040640  -0.6077 5.436e-001
AR(2; 2, 2) -0.040369  0.053885  -0.7492 4.540e-001
AR(2; 3, 2) -0.026620  0.036738  -0.7246 4.689e-001
AR(2; 1, 3)  0.061172  0.049344   1.2397 2.155e-001
AR(2; 2, 3) -0.107542  0.058522  -1.8376 6.652e-002
AR(2; 3, 3)  0.074875  0.062267   1.2025 2.296e-001
AR(3; 1, 1)  0.023048  0.038363   0.6008 5.482e-001
AR(3; 2, 1) -0.060211  0.037228  -1.6173 1.062e-001
                  Value Std.Error  t value   Pr(>|t|)
   AR(3; 3, 1) -0.012838  0.029071 -0.44159 6.589e-001
   AR(3; 1, 2) -0.071603  0.030375 -2.35731 1.867e-002
   AR(3; 2, 2) -0.014849  0.038279 -0.38792 6.982e-001
   AR(3; 3, 2) -0.001338  0.028970 -0.04619 9.632e-001
   AR(3; 1, 3) -0.064877  0.041201 -1.57464 1.158e-001
   AR(3; 2, 3)  0.024018  0.042614  0.56362 5.732e-001
   AR(3; 3, 3) -0.157419  0.044035 -3.57487 3.735e-004
       A(1, 1)  0.022257  0.007188  3.09660 2.032e-003
       A(2, 2)  0.029432  0.008899  3.30724 9.884e-004
       A(3, 3)  0.015534  0.004087  3.80113 1.561e-004
 ARCH(1; 1, 1)  0.182588  0.051191  3.56681 3.849e-004
 ARCH(1; 2, 2)  0.214782  0.061093  3.51567 4.657e-004
 ARCH(1; 3, 3)  0.252768  0.055259  4.57424 5.615e-006
GARCH(1; 1, 1)  0.491430  0.128102  3.83625 1.357e-004
GARCH(1; 2, 2)  0.407957  0.138514  2.94523 3.330e-003
GARCH(1; 3, 3)  0.464077  0.094584  4.90652 1.144e-006
```

```
Estimated Conditional Constant Correlation Matrix:
------------------------------------------------------------
       1.1     1.2     1.3
1.1 1.0000 0.13900 0.13215
1.2 0.1390 1.00000 0.04541
1.3 0.1321 0.04541 1.00000

Standard Errors:
        [,1]    [,2]    [,3]
[1,]      NA 0.03939 0.04364
[2,] 0.03939      NA 0.04270
[3,] 0.04364 0.04270      NA

------------------------------------------------------------


###############
#
#  JANUARY05
#
###############

Conditional Distribution:  t
 with estimated parameter 11.94504 and standard error 2.543177

------------------------------------------------------------

Estimated Coefficients:
------------------------------------------------------------
                   Value Std.Error  t value   Pr(>|t|)
   AR(1; 1, 1)  0.801032  0.019458  41.1680 0.000e+000
   AR(1; 2, 1)  0.066606  0.018235   3.6526 2.780e-004
   AR(1; 3, 1)  0.019265  0.015404   1.2506 2.115e-001
   AR(1; 1, 2)  0.076020  0.017845   4.2599 2.310e-005
   AR(1; 2, 2)  0.809481  0.019447  41.6249 0.000e+000
   AR(1; 3, 2) -0.020366  0.014368  -1.4174 1.568e-001
   AR(1; 1, 3)  0.133111  0.017793   7.4811 2.105e-013
   AR(1; 2, 3)  0.110551  0.018519   5.9697 3.696e-009
   AR(1; 3, 3)  0.972236  0.015176  64.0638 0.000e+000
       A(1, 1)  0.005020  0.001878   2.6728 7.689e-003
```

```
       A(2, 2)  0.009699  0.003777   2.5678 1.043e-002
       A(3, 3)  0.013499  0.005702   2.3672 1.818e-002
 ARCH(1; 1, 1)  0.161899  0.038536   4.2013 2.980e-005
 ARCH(1; 2, 2)  0.130353  0.037264   3.4981 4.967e-004
 ARCH(1; 3, 3)  0.099715  0.047529   2.0980 3.625e-002
GARCH(1; 1, 1)  0.776455  0.047300  16.4154 0.000e+000
GARCH(1; 2, 2)  0.744753  0.072035  10.3388 0.000e+000
GARCH(1; 3, 3)  0.581219  0.157736   3.6848 2.457e-004

Estimated Conditional Constant Correlation Matrix:
------------------------------------------------------------
        1.1     1.2     1.3
1.1 1.00000 0.02497 0.11986
1.2 0.02497 1.00000 0.05689
1.3 0.11986 0.05689 1.00000

Standard Errors:
        [,1]    [,2]    [,3]
[1,]      NA 0.04180 0.03922
[2,] 0.04180      NA 0.04366
[3,] 0.03922 0.04366      NA

------------------------------------------------------------


###############
#
#  JANUARY06
#
##############
Conditional Distribution:  t
 with estimated parameter 26.85611 and standard error 0.03861663

------------------------------------------------------------

Estimated Coefficients:
------------------------------------------------------------
                  Value Std.Error  t value   Pr(>|t|)
   AR(1; 1, 1)  0.739084  0.023000  32.1344 0.000e+000
   AR(1; 2, 1)  0.096804  0.024799   3.9036 1.035e-004
```

```
   AR(1; 3, 1) -0.010540  0.018169  -0.5801 5.620e-001
   AR(1; 1, 2)  0.119526  0.019140   6.2448 7.170e-010
   AR(1; 2, 2)  0.831742  0.020470  40.6328 0.000e+000
   AR(1; 3, 2)  0.007411  0.014709   0.5038 6.145e-001
   AR(1; 1, 3)  0.124918  0.015580   8.0176 4.219e-015
   AR(1; 2, 3)  0.038000  0.016103   2.3598 1.855e-002
   AR(1; 3, 3)  0.981417  0.012641  77.6393 0.000e+000
       A(1, 1)  0.038421  0.017846   2.1530 3.164e-002
       A(2, 2)  0.028190  0.009505   2.9657 3.117e-003
       A(3, 3)  0.008477  0.002620   3.2350 1.271e-003
 ARCH(1; 1, 1)  0.102828  0.052134   1.9724 4.894e-002
 ARCH(1; 2, 2)  0.161652  0.055048   2.9366 3.422e-003
 ARCH(1; 3, 3)  0.122475  0.040557   3.0198 2.617e-003
GARCH(1; 1, 1)  0.095691  0.374320   0.2556 7.983e-001
GARCH(1; 2, 2)  0.368885  0.176968   2.0845 3.746e-002
GARCH(1; 3, 3)  0.593769  0.102813   5.7753 1.133e-008

Estimated Conditional Constant Correlation Matrix:
------------------------------------------------------------
       1.1       1.2       1.3
1.1 1.0000  0.199055  0.169191
1.2 0.1991  1.000000 -0.002374
1.3 0.1692 -0.002374  1.000000

Standard Errors:
        [,1]    [,2]    [,3]
[1,]      NA 0.03639 0.03851
[2,] 0.03639      NA 0.04062
[3,] 0.03851 0.04062      NA

------------------------------------------------------------

#############
#
# JANUARY07
#
############

Conditional Distribution:  t
 with estimated parameter 8.220502 and standard error 1.26982
```

```
-----------------------------------------------------------------

Estimated Coefficients:
-----------------------------------------------------------------
                  Value Std.Error   t value   Pr(>|t|)
AR(1; 1, 1)  0.8108232  0.039752  20.39683 0.000e+000
AR(1; 2, 1)  0.2042492  0.037104   5.50477 5.122e-008
AR(1; 3, 1)  0.0961658  0.031667   3.03683 2.476e-003
AR(1; 1, 2)  0.1511063  0.029715   5.08526 4.671e-007
AR(1; 2, 2)  0.8017278  0.039459  20.31809 0.000e+000
AR(1; 3, 2)  0.0631722  0.027121   2.32924 2.012e-002
AR(1; 1, 3)  0.2362681  0.039037   6.05245 2.280e-009
AR(1; 2, 3)  0.1158365  0.036911   3.13827 1.768e-003
AR(1; 3, 3)  0.9467887  0.040812  23.19859 0.000e+000
AR(2; 1, 1) -0.0371993  0.050628  -0.73476 4.627e-001
AR(2; 2, 1) -0.0721002  0.050210  -1.43597 1.514e-001
AR(2; 3, 1) -0.0259358  0.040038  -0.64778 5.173e-001
AR(2; 1, 2) -0.0162113  0.035577  -0.45567 6.488e-001
AR(2; 2, 2)  0.0269136  0.053554   0.50255 6.154e-001
AR(2; 3, 2) -0.0133221  0.033996  -0.39187 6.953e-001
AR(2; 1, 3)  0.0424936  0.054122   0.78515 4.326e-001
AR(2; 2, 3)  0.0132476  0.051978   0.25487 7.989e-001
AR(2; 3, 3)  0.0550571  0.055800   0.98669 3.241e-001
AR(3; 1, 1)  0.0274102  0.035724   0.76728 4.432e-001
AR(3; 2, 1) -0.0643807  0.036010  -1.78785 7.421e-002
                  Value Std.Error t value   Pr(>|t|)
   AR(3; 3, 1) -0.02860  0.029945 -0.9550 3.399e-001
   AR(3; 1, 2) -0.10080  0.028590 -3.5256 4.490e-004
   AR(3; 2, 2)  0.01065  0.038254  0.2785 7.807e-001
   AR(3; 3, 2) -0.03130  0.024038 -1.3022 1.933e-001
   AR(3; 1, 3) -0.11308  0.039747 -2.8451 4.565e-003
   AR(3; 2, 3) -0.05291  0.041826 -1.2649 2.063e-001
   AR(3; 3, 3) -0.10391  0.040474 -2.5674 1.045e-002
       A(1, 1)  0.03137  0.012706  2.4693 1.376e-002
       A(2, 2)  0.01802  0.005659  3.1839 1.515e-003
       A(3, 3)  0.02726  0.006885  3.9600 8.229e-005
 ARCH(1; 1, 1)  0.14971  0.063258  2.3667 1.821e-002
 ARCH(1; 2, 2)  0.14875  0.044681  3.3291 9.150e-004
 ARCH(1; 3, 3)  0.22407  0.058222  3.8485 1.292e-004
```

```
GARCH(1; 1, 1)  0.39985  0.204684  1.9535 5.114e-002
GARCH(1; 2, 2)  0.63634  0.089244  7.1304 2.412e-012
GARCH(1; 3, 3)  0.30431  0.141213  2.1550 3.149e-002

Estimated Conditional Constant Correlation Matrix:
-------------------------------------------------------------
       1.1     1.2     1.3
1.1 1.0000 0.13632 0.16333
1.2 0.1363 1.00000 0.09073
1.3 0.1633 0.09073 1.00000

Standard Errors:
        [,1]    [,2]    [,3]
[1,]      NA 0.04205 0.04240
[2,] 0.04205      NA 0.04393
[3,] 0.04240 0.04393      NA

-------------------------------------------------------------
```

Bibliography

[1] `http://www.knmi.nl/samenw/hydra/index.html`

[2] `http://www.knmi.nl/samenw/hydra/documents/index.html`

[3] G.E.P. Box *Time series analysis-forecasting and control*

[4] Ch. Gourieroux *ARCH Models and Financial Applications*

[5] J. E. Payne *Further Evidence on Modeling Wind Speed and Time-Varying Turbulence*

[6] Bradley T.Ewing, Jamie Brown Kruse, John L. Schroeder *Time series analysis of wind speed with time-varying turbulence*

[7] R. F. Engle *Autoregressive Conditional Heteroskedasticity with Estimates of the Variance of United Kingdom Inflation*

[8] T.Bollerslev *Generalized Autoregressive Conditional Heteroskedasticity*

[9] Eric Zivot and Jiahui Wang *Modeling Financial Time Series with S-PLUS*

[10] J.D. Cryer, K.S.Chan *Time series analysis with application in R*

[11] P.J.Brockwell, R.A.Davis *Introduction to Time Series and Forecasting*

[12] G.Janacek, L.Swift *Time series, forecasting, simulation, applications*

[13] A.C.Harvey *Time Series Models*

[14] B. Brown, R. W. Katz, A.H. Murphy *Time Series models to simulate and forecast wind speed and wind power*

[15] B. Klöckl *Impacts of Energy Storage on Power Systems with Stochastic Generation*

[16] P. Schavemaker, L. van der Sluis *Electrical power system essentials*

[17] Daniel Pena, George C.Tiao, Ruey S. Tsay *A Course in Time Series Analysis*

[18] Karen C. Chou, Ross B. Corotis *Simulation of hourly wind speed and array wind power*

[19] H. Nfaoui, J.Buret, A.A.M. Sayigh *Stochastic simulation of hourly average wind speed sequences in tangries (Morocco)*

[20] P.Poggi, M.Museli, G. Notton, C.Cristofaru, A.Louche *Forecasting and simulating wind speed in Corsica by using an autoregressive model*

[21] S.Bivona, R.Burlon, C.Leone *Hourly wind speed analysis in Sicily*

[22] Lalarukh Kamal, Yasmin Zahra Jafri *Time series models to simulate and forecast hourly averaged wind speed in Queta, Pakistan*

[23] A.R.Daniel, A.A.Chen *Stochastic simulation and forecasting of hourly average wind speed sequences in Jamaica*

[24] J.Ch.Bouette, J.F.Chassagneux, D. Sibai, R. Terron, A. Charpentier *Wind in Ireland: long memory or seasonal effect*

[25] S.Hussain, A.Elbergali, A.Al-Masri, G.Shukur *Parsimonious modelling, testing and forecasting of long-range dependence in wind speed*

[26] S.C.Pryor, R.J. Barthelmie *Statistical analysis of flow characteristics in the coastal zone*

[27] A.B.Sigl, R.B.Corotis, D.J.Won *Run Duration of Surface Wind Speeds for Wind Energy Application*

[28] S.C. Pryor, R.J.Barthelmie *Comparison of Potenatial Power Production at On-and Offshore Sites*

[29] K.Kocak *Practical ways of evaluating wind speed persistence*

[30] K.Kocak *A method for determination of wind speed persistence and its application*

[31] C.Tsekos, K.Anastasiou *Persistence of marine environmental parameters from Markov theory*

[32] T.Bollerslev *Modelling the Coherence in Short-Run Nominal Exchange Rates*

[33] L.Bauwens, S.Laurent, J.V.K.Rombouts *Multivariate GARCH models*

[34] A.Silvennoinen, T.Teräsvirta *Multivariate GARCH models*

[35] H. Lüthepohl *New Introduction to Multiple Time Series Analysis*

[36] R. F.Engle V. Ng *Measuring and Testing the Impact of News on Volatility*

[37] P. Norgaard, H.Holttinen *A Multi-Turbine power curve approach*