Technical University of Crete

School of Mineral Resources Engineering

Geostatistical Analysis of Installed Wind Power Production Data

Author: Panagiota Gkafa

Advisor: Prof. Dionissios Hristopoulos, Technical University of Crete

Steering Committee Members:

Prof. Michalis Galetakis, Technical University of Crete

Prof. Nikolaos Thomaidis, Aristotle University of Thessaloniki

June 6, 2020

I would like to dedicate to my lovely family!

Acknowledgments

Firstly, I would like to thank my advisor Prof. Dionissios Hristopoulos for showing confidence in me by assigning me this master's thesis. Also I would like to thank him for his patience, caring, motivation, and guidance. I would also like to thank the other two members of my thesis committee: Prof. Nikolaos Thomaidis, for sharing with me the data of the installed power production from the Netherlands, and his comments and his guidance, and Prof. Michail Galetakis for his comments and his guidance.

I would also like to thank the members, and my friends in the Geostatistical Laboratory of the School of Mineral Resources Engineering at the Technical University of Crete, Ms. Vicky Agou, Ms. Michaela Vasiliadi, and Dr. Andreas Pavlides for their support and their friendship. I would especially like to thank Vicky and Andreas for many helpful conversations during my studies.

I would like to express my gratitude to all my friends, for their support and encouragement during all these years.

Last but no least, I would like to thank my family. My father, my mother, and my brother have always been by my side all these years with their support, and their encouragement.

Abstract

In recent years, an increasing number of countries are attempting to reduce their reliance on fossil fuels and enhance the contribution of renewable energy sources in their energy production plans. Renewable energy sources include wind, sun, geothermal sources and tidal energy. Wind is the most common renewable energy source, both for domestic and industrial use. Hence, the prediction of wind speed and aeolian energy potential is an important topic of research.

This thesis focuses on the investigation of the variability of aeolian energy production in the Netherlands. Spatial and temporal models for aeolian energy are defined and estimated using geostatistical and time-series forecasting methods respectively. The available data are average daily measurements of aeolian power produced by 46 stations distributed across the Netherlands. The data are recorded during the six-year time time period from 2001 until 2006. Most of the available studies in the literature analyse wind speed data. In this approach, the wind speed is first predicted at unmeasured points in space or time. Then, the respective aeolian power is estimated using a standard "power curve", which relates the wind speed to power production. In contrast, the models investigated herein (both the geostatistical models for spatial prediction and the time series models for forecasting) are directly based on data of aeolian power production.

Wind speed typically depends on altitude. However, in the spatial model used herein a topographic trend is not necessary, due to the flat topography of the Netherlands. In order to investigate the spatial variability of aeolian power production, the empirical variogram is calculated from the annual mean installed power production. Then, the empirical variogram is fitted to three theoretical models (Gaussian, exponential, and spherical). The spherical variogram is selected as the optimal model because it produces the minimum sum of weighted squared errors. Ordinary kriging is then applied to the aeolian power production data, in order to generate an interpolated

map of aeolian power potential over the entire country and a respective variance map for each year studied. To validate the performance of the spatial model, the method of leave-one-out cross-validation is used. The spatial model performs well, as evidenced by the high values of Pearson's correlation coefficient (85%) between the data and the predictions. The kriging-generated map gives a visual representation of aeolian power potential and its uncertainty over the Netherlands. The highest wind power predictions are in the West area of Netherlands (near the North Sea), while the lowest power estimates are in the Eastern part of the country. In addition, the uncertainty of the predictions is lower in the West and higher in the East. These spatial patterns are consistently observed for all the years (2001–2006) in the study.

In the temporal analysis we focus on the time series of average monthly wind power production at each station. The methodology is illustrated for two stations, one onshore and one in the North Sea, off the Netherlands' coast. The temporal variation of wind power production exhibits seasonal behavior with an annual cycle. We follow two different modeling approaches: In the first approach, we fit an explicit periodic function to the data and then apply a SARIMA time series model to the stochastic residuals. In the second approach, a SARIMA model is directly fitted to the average monthly wind power data. The optimal parameters are used to predict wind power production for the following 12 months. Thus, the prediction involves the monthly average power production for the year 2007. To validate the performance of the models, cross-validation using the method of one-step-ahead forecast is used. The temporal models show good performance with respect to the root mean square error (RMSE)—the RMSE is in the range 0.04–0.22 MW (about 21%–43% of the average monthly wind power) at each station.

Περίληψη

Τα τελευταία χρόνια πολλές χώρες προσπαθούν να μειώσουν την εξάρτησή τους από τα ορυκτά καύσιμα και να ενισχύσουν την συμβολή των ανανεώσιμων πηγών στην παραγωγή ενέργειας. Οι ανανεώσιμες πηγές ενέργειας περιλαμβάνουν τον άνεμο, τον ήλιο, τις γεωθερμικές πηγές και την κυματική ενέργεια. Ο άνεμος είναι η πλέον συνήθης πηγή ανανεώσιμης ενέργειας, τόσο για οικιακή όσο και για βιομηχανική χρήση. Ως εκ τούτου, η ανάλυση της μεταβλητότητας και η πρόβλεψη της ταχύτητας του ανέμου καθώς και της δυνητικής παραγωγής ισχύος είναι σημαντικά ερευνητικά θέματα.

Η συγκεκριμένη μεταπτυχιακή εργασία διερευνά την μεταβλητότητα της παραγόμενης αιολικής ενέργειας στην Ολλανδία. Για την εκτίμηση των χωρικών και των χρονικών μοντέλων χρησιμοποιούνται γεωστατιστικές μέθοδοι και μέθοδοι χρονοσειρών αντίστοιχα. Τα διαθέσιμα δεδομένα είναι οι μέσες ημερήσιες μετρήσεις της παραγωγής ενέργειας από 46 σταθμούς στην Ολλανδία. Τα δεδομένα καταγράφονται κατά την εξαετή χρονική περίοδο από το 2001 έως το 2006. Οι περισσότερες διαθέσιμες έρευνες στη βιβλιογραφία αναλύουν δεδομένα που αφορούν την ταχύτητα του ανέμου. Σε αυτήν την περίπτωση, γίνεται εκτίμηση αρχικά της ταχύτητας του ανέμου στον χώρο ή στον χρόνο. Στη συνέχεια εκτιμάται η αντίστοιχη αιολική ενέργεια, χρησιμοποιώντας μία τυπική 'καμπύλη ενέργειας', η οποία συσχετίζει την ταχύτητα του ανέμου με την παραγόμενη ισχύ. Σε αντίθεση, τα μοντέλα που ερευνήθηκαν στη συγκεκριμένη εργασία (τόσο τα γεωστατιστικά μοντέλα για την χωρική εκτίμηση, όσο και τα χρονικά μοντέλα για την πρόβλεψη στο χρόνο) βασίζονται άμεσα σε δεδομένα παραγόμενης ισχύος.

Η ταχύτητα του ανέμου συνήθως εξαρτάται από το υψόμετρο. Ωστόσο, λόγω της επίπεδης τοπογραφίας της Ολλανδίας, δεν είναι απαραίτητο να ληφθεί υπόψη κάποια τοπογραφική τάση στο χωρικό μοντέλο. Προκειμένου να διερευνηθεί η χωρική μεταβλητότητα της αιολικής ισχύος, υπολογίζεται το εμπειρικό βαριόγραμμα παραγόμενης ισχύος από τα ενεργειακά δεδομένα. Στη συνέχεια το εμπειρικό βαριόγραμμα προσαρμόζεται σε τρία θεωρητικά μοντέλα (Γκαουσσιανό, Εκθετικό και Σφαιρικό). Το σφαιρικό μοντέλο βαριογραμμάτος επιλέγεται ως το βέλτιστο, βάσει του ελάχιστου αθροίσματος των σταθμισμένων τετραγωνικών σφαλμάτων. Στη συνέχεια το Κανονικό Kriging εφαρμόζεται στα δεδομένα με σκοπό τη δημιουργία των χαρτών παρεμβολής για το αιολικό ενεργειακό δυναμικό σε όλη την έκταση της χώρας, και την κατασκευή των αντίστοιχω χαρτών αβεβαιότητας. Για να εξεταστεί η απόδοση του χωρικού μοντέλου, χρησιμοποιήθηκε η μέθοδος της διασταυρωτικής επιβεβαίωσης και συγκεκριμένα της αφαίρεσης ενός σημείου εκ περιτροπής (leave-one-out cross-validation). Σύμφωνα με

το συντελεστή συσχέτισης του Pearson, ο οποίος είναι ίσος με 85% τα χωρικά μοντέλα παρουσιάζουν μια αρκετά καλή απόδοση. Οι παραγόμενοι χάρτες βάσει του kriging, απεικονίζουν το αιολικό ενεργειακό δυναμικό και την αβεβαιότητά του. Οι υψηλότερες τιμές της εκτιμώμενης αιολικής ισχύος παρατηρούνται στη Δυτική περιοχή της χώρας (δίπλα στη Βόρεια Θάλασσα (North Sea», ενώ οι χαμηλότερες στην Ανατολή. Σε αντίθεση, η αβεβαιότητα είναι χαμηλή στη Δύση και υψηλή στην Ανατολή. Το ίδιο σταθερό μοτίβο παρατηρείται για όλα τα χρόνια της μελέτης (2001–2006).

Στη χρονική ανάλυση, χρησιμοποιούμε τη μέση μηνιαία ισχύ ανά σταθμό. Η εφαρμογή της μεθοδολογίας παρουσιάζεται για έναν χερσαίο και έναν υπεράκτιο σταθμό στη Βόρεια Θάλασσα. Τα δεδομένα εμφανίζουν μία εποχικότητα με ετήσιο κύκλο. Για την μοντελοποίηση των δεδομένων, χρησιμοποιήθηκαν δύο διαφορετικές προσεγγίσεις. Στην πρώτη προσέγγιση, προσαρμόζουμε ενα αιτιοκρατικό περιοδικό μοντέλο, και στη συνέχεια εφαρμόζουμε ένα μοντέλο SARIMA στα στοχαστικά υπόλοιπα. Στη δεύτερη προσέγγιση, εφαρμόζουμε τα μοντέλα SARIMA απευθείας στους μηνιαίους μέσους όρους της αιολικής ισχύος. Στη συνέχεια, γίνεται εκτίμηση των παραμέτρων του μοντέλου βάσει των υπαρχόντων χρονοσειρών. Οι βέλτιστες παράμετροι, χρησιμοποιούνται για να γίνει πρόβλεψη τους επόμενους 12 μήνες. Έτσι η πρόβλεψη αποτελείται από μηνιαίους μέσους όρους αιολικής ισχύος για το έτος 2007. Για να εξεταστεί η αποδοτικότητα του μοντέλου, χρησιμοποιούμε τη μέθοδο της διασταυρωτικής επιβεβαίωσης βασισμένη στην πρόβλεψη της επόμενης χρονικής στιγμής (one-step-ahead forecast). Σύμφωνα με τη ρίζα του μέσου τετραγωνικού σφάλματος (RMSE), (το RMSE έχει εύρος 0.04–0.22 MW, δηλαδή ανέρχεται στο 21%–43% της μέσης τιμής των μηνιαίων μέσων όρων σε κάθε σταθμό), τα μοντέλα SARIMA παρουσιάζουν σχετικά καλή απόδοση.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In the last decades, geostatistics has many applications in the environmental sciences. Hence, it is important to test the spatial and temporal properties of them. Geostatistics is a solution to these problems because it develops appropriate models that can produce accurate spatial and temporal predictions, and estimate the uncertainty of the results. The common case in climate studies is the prediction of the installed power capacity of renewable energy sources. Specifically, the spatial and temporal analysis in different time scales is a problem for many studies.

This thesis is motivated by the need for models that can accurately capture the spatiotemporal variability of wind power. As such, spatial methods are used to map the mean annual power capacity from offshore and onshore wind turbines. Ordinary kriging is used for the predictions. For forecasting of the future annual mean power production was used.

1.1 Renewable Energy

During the recent decades more and more countries have been using renewable energy, both for domestic and industrial use. This kind of energy covers forms such as wind, sun, geothermal, tidal [48]. According to recent research, renewable energy accounts for 24.5% of global electricity production.

The global investment in renewable technologies amounts to 286 billion dollars in 2015 [53]. Also, there are 7.7 million employees working in renewable energy. The renewable energy production systems constantly become more productive and cheaper, and total energy consumption is increasing. Since 2019, more than 2/3 of installed

electricity capacity is renewable energy production facilities.

The rapid increase of renewable energy technologies, contributes to decrease of environmental pollution [58], and financial growth. In 30 countries, renewable energy contributes to 20% of energy supply. Hence, the use of renewable energy is expected to increase in the following years [54]. Renewable sources can be used in many areas across the world, in contrast with fossil fuels, whose reserves are limited and strategically located.

The most important renewable energy resources are wind, solar, geothermal, hydropower, and bio-energy/bio-mass [48].

- **Wind Energy** becomes from wind turbines, which convert the wind's kinetic energy to electric. The ideal areas to install wind farms are offshore areas, which have constant and strong winds, and high elevation.

- **Solar Energy** becomes through the transformation of solar energy to electric, either directly from photo-voltaic panels, or through solar panels which collect sun's rays to achieve high temperatures and utilize these temperatures for energy production.

- **Geothermal energy** becomes from energy which is stored inside the earth. The geothermal gradient (the difference in temperature based on depth inside the crust), is utilized for continuous conduction of thermal energy in the form of heat from the depths of the earth to the surface.

- **Hydropower Energy** has two forms a) wave energy and b)tidal energy. Wave energy exploits the kinetic energy of sea or ocean surface waves. Tidal energy, exploits the energy of tidal waves (the interaction of moon's gravity on the sea level). As the water moves, it is forced to pass through a turbine producing energy.

- **Bio-Energy** is the energy that comes from alive or recently alive organisms. It often refers to all plants, but wood remains the largest source of bio-energy. To produce Bioenergy, biomass is transformed to bio-fuel through thermal or chemical procedures.

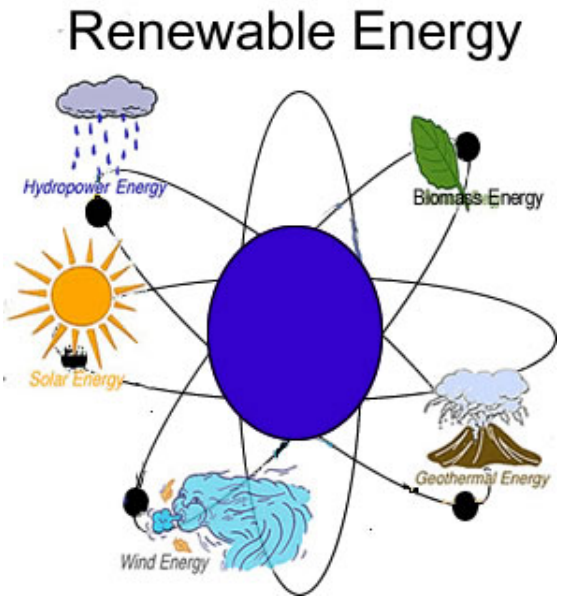


Figure 1.1: Schematic representation of different forms of renewable energy sources. The schematic in this figure is taken from [9].

1.2 Wind Energy

Wind energy is generated from wind turbines, in wind farms. These turbines convert wind kinetic energy to electricity. A group of wind turbines in the same location is considered a wind farm. Large wind farms may consist of hundreds of turbines, covering a large area of hundreds squares miles, using the intermediate land for agricultural or other purposes. If this wind farm is inland, the farm is called onshore wind farm. In contrast, the wind farms which are located in the sea are named as offshore wind farms.

Onshore wind farms may not have any effects in environment, but they affect the landscape of area where they are installed due to the large space they occupy. The largest offshore and onshore wind farms in the world are shown in Tables (1.1) and (1.2) [63] .

The wind conditions of the area, easy access to electricity network for transportation of electricity and the local prices are some of the conditions that need to be considered to create a wind farm. Wind speed is a parameter which is related to power production. This means that the higher the average wind speed, the higher the energy generated by the wind turbine. Hence, stronger winds have economic benefits for the wind farm development. To avoid damages from strong winds and high turbulence, more durable wind turbines need to be used. However, the mean power

Table 1.1: Capacity (MW) for ten of the largest *offshore* wind farms globally. The table is taken from [63].

Wind Farm	**Country**	**Capacity (MW)**
Wanley Extension	United Kingdom	659
London Array	United Kingdom	630
Gemini Wind Farm	Netherlands	600
Gode Wind (phases 1 and 2)	Germany	582
Gwynt y Mor	United Kingdom	576
Race Bank	United Kingdom	573
Greater Cabbard	United Kingdom	504
Dudgeon	United Kingdom	402
Veja Mate	Germany	402

Table 1.2: Capacity (MW) for ten of the largest *onshore* wind farms in the world. The table is taken from [63].

Wind Farms	**Country**	**Capacity (MW)**
Gansu Wind Farm	China	7965
Alta Wind Energy Center	United States	1548
Muppandal Wind Farm	India	1500
Jaisalmer Wind Park	India	1064
Los Vientos Wind Farm	United States	912
Shepherds Flat Wind Farm	United States	845
Meadow Lake Wind Farm	United States	801
Roscoe Wind Farm	United States	781.5
Horse Hollow Wind Energy Center	United States	735.5

is not analogous to mean wind speed. As such, ideal wind conditions are strong and constant winds with low turbulence and same direction. Hence the ideal land to install a wind farm is the mountain passes, which work like a channel directing the wind.

1.2.1 Wind Energy Capacity

The wind power production was at 100 GWatt in the European Union in 2012 while the respective figure in the USA was 75 Gwatt in 2015. In 2018, the global wind power capacity increased by 51Gwatt to 591 GWatt. In several countries, the installed power production has been in high levels. In 2018, Denmark had the highest

wind energy production at 41,4%, followed by 28% in Ireland, 24% in Portugal, 21% in Germany, and 19% in Spain [33]. In the Netherlands, the installed power capacity was at 4.341MW at the end of 2017. Figure 1.2 shows the annual growth of wind power capacity globally.

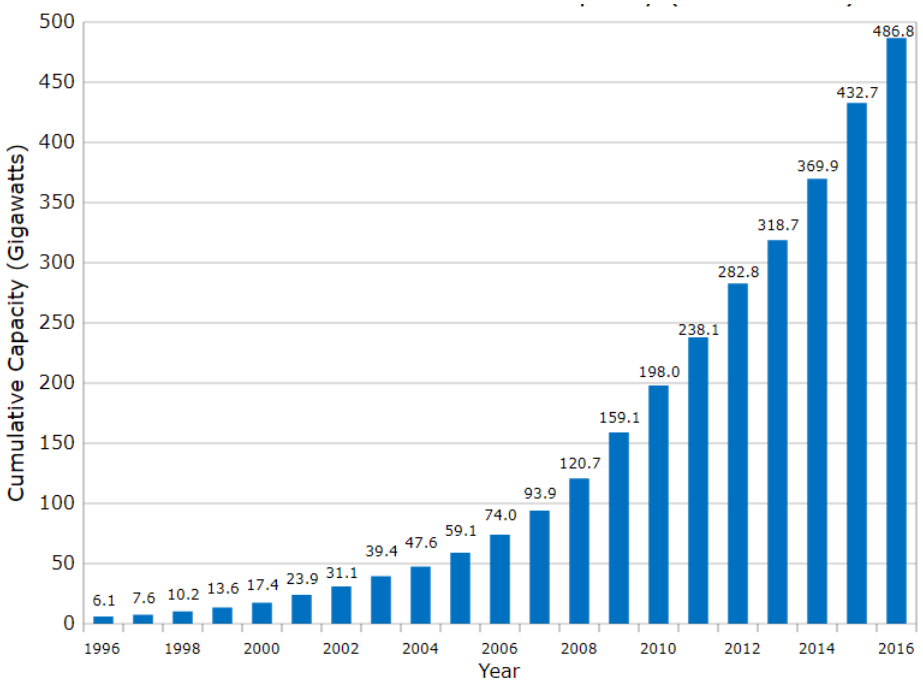


Figure 1.2: Evolution of the global annual wind power cumulative capacity (GW) for the period 1996–2018. The figure is from [64].

1.2.2 Wind Power Forecasting

The prediction of wind power and speed is necessary, due to the development of wind farms, and the increase of global consumption of wind energy. Spatial, temporal, and spatio-temporal models have been created to predict these variables. The prediction could be implemented either with physical or statistical methods. Numerical Weather Prediction (NWP), is a physical method, which combines mathematical and physics equations to predict the wind variables (power and speed) [35, 60]. The main feature of this method is the short-term forecast in large-areas. The needed meteorological variables for this method are wind speed, temperature, direction, humidity, and pressure. Models of NWP are the following [60]:

- Uk Meteorological Office mesoscale (MESO) model.
- Danish Meteorological HIRLAM model.

- Regional Atmospheric Modeling System (RAMS) model.
- Weather Research and Forecasting Model (WRF).

In statistical methods, there are two types for spatial interpolation: deterministic and stochastic or geostatistical methods. Interpolation methods (stochastic or deterministic) create surfaces which fill gaps using the whole data set or neighborhoods in data set. Deterministic methods use models that produce the same output from a given starting condition. Some of these models are inverse distance weighting, polynomial regression, triangular irregular network, nearest neighbor. Stochastic methods estimate the correlation between the measurements and the spatial structure of data. Some of the best known stochastic methods are the kriging family of methods (universal, simple, ordinary, and co-kriging). Other stochastic methods include the Stochastic Local Interaction models (SLI) [28]. Stochastic methods for time series forecasting include methods that create models (AR, MA, ARMA, etc) to predict the next time scale, and artificial neural networks (ANN) (radial basis function, Multilayer Perception, Feed-forward) [39].

Related studies have been presented by Amanda Lenzi et al [42] which produce a spatial model to estimate power production at two different time scales. Also, Alexiadis M.C. et al [5] presents a technique to forecast the wind speed and power, based on cross-correlation at neighboring sites. Kariniotakis et al [39] use ANN modeling to forecast the wind power of a wind farm study area.

In modeling wind power generation data there are some factors that should be considered. Hourly time series at multiple heights with relevant variables, corrected for site characteristics, can be used as long term references (reanalysis, meteorological stations). Land cover classification maps for the selected area and period at high resolution (i.e. 10m) should be consulted when modeling wind data. Other useful factors are constraints (exclusion zones, grid connection, roads etc), elevation, and topography (especially roughness).

Probabilistic Forecasting

Wind energy forecasts can be made with deterministic methods or stochastic interpolation methods. To make such forecasts, temporal analysis of time series is used often in combination with spatial analysis. Probabilistic forecasting is the recommended approach for wind energy forecasting. Also using a probabilistic forecast model, the energy management would improve resulting in a decrease of fossil fuel

dependency [37].The power output is affected by several environmental factors, such as wind direction, wind speed, air density, humidity, turbulence intensity, and wind shears. Several methods use the power curve, which relates wind power to wind speed [22].

Power Curve

The power curve is a way to relate the power output with all the factors mentioned in the previous section. However, most of power curve models relate only the wind speed (and sometimes with direction) with wind power. The wind power industry uses the power curve for several reasons. The main purpose is to forecast the wind power in two steps: firstly the wind speeds are forecast and then through the power curve the wind power is estimated. Another purpose of the power curve is for wind turbine performance assessment and another one health monitoring. The behavior of the power curve could be different when the wind speed changes [23].

Each wind turbine has a unique power performance curve because the output power of a wind turbine varies even with the same wind speed. The power curve includes a)"cut-in-speed", in which turbine blades begin to rotate, b) "a rated speed", which is the lowest speed at which the maximum power output of turbine is generated, and c) "cut-out-speed", in which the turbine is shut down to prevent damage. The output power is captured by the power curve as a function of the hub height wind speed, and is defined as:

$$P = \frac{1}{2}\rho\pi R^2 C_{_P} u^3, \tag{1.1}$$

where ρ is the air density, R is the radius of the rotor, C_P is the power coefficient (is the percentage of power captured by the turbine), and the u is the wind speed [43]. The power curve is depended from the wind speed and the power coefficient, as the air density remains constant at hub height. The power coefficient depends on the tip speed ratio (λ) and the blade-pitch angle (β). The tip-speed ratio λ for wind turbines is the ratio between the actual speed of the wind u, and the tangential speed of the tip of a blade [55]. A typical power curve is presented at the following figure:

For their study, Jooyoung Leon and James W. Taylor [37] use a bivariate vector autoregressive moving average-generalized autoregressive conditional heteroscedastic

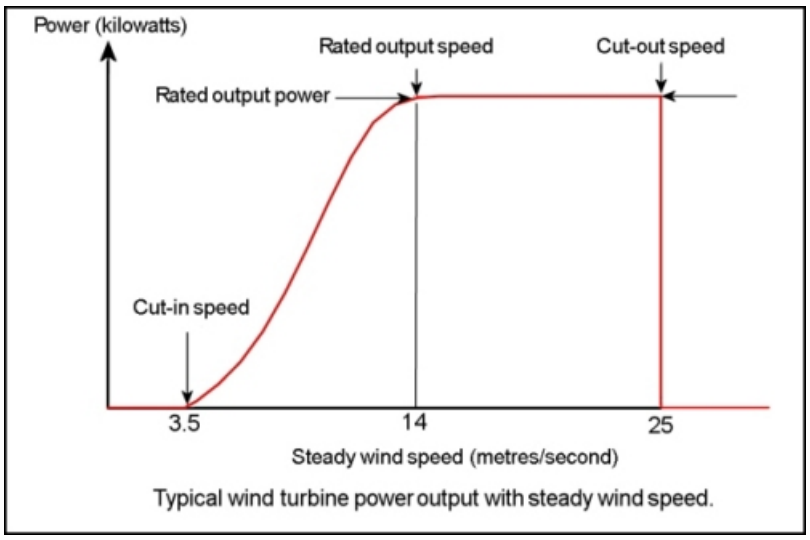


Figure 1.3: A typical wind turbine power curve. The x axis is the steady wind speed in m/s and in y axis is the output power in Kw. (Figure from [1])

(VARMA-GARCH) model, which improves wind speed prediction through the joint modeling of wind speed and direction. The investigators convert these predictions to wind power density predictions. Towards that purpose, Monte Carlo simulations are used, along with conditional kernel density estimation. The suggested method was able to outperform simpler models based on the deterministic power curve as well as simple benchmark methods. Also, Giwhyun Lee et al [22], provide an additive multivariate kernel (AMK) method to model the power curve with the possibility to include more environmental factors to create a new power curve model. This model can describe the nonlinear relationships between the power output and the multitude of environmental factors, and the high interaction effects. Their model performs better (based on their validation measures) than other methods such as Bayesian additive regression trees and smoothing spline analysis of variance. As such, it would give a better estimation of power production. Also, the AMK model is faster than the other methods.

1.3 Wind Resource and Energy Yield Assessment

An important part in wind data analysis is the wind resource assessment (WRA). The WRA is useful to map the wind resources and to define the financial feasibility for a wind project in order to be acceptable by banks and investors[10]. Nevertheless, some energy markets do not follow the requirements of the international standards (as IEC 61400–12–1)[6]. These standards refer to a credible estimate of losses and

uncertainties, auditable data acquisition, processing, and archiving, rigorous modelling of flow based on a linear model for simple terrain and a computational fluid dynamics model for complex terrain, and a on-site wind measurement with tall towers and high-quality machines for at least one year. If a project does not conform to international standards, it could lead to higher risks and uncertainties. Thus, the investors often request a higher share of equity capital and higher interest rates [6].

The average annual energy production (AEP), represents the output of a bankable WRA. In any kind of financial analysis for the project both AEP and WRA are used. Uncertainty analysis is the most important part of the analysis. It is important to estimate the uncertainty of different levels of confidence. Certain levels of confidence (notated as P50, P90) have been identified as the most important for the financial evaluation of the project and risk assessment [7].

Specifically the P50 level confidence is used as reference for the annual average production. P50 indicates that the probability of predicted value to be overestimated or underestimated is 50% on long term. So P50 is also called as AEP . P90 is the energy production that will be generated at 90% probability [6]. There is a 90% probability to generate P90 electrical energy or more in a given year, and only 10% to generate less [10]. A probability of 10% is an acceptable risk for the investors and the banks. These probabilities are estimated from the dataset's (or the simulated data) cumulative distribution function (CDF). The difference between the P90 and P50 level of energy production is affected by the level of uncertainty[7].

In order to attract financing, the project must have low uncertainties. Thus, to minimize uncertainties due to measurement, modelling, and other factors, it is important to make a through investigation during the study. The AEP, which is the primary output of the WRA, influences directly the revenue of the project. Hence, the revenue is low when the AEP is low. The second most important output is the uncertainty associated with AEP, which directly influences the risk of the project [10].

1.4 Probabilistic Analysis

It is easy to observe that the wind speed has high variability. To determine the annual production of power from a turbine, it is necessary to know the long-term mean wind speed because power is a non-linear function of wind speed. The most

commonly used probability distribution function for wind data is the Weibull distribution.

1.4.1 General Form of Weibull Distribution

The general form the probability density function of Weibull distribution is written as:

$$f(x;\lambda,k) = \begin{cases} \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k} & x \geq 0 \\ 0 & x < 0. \end{cases} \tag{1.2}$$

The parameter k is the shape parameter and the λ is the scale parameter. The Weibull distribution interpolates between the exponential distribution if k=1, and the Rayleigh distribution if k=2.

The cumulative distribution function for the Weibull distribution is written as:

$$F(x;k,\lambda) = 1 - e^{-(x/\lambda)^k},\ x \geq 0. \tag{1.3}$$

Figure (1.4) presents the probability function and the cumulative density function for different values of shape and scale for the Weibull distribution [34].

Moments of the Weibull Distribution

Mean Value

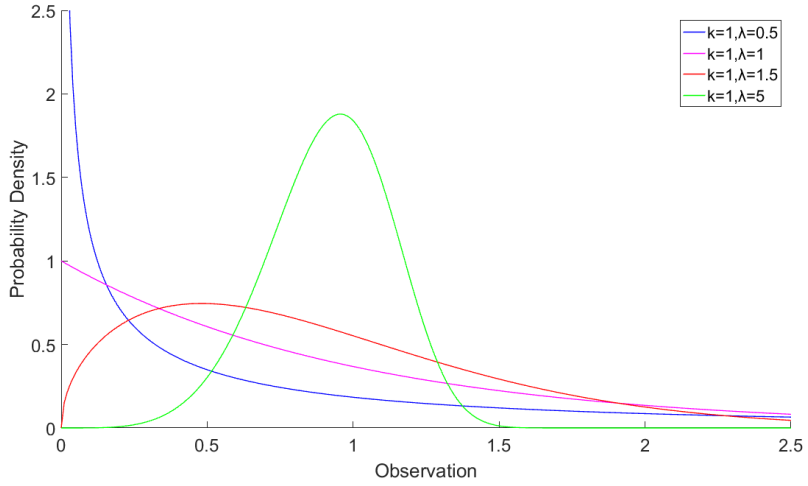
$$\mu = \lambda\Gamma\left(1+\frac{1}{k}\right). \tag{1.4}$$

Variance

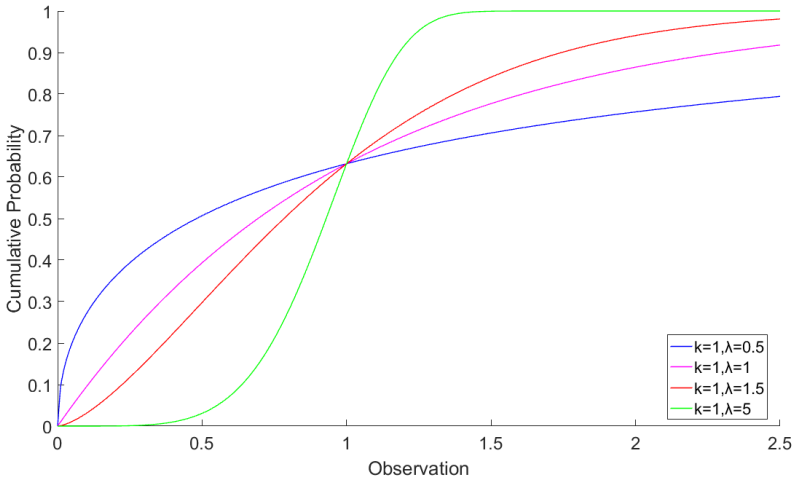
$$\sigma^2 = \lambda^2\left[\Gamma\left(1+\frac{2}{k}\right) - \left(\Gamma\left(1+\frac{1}{k}\right)\right)^2\right]. \tag{1.5}$$

Skewness

$$\frac{\Gamma(1+3/k)\lambda^3 - 3\mu\sigma^2 - \mu^3}{\sigma^3}. \tag{1.6}$$



(a)



(b)

Figure 1.4: Probability density function (Figure 1.4a) and cumulative distribution function (Figure 1.4b) of the Weibull distribution for different values of the shape k and scale λ parameters.

1.4.2 Weibull Distribution for Wind Data

To estimate the mean power from a wind turbine over a range of (mean) wind speeds, the following non-dimensional form of the probability density for the wind data is used [57]:

$$p\left(\frac{u}{U}\right) = k\Gamma\left(1+\frac{1}{k}\right)\left\{\frac{u}{U}\Gamma\left(1+\frac{1}{k}\right)\right\}^{k-1} e^{-\left\{\frac{u}{U}\Gamma\left(1+k^{-1}\right)\right\}^{k}}, \tag{1.7}$$

where u is the fluctuating wind speed component, U is the mean value of the wind speed, $p(\frac{u}{U})$ is the non-dimensional probability density function (pdf), k is

the Weibull shape factor, and $\Gamma\left(1+k^{-1}\right)$ denotes the gamma function with argument $1+k^{-1}$.

The standard deviation of the wind speed which corresponds to the above pdf is given by:

$$\sigma = U\sqrt{\frac{\Gamma\left(1+\frac{2}{k}\right)}{\left[\Gamma\left(1+\frac{1}{k}\right)\right]^2}-1}. \tag{1.8}$$

Weibull parameters are estimating using several estimation methods, such as maximum likelihood. Camilo Carillo et.al, present a different method to estimate the Weibull parameters for wind energy analysis [12].

Chapter 2

Geostatistical Methods

Earth science data are distributed over space and time. The analysis and prediction of spatial properties, such as porosity, pollutant concentrations, is carried out using geostatistical methods [16, 24]. Geostatistics includes methods that can be used to characterize spatial properties based on the theory of random fields. Random fields are considered a good numerical framework for spatial data analysis similar to how time series analysis is used for temporal data. The number of variables required to represent a spatial or a temporal process is infinite, even for areas of finite size, due to the constant variation of geographical location [47].

Geostatistical methods are applied to studies such as meteorology [2], topographic analysis, prospecting [38], mapping and mapping of pollutant concentrations in various environmental media (air, subsurface, surface aquifers) [24], and coal mining [50]. Time series are applied to studies such as economics[2], meteorology(precipitation and wind) [2, 30], in order to forecast the variable which shows the progress of a process.

2.1 Randomness

Randomness characterizes phenomena for which the value cannot be known with absolute precision. The reasons that may contribute to this are either intrinsic (strong spatial and temporal variability of the phenomenon), or come from the experimental process (random errors, limited resolution), or from environmental changes (variations in temperature and humidity) [27].

2.2 Random Variable-Random Field

A random variable X can take values from a set of possible values. A random variable is called discrete if it takes x_i values, where $i = 1, \ldots, N$, in an integer set, i.e. the frequency of occurrence of each value is determined by a probability function. A random variable is constant when it takes values from a continuous set. The probability the variable X could take values from a very small interval around x is determined by the probability density function [49, 17]. The expected value $\mathbb{E}[X]$ of a random variable X is the average of the random variable for all states.

If the probability distribution of X follows a probability density function (pdf) $f(x)$, then the expected value is

$$\mathbb{E}[X] = \int_{-\infty}^{+\infty} x\ f(x)\ dx. \tag{2.1}$$

A stochastic process is a collection of random variables, which represent the evolution of a system of random values through the time. In such a process there are many directions in which this process can evolve. A random field $X(\mathbf{s})$ is a collection of random values that are distributed in space with the vector $\mathbf{s}$ corresponding to the location of each point in the study area. When the random field is discrete, then it consists of a list of random numbers where their pointers are mapped to an n -dimensional space.

If the random variables are distributed in space, then the mathematical properties from which the random variables are described are extended. A random field consists of a set of random variables that describe spatial change in at least one of its mathematical properties. Thus a random field can be considered as a multidimensional random variable. Random fields have unique mathematical properties, which sets them apart from a set of independent random variables, because of the interdependence of physical sizes at different locations in space[27, 62].

Fluctuations are the stochastic component around the field's expectation, and is defined by:

$$X'(\mathbf{s}) = X(\mathbf{s}) - \mathbb{E}[X(\mathbf{s})]. \tag{2.2}$$

2.3 Probability Density Function

The probability density function (pdf) of a random field is indicated $f_X[x(\mathbf{s})]$, where the integer denotes the field , and the function argument is the values of state $x(\mathbf{s})$. For a single random variable, the pdf f_X represents a point. In contrast the pdf of a random field contains values throughout the space, where the field is defined. This means that the joint pdf for any number of points in the field is described by the pdf f_X. Therefore the pdf of a random field contains more information than the pdf of a single variable.

The one-dimensional pdf of a field at point $\mathbf{s_1}$ is defined as $f_X(x_1; \mathbf{s_1})$ and expresses the possible states of a field at the point s_1. Accordingly, the two-dimensional pdf of the field is defined as $f_X(x_1, x_2; \mathbf{s_1}, \mathbf{s_2})$, and expresses the interdependence of possible states in two points. Similarly defined is the multidimensional pdf $f_X(x_1, \ldots, x_N; \mathbf{s}_1, \ldots, \mathbf{s}_N)$, which describes the interdependence of possible states for a set of N points.

2.4 Statistical Homogeneity

Some assumptions that limit the properties of a random field can lead to a more effective geostatistical analysis. The most widely used simplistic assumption is statistical homogeneity, which is an extension of the classical definition of homogeneity. A property is homogeneous if the corresponding variable has a constant value in space. Therefore, one random field $X(\mathbf{s})$ is statistical homogeneous if the mean is constant, $m_{_X}(\mathbf{s}) = m_{_X}$, the coefficient function is solely defined and dependent on the vector of the distance $\mathbf{r} = \mathbf{s}_1 - \mathbf{s}_2$ between the two points and not from their position, $c_{_X}(\mathbf{s}_1, \mathbf{s}_2) = c_{_X}(\mathbf{r})$, and the variance of a statistically homogeneous field is constant.

The above conditions define statistical homogeneity in the weak sense. A random field described by strong statistically homogeneous, when the multidimensional Probability Density Function for N points (where N is a positive integer) remains constant when the distance between the points does not change, although any transformation could change the position of them. Therefore, the concept of statistical homogeneity exists when the statistical properties of a random field are not dependent on the spatial coordinates of the points, so they are independent of the reference system. In practice, statistical homogeneity assumes that there are no systematic

trends, and thus the variation of field values can be attributed to fluctuations around a constant level equal to the mean value [50].

2.5 Statistical Isotropy

Statistical isotropy is a useful tool in a geostatistical analysis. A field is statistically isotropic only if is statistically homogeneous and the covariance function depends only on the Euclidean distance and not on the direction of the distance vector $\mathbf{r}$. If a covariance function is statistically isotropic is by definition statistically homogeneous, but not conversely.

In contrast, if the spatial variability depends on the direction, then the random field is defined by anisotropy. The covariance of an anisotropic random field depends on both the distance r and the direction of the vector $\mathbf{r}$. A random field is anisotropic when the directional covariance functions have different values either in the variance or the correlation length [47, 50]. The variance is a measure of the amplitude of the fluctuations in the field. The correlation length defines the interval within there is interdependence, i.e. defines the distance within which a value of a point affects the value at another point, in the field. In the case of statistically isotropic fields the two most important parameters are the variance $\sigma_X^2 = c_X(0)$ and the correlation length ξ.

2.6 Spatial Analysis

2.6.1 Moments

Statistical moments are deterministic functions, which represent mean values, for all possible field states, of different combinations of field values at one or more locations. The mean value of a quantity A(X)is denoted by $\mathbb{E}[A(X)]$. For a multidimensional moment $\mathbb{E}[X^{k_1}(\mathbf{s_1}) \dots X^{k_N}(\mathbf{s_N})]$, where $k_1 + \dots + k_N = K$, given by the following k-dimensional integral

$$\mathbb{E}[X^{k_1}(\mathbf{s_1}) \dots X^{k_N}(\mathbf{s_N})] = \int dx_1 \int dx_N f_X(x_1, \dots, x_N; \mathbf{s_1}, \dots, \mathbf{s_N})\, \mathbf{x_1}^{\mathbf{k_1}} \dots \mathbf{x_N}^{\mathbf{k_N}}. \tag{2.3}$$

Mean Value

Mean value is similar to the arithmetic average of the data values in a sample, and it defined as

$$m_{_X}(\mathbf{s}) = \mathbb{E}[X(\mathbf{s})]. \tag{2.4}$$

where $X(\mathbf{s})$ is the random field and $\mathbb{E}[]$ is the expectation, calculated over all the states of the field, i.e.

$$\mathbb{E}[X(\mathbf{s})] = \int dx\ f_{_X}(x;\mathbf{s})\ x, \tag{2.5}$$

where x are the values that correspond to a given state. Integral limits depend from the space wherein the field is defined.

In Equation (2.5), the mean value may depend on the location $\mathbf{s}$, which derives from a possible dependence of the one-dimensional probability density function on the position. In practice, the probability density function is not known, and therefore the mean must be calculated from the data by statistical methods. The same applies to the other parameters of the pdf.

Variance

The variance of a random field is estimated by the mean value of the squared fluctuation, according to the following equation:

$$\sigma_x^2(\mathbf{s}) = \mathbb{E}[\{X(\mathbf{s}) - m_x(\mathbf{s})\}^2] = \mathbb{E}[X'^2(\mathbf{s})]. \tag{2.6}$$

The variance can vary from point to point. If the field is statistically homogeneous, the variance is constant at all points.

Errors

In the fields of science, engineering and statistics, measurement accuracy is the degree of proximity of measurements to their true value. While the accuracy of the measurement related to repeatability is the degree to which the repetitions of measurements made under the same conditions produce the same result (error). Statistical bias is a characteristic of statistics, according to which the expected value of

the results differs from the actual value of the parameter being estimated. The errors can be systematic or random . Systematic errors are introduced from method or instrument flaws, so the resulting measurements are inaccurate and biased. Random errors are caused by uncontrolled fluctuations and affect the measurements randomly. In this case, the inaccuracy is due to random fluctuations and not method flaws [27].

Standard Error

The standard error (SE) is the variance of the estimated parameter's distribution. The SE is calculated as:

$$SE = \sqrt{\frac{\sigma_X^2}{n}}, \tag{2.7}$$

where the σ_X^2 is the variance, and n is the size of sample.

2.6.2 Covariance Function

The covariance function $c_X(\mathbf{s}_1, \mathbf{s}_2)$ of a random field $X_\mathbf{s}$, expresses the influence of the value at $\mathbf{s_1}$ on the fluctuation of the value at $\mathbf{s_2}$.

The centered covariance function is defined by the following formula:

$$c_X(\mathbf{s}_1, \mathbf{s}_2) = \mathbb{E}[X(\mathbf{s}_1) \cdot X(\mathbf{s}_2)] - \mathbb{E}[X(\mathbf{s}_1)]\,\mathbb{E}[X(\mathbf{s}_2)]. \tag{2.8}$$

The random field $X'(\mathbf{s}_1) = X(\mathbf{s}_1) - m_x(\mathbf{s}_1)$, represent the fluctuation of the field $X(\mathbf{s_1})$ around the mean value in the point $\mathbf{s_1}$. The mean value of the fluctuation field is equal to zero,

$$\mathbb{E}[X'(\mathbf{s}_1)] = 0. \tag{2.9}$$

Based on the previous equations, the centered covariance function is defined by

$$c_X(\mathbf{s}_1, \mathbf{s}_2) = \mathbb{E}[X'(\mathbf{s}_1)X'(\mathbf{s}_2)]. \tag{2.10}$$

In conclusion, the centered covariance function quantitatively describes the dependence of the fluctuations on two different points [15].

Covariance and Variance

If two points in the random field coincide, the the value of the covariance function is equal to the variance of the field at that point:

$$c_X(\mathbf{s}_1, \mathbf{s}_1) = \sigma_X^2(\mathbf{s}_1). \tag{2.11}$$

Basic Concepts of Covariance Function

In statistically homogeneous and isotropic fields, the two important parameters of covariance are: 1) the **variance** $\sigma_X^2 = c_X(0)$ measures the magnitude of the field fluctuations, and 2) the **correlation length** (ξ), which normalizes the distance (in the covariance function the distance is defined by the ratio r/ξ. The correlation length defines the distance in which the values are correlated. In case of anisotropic dependence, there are different correlations lengths over the direction of anisotropy.

Bochner's Theorem

Not every function can be considered as a covariance function. The permissibility conditions are defined by the Bochner's theorem [11], which is defined by the **spectral density**, given by the Fourier transformation. The Fourier transformation is given by the following equation:

$$\tilde{c}_x(\mathbf{k}) = \int d\mathbf{r}\; e^{-\mathbf{ikr}}\; c_x(\mathbf{r}), \tag{2.12}$$

where **r** is the distance vector between two points and k is the vector of spatial frequency (or wave-number).

The inverse transformation estimated by the following integral:

$$c_x(\mathbf{k}) = \frac{1}{(2\pi)^d} \int d\mathbf{r}\; e^{\mathbf{ikr}}\; \tilde{c}_x(\mathbf{k}). \tag{2.13}$$

The Bochner's Theorem mentioned that: A function $c_X(\mathbf{r})$ is accepted as a covariance function if:

1. It admits the Fourier transform $\tilde{c}_x(\mathbf{k})$.
2. $\tilde{c}_x(\mathbf{k})$ is non-negative over the entire frequency domain.

3. The integral of $\tilde{c}_x(\mathbf{k})$ over the entire frequency domain exists and is bounded.

2.6.3 Variogram Function

The variogram function describes the variance of the difference between the values of a random field $X(\mathbf{s})$ at locations $\mathbf{s_1}$, and $\mathbf{s_2}$ and is defined by the following equation [27]:

$$\gamma_X(\mathbf{s_1}, \mathbf{s_2}) = \frac{1}{2} Var[X(\mathbf{s_1}) - X(\mathbf{s_2})]. \tag{2.14}$$

If the random field has a constant mean value the variogram is defined by means of the expectation:

$$\gamma_X(\mathbf{s_1}, \mathbf{s_2}) = \frac{1}{2} \mathbb{E}\left[X(\mathbf{s_1}) - X(\mathbf{s_2})\right]^2. \tag{2.15}$$

The variogram is defined for a pair of points, using the expectation of the field of squared differences, which is defined as $\delta X(\mathbf{s_1}; \mathbf{s_2}) = X(\mathbf{s_1}) - X(\mathbf{s_2})$.

If the field is statistically homogeneous, then the variogram is directly connected to the covariance function by the equation:

$$\gamma_X(\mathbf{r}) = \sigma_X^2 - c_X(\mathbf{r}). \tag{2.16}$$

Hence the variogram's upper bound of a random field is the variance of the field i.e. c_0. Also from the above equation is mentioned that the variogram tends asymptotically to the variance. For statistically homogeneous fields the variogram contains the same information as the covariance [15]. The variogram function as it seems from the Equation (2.15) is a non-negative function, i.e., $\gamma_X \geq 0$, but the reverse does not apply, i.e., every non-negative function is not necessarily a variogram function. The conditions of Bochner's theorem need to apply.

Statistically homogeneous field

In a homogeneous field with isotropic spatial dependence, the variogram is estimated from two parameters: the sill (upper bound of variogram) and the correlation length. Specifically the value of the variogram, for long distances $\mathbf{r}$ tend asymptotically to an upper bound, equal to variance σ_X^2, of the field. This property is based on the

fact that covariance function is zero for long distances and on the relation $\gamma_X(\mathbf{r}) = \sigma_X^2 - c_X(\mathbf{r})$. The correlation length defines the time in where the variogram approaches the sill, and the range within that two points are correlated.

If correlations characteristics differ with different directions in space, the dependence is anisotropic. In geometrical anisotropy, the sill is independent of the direction, but the speed of approach to the sill depends on the direction. In this case the variogram defined as function $\gamma_X\left(\frac{r_1}{\xi_1}, \ldots, \frac{r_d}{\xi_d}\right)$ of dimensionless distances $\frac{r_1}{\xi_1}, \ldots, \frac{r_d}{\xi_d}$, where the $\xi_1, ..., \xi_d$ are the correlation lengths in corresponding directions.

In the case of zone anisotropy, the upper bound depends from the spatial direction. Then the variogram function can be expressed as:

$$\gamma_X(\mathbf{r}) = \gamma_{X,1}(\mathbf{r}) + \gamma_{X,2}(\widehat{\mathbf{r}}), \tag{2.17}$$

where the $\gamma_{X,1}(r)$ represents the isotropic dependence and the $\gamma_{X,2}(\widehat{\mathbf{r}})$ the anisotropic dependence of the sill in the direction of the unit vector $\widehat{r}$.

Variogram Models

To estimate the variogram at any distance, a theoretical variogram model should be fitted in the experimental. The commonly used theoretical variogram model are exponential, gaussian, spherical, and power-law. In functions σ_X^2 represents the variance of the spatial field, $\|\mathbf{r}\|$ is the Euclidean norm of the lag vector $\mathbf{r}$, and ξ is the correlation length [8, 46].

1. **Exponential**

$$\gamma_X(r) = \sigma_X^2 \left[1 - \exp\left(-\frac{\|\mathbf{r}\|}{\xi}\right)\right] \tag{2.18}$$

2. **Gaussian**

$$\gamma_X(r) = \sigma_X^2 \left[1 - \exp\left(-\frac{\|\mathbf{r}\|^2}{\xi^2}\right)\right] \tag{2.19}$$

3. **Spherical**

$$\gamma_X(r) = \begin{cases} \sigma_X^2 \left[1.5\left(\frac{\|\mathbf{r}\|}{\xi}\right) - 0.5\left(\frac{\|\mathbf{r}\|}{\xi}\right)^3\right] & \|\mathbf{r}\| \leq \xi \\ \sigma_X^2 & \|\mathbf{r}\| \geq \xi \end{cases} \tag{2.20}$$

4. **Power-law**

$$\gamma_X(\|\mathbf{r}\|) = \alpha\|r\|^{2H}, \quad 0 < H < 1, \quad \alpha > 0 \tag{2.21}$$

A theoretical variogram model is accepted as variogram function if it is conditionally negative definite function. This means that for any linear coefficient λ_α, which satisfies the following condition

$$\sum_{\alpha=1}^{n} \lambda_\alpha = 0, \tag{2.22}$$

must satisfy the following inequality:

$$-\sum_{\alpha=1}^{n}\sum_{\beta=1}^{n} \lambda_\alpha\lambda_\beta\gamma_X(\mathbf{s}_\alpha - \mathbf{s}_\beta) \geq 0. \tag{2.23}$$

In practice, inequality control is not feasible for every possible combination of coefficients λ_α, so the acceptance criterion is expressed with Bochner theorem. According to Bochner theorem, the function $\gamma_X(\mathbf{r})$ is admissible as variogram function in d dimensions if the following is applied [27]:

1. $\gamma_X(0) = 0$,
2. the generalized Fourier trasformation $\tilde{\gamma}_X(\mathbf{k})$, exists,
3. $\tilde{\gamma}_X(\mathbf{k})$ satisfies the inequality: $-k^2\tilde{\gamma}_X(\mathbf{k}) \geq 0$ and
4. $\lim \gamma_X(\mathbf{r})/r^2 = 0 \quad r \to \infty$.

If the random field is statistically homogeneous, is easy to test the acceptance of a theoretical variogram, using the covariance $\sigma_X^2 - \gamma_X(\mathbf{r})$. If the function $\gamma_X(\mathbf{r})$ represents an acceptable variogram, then the function $c_X(\mathbf{r}) = \sigma_X^2 - \gamma_X(\mathbf{r})$ is permissible covariance function and vice versa [16].

2.6.4 Spatial Estimation

A significant problem in geostastics is the estimation of a variable of interest over an entire area on the basis of values observed at a limited number of points. Estimation aims to provide information about points in space where no measurements are available. From a deterministic viewpoint, this is an interpolation problem. The variable is estimated by a parametric function, either explicitly or implicitly. The estimation

of the variable can be either local, if it refers to a specific point , or globally if it refers to the calculation of a characteristic value for an entire area. The estimation of the field, requires a model containing the spatial dependence,so it is possible to estimate points, where there no measurements, using the neighboring measured field values. The common used methods which is based on the minimization of the square error of estimate, and the linear interpolation are known as "kriging" [40].

The problem of local estimation is described as follows: In a data set $X(\mathbf{s_i})$, at points $\mathbf{s_i}\,(i = 1, \ldots, N)$, which are located in a region Ω define the value of the field at the estimation point $\mathbf{u} \in \Omega$, which does not coincide with any of the $\mathbf{s_i}$. The estimate at the point $\mathbf{u}$ is denoted as $\widehat{X}(\mathbf{u})$. The estimation process is repeated at every node of the grid, which is defined from the particular application.

To reduce computing intensity in kriging methods a neighborhood $\omega(\mathbf{u})$ around the point $\mathbf{u}$ is often determined. This neighborhood includes $n(\mathbf{u}) \leqslant N$ points at $\mathbf{s_i}, (i = 1, \ldots, N)$. The size of the neighborhood defined from the correlation length. In linear interpolation methods, the field fluctuation at the estimate point is expressed by the following linear combination:

$$\widehat{X}(\mathbf{u}) - m_{_X}(\mathbf{u}) = \sum_{\alpha=1}^{n(\mathbf{u})} \lambda_\alpha [X(\mathbf{s}_\alpha) - m_{_X}(\mathbf{s}_\alpha)]. \tag{2.24}$$

The coefficients λ_α, represents the linear weights. Hence the Equation (2.24), estimates the fluctuation at the prediction point.

The estimate of the field is given by the following equation:

$$\widehat{X}(\mathbf{u}) = m_{_X}(\mathbf{u}) + \sum_{\alpha=1}^{n(\mathbf{u})} \lambda_\alpha [X(\mathbf{s}_\alpha) - m_{_X}(\mathbf{s}_\alpha)]. \tag{2.25}$$

In stochastic methods, the estimator $\widehat{X}(\mathbf{u})$ is a random variable as is the estimation error $\epsilon(\mathbf{u}) = \widehat{X}(\mathbf{u}) - X(\mathbf{u})$. Kriging methods estimate the optimal value $X(\mathbf{u})$, using the weight, which minimize the variance of estimate error. Kriging is the best linear unbiased estimator since it minimizes the square of the prediction error.

The common used kriging methods are described below.

2.6.5 Simple kriging

Simple kriging is used when the mean value $m_{_X}$ of the random field is known and constant throughout the entire field. In this case, the kriging estimator $\widehat{X}(\mathbf{u})$ is determined by the equation:

$$\widehat{X}(\mathbf{u}) = \sum_{\alpha=1}^{n(\mathbf{u})} \lambda_\alpha X(\mathbf{s}_\alpha) - m_{_X} \left[\sum_{\alpha=1}^{n(\mathbf{u})} \lambda_\alpha X(\mathbf{s}_\alpha) - 1 \right]. \tag{2.26}$$

The linear weights, λ_α are estimated by minimizing the error variance, given by the equation:

$$\sigma^2_{E,SK}(\mathbf{u}) = Var[X(\mathbf{u}) - \widehat{X}(\mathbf{u})] = Var[\widehat{X}(\mathbf{u}) - m_{_X} - X'(\mathbf{u})]. \tag{2.27}$$

The equation of the estimator $\widehat{X}(\mathbf{u})$ leads to the following relation for the fluctuation of the random variable $\widehat{X}(\mathbf{u})$:

$$\widehat{X}(\mathbf{u}) - m_{_X} = \sum_{\alpha=1}^{n(\mathbf{u})} \lambda_\alpha [X(s_\alpha) - m_{_X} = \sum_{\alpha=1}^{n(\mathbf{u})} \lambda_\alpha X'(s_\alpha). \tag{2.28}$$

The linear function $\sum_{\beta=1}^{n(\mathbf{u})} \lambda_\beta c_{_X}(\mathbf{s}_\alpha - \mathbf{s}_\beta) = c_{_X}(\mathbf{s}_\alpha - \mathbf{u}), \quad \alpha = 1, ..., n(\mathbf{u})$, can be expressed as

$$\mathbf{C}_{\alpha,\beta}\lambda_\beta = \mathbf{C}_{\alpha,u}, \tag{2.29}$$

where the matrix $\mathbf{C}_{\alpha,\beta}$ represents the covariance matrix, with elements $\mathbf{C}_{\alpha,\beta} = c_X(\mathbf{s}_\alpha - \mathbf{s}_\beta)$. The vector $\mathbf{C}_{\alpha,u}$ represents the values of the covariance function between the sample points and the estimation points $\mathbf{C}_{\alpha,u} = c_{_X}(\mathbf{s}_\alpha - \mathbf{u})$.

Considering the equation $c_{_X}(0) = \sigma^2_{_X}$, the linear system is written in the form of matrices as follows:

$$\begin{bmatrix} \sigma^2_{_X} & \dots & \dots & c_{_X}(\mathbf{s}_1 - \mathbf{s}_n) \\ c_{_X}(\mathbf{s}_2 - \mathbf{s}_1) & \dots & \dots & c_{_X}(\mathbf{s}_2 - \mathbf{s}_n) \\ \vdots & \vdots & \vdots & \vdots \\ c_{_X}(\mathbf{s}_n - \mathbf{s}_1) & \dots & \dots & \sigma^2_{_X} \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_n \end{bmatrix} = \begin{bmatrix} c_{_X}(\mathbf{s}_1 - \mathbf{u}) \\ c_{_X}(\mathbf{s}_2 - \mathbf{u}) \\ \vdots \\ c_{_X}(\mathbf{s}_n - \mathbf{u}) \end{bmatrix} \tag{2.30}$$

The solution of the linear system is given by the following equation:

$$\lambda_\beta = C^{-1}_{\beta,\alpha} C_{\alpha,u}, \forall \beta = 1, ..., n(\mathbf{u}). \tag{2.31}$$

In the case of stationarity, variogram and covariance are connected through the following equation

$$c_X(\mathbf{s}_\alpha, \mathbf{s}_\beta) = \sigma^2_X - \gamma_X(\mathbf{s}_\alpha, \mathbf{s}_\beta). \tag{2.32}$$

Solving the above linear system, gives the λ_α coefficient values, if the covariance function is permissible and each point has a unique value. The values of the linear weights are independent from the sill of variogram, but they depend from the correlation length. Kriging is an exact interpolator, i.e. at every point where a measurements is available, the krigng estimate coincides with the sample value.

The uncertainty of the estimation is determined by the squared root of the variance of the estimation error. The variance $\sigma^2_{E,SK}(\mathbf{u})$ is defined by the following equation:

$$\sigma^2_{E,SK}(\mathbf{u}) = \sigma^2_X - \sum_{\alpha=1}^{n(\mathbf{u})} \sum_{\beta=1}^{n(\mathbf{u})} C_{u,\alpha} C^{-1}_{\alpha,\beta} C_{\beta,u}. \tag{2.33}$$

According to the Equation (2.33), the error variance increases proportionally to the random field variance σ^2_X. The error increases as the distance $\|\mathbf{u} - \mathbf{s}_\alpha\|$ between the estimation point and the data points [15, 24].

2.6.6 Ordinary Kriging

In ordinary kriging, the mean value is considered constant inside the local neighborhood, but may vary from neighborhood to neighborhood. The mean value is not necessarily known. In this case, the mean value is not calculated from the average of the sample values, but is calculated providing that the coefficient function is known. The estimate is calculated from the following equations:

$$\hat{X}(\mathbf{u}) = \sum_{\alpha=1}^{n(\mathbf{u})} \lambda_\alpha X(\mathbf{s}_\alpha), \tag{2.34}$$

and

$$\sum_{\alpha=1}^{n(\mathbf{u})} \lambda_\alpha = 1. \tag{2.35}$$

The Equation (2.35) represents the non-bias condition. In ordinary kriging minimum mean square error should be calculated using the restriction imposed by the non-bias constraint. The minimization of the error variance under the non-bias condition $\sum_{\alpha=1}^{n(\mathbf{u})} \lambda_\alpha = 1$ uses the Lagrange multipliers method for constrained minimization.

To calculate the linear weights following equation is used:

$$\sum_{\beta=1}^{n(\mathbf{u})} \lambda_\beta c_X(\mathbf{s}_\alpha - \mathbf{s}_\beta) + \mu = c_X(\mathbf{s}_\alpha - \mathbf{u}), \quad \alpha = 1, ..., n(\mathbf{u}), \tag{2.36}$$

where μ is the Lagrange coefficient, and

$$\sum_{\alpha=1}^{n(\mathbf{u})} \lambda_\alpha = 1. \tag{2.37}$$

The linear system of Equations (2.36), and (2.37), is possible to be written in the form of matrices as follows:

$$\begin{bmatrix} \sigma_X^2 & c_X(\mathbf{s}_1 - \mathbf{s}_2) & \dots & c_X(\mathbf{s}_1 - \mathbf{s}_n) & 1 \\ c_X(\mathbf{s}_2 - \mathbf{s}_1) & \sigma_X^2 & \dots & c_X(\mathbf{s}_2 - \mathbf{s}_n) & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ c_X(\mathbf{s}_n - \mathbf{s}_1) & c_X(\mathbf{s}_n - \mathbf{s}_2) & \dots & \sigma_X^2 & 1 \\ 1 & 1 & \dots & 1 & 0 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_n \\ \mu \end{bmatrix} = \begin{bmatrix} c_X(\mathbf{s}_1 - \mathbf{u}) \\ c_X(\mathbf{s}_2 - \mathbf{u}) \\ \vdots \\ c_X(\mathbf{s}_n - \mathbf{u}) \\ 1 \end{bmatrix} \tag{2.38}$$

The solution of the linear system is given by the following equation:

$$\lambda_\beta = \mathbf{C}_{\beta,\alpha}^{-1} \mathbf{C}_{\alpha,u}, \quad \forall \beta = 1, \dots, n(\mathbf{u}). \tag{2.39}$$

The optimal estimate of the kriging error variance is respectively given by the equation:

$$\sigma^2_{E,OK}(\mathbf{u}) = \sigma^2_X - \sum_{\alpha=1}^{n(\mathbf{u})} \lambda_\alpha c_X(\mathbf{u}, \mathbf{s}_\alpha) - \mu, \tag{2.40}$$

where the Lagrange multiplier takes values $\mu < 0$.

2.7 Spatial Model Validation

For assessing the performance of a different parameters for the same model or to compare the performance of different models, validations methods are used. Methods like likelihood maximization, empirical contrast minimization, and least squares quantify the fit of the data to spatial models. Validation includes methods that estimate the predictive performance of the model based on the sample data. The most used validation method is Cross-Validation (CV).

CV methods separate the data, once or several times, to estimate the reliability and the accuracy of the model based on the subsets. The data is first split in a training set, and a validation set. The validation set is used to estimate the predictive performance of the model. The optimal model is the one with the best validation measures as chosen by the investigator.

The most commonly used cross-validation methods are **leave-p-out cross validation**, and **leave-one-out cross validation**. In **leave-p-out cross validation** (LPO CV), the original data are split in training set with N-p sample points (N refers to the number of original data), and the validation set with p remaining sample points. Afterwards the spatial model is tested based on the training set, and the predictions are compared with the validation set. This process can be repeated as many times as the possible partition of the set of N into two sets. **Leave-one-out cross validation** (LOOV CV), is a specific case of LPO CV where p=1, i.e the training set contains N-1 points in each iteration, and the validation set is a single point. The process is implemented N times, and the CV measure is calculated as the average of N cases [27].

2.7.1 Cross-Validation Measures

Assume a random field $X_\mathbf{s}$ with known values at locations s_i, $i = 1, \ldots, N$. The statistical measures are evaluated, in order to asses the model performance. These measures include: the mean error (bias) (ME), the mean absolutely error (MAE),

the root mean square error (RMSE), and Pearson's linear correlation coefficient (ρ). For the following measures, $\widehat{x}(s_i)$ and $x(s_i)$ are the estimated and true value of the field at point s_i, $\overline{x(s_i)}$ is the spatial average value of the data, $\overline{\widehat{x}(s_i)}$ is the spatial average of the estimates, and N is the number of the observation.

Mean Error (bias)

The mean error is estimated as follows:

$$ME = \frac{1}{N}\sum_{i=1}^{N}\left[\hat{x}(\mathbf{s_i}) - x(\mathbf{s_i})\right]. \tag{2.41}$$

High and positive or negative values of mean error denotes bias.

Mean Absolute Error (MAE)

The mean absolute error is estimated as follows:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|\hat{x}(\mathbf{s_i}) - x(\mathbf{s_i})|. \tag{2.42}$$

The mean absolute error estimates the accuracy and the precision of the estimation.

Root Mean Square Error (RMSE)

The root mean square error is estimated as follows:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left[\hat{x}(\mathbf{s_i}) - x(\mathbf{s_i})\right]^2}. \tag{2.43}$$

The root mean square error calculates the accuracy and the precision of the estimation as mean error. Also, because of the squaring the errors, RMSE gives higher weights on large deviations.

Pearson's Correlation Coefficient (ρ)

The Correlation Coefficient is the most commonly used to measure the relationship between the data and the estimates, and is calculated as follows:

$$\bar{\varrho}_{X,\hat{X}} = \frac{\sum_{i=1}^{N}\left[x(\mathbf{s_i}) - \overline{x(\mathbf{s_i})}\right]\left[\hat{x}(\mathbf{s_i}) - \overline{\hat{x}(\mathbf{s_i})}\right]}{\sqrt{\sum_{i=1}^{N}\left[x(\mathbf{s_i}) - \overline{x(\mathbf{s_i})}\right]^2}\sqrt{\sum_{i=1}^{N}\left[\hat{x}(\mathbf{s_i}) - \overline{\hat{x}(\mathbf{s_i})}\right]^2}}. \tag{2.44}$$

Pearson's Correlation Coefficient estimates the linear relationship between two variables. The coefficient can be described by a scatterplot. If $\rho = +1$ or $\rho = -1$ the scatterplot is a straight line with positive or negative slop respectively. If $\mid \rho \mid < 1$ the values appear as a cloud of points, which becomes more diffuse as $\mid \rho \mid$ decreases from 1 to 0 [32].

Chapter 3

Time Series Analysis

Time series is a data set which is collected over time and expresses the evolution of a variable's values over time. Specifically, time series include a set of observations which they are usually collected using a set time step. Time series is a stochastic process, because the values are affected from random factors, although the value of each step is random variable. So time series is a collection of random variables i.e. $X_t, t \in T$, where T is a set of time observations [59]. To investigate the behavior of the variable, only values from previous time periods are needed. Forecasting is implemented based on such values to predict the values in the following time periods. The random variables X_t are distributed according to a univariate cumulative distribution function F_t and a respective probability density function f_t.

Methods of time series analysis have many applications in different disciplines. For example, such methods are used in economics [25], climate studies (related to global warming) [26], and earthquakes [19].

White noise: The simplest case of a time series model is white noise, which is a collection of independent and identically distributed (iid) random variables, W_t, with zero mean value and variance σ_W^2. If the noise values follow the Gaussian distribution, i.e., $\epsilon_t \sim N(0, \sigma_\epsilon^2)$, it is known as Gaussian White Noise (GWN).

3.1 Moments

Mean value

The mean value (expected value) of a time series at time t is defined by the following equation:

$$\mu_x(t) = \mathbb{E}[X_t] = \int_{-\infty}^{\infty} x f_t(x) dx, \tag{3.1}$$

where $\mathbb{E}[\cdot]$ denotes the expectation and is calculated over all the possible states of X.

Autocovariance Function

The autocovariance function constitutes the second moment product and is defined as:

$$\gamma_X = \text{cov}(x_i, x_j) = \mathbb{E}[(X_i - \mu_i)(X_j - \mu_j)], \tag{3.2}$$

for all i and j [1]. If there is no reference in which time series refers to, is possible to write $\gamma_x(i,j)$ as $\gamma(i,j)$. Note that $\gamma_x(i,j) = \gamma_x(i,j)$, for all time points t_i and t_j [14]. The autocovariance function defines the linear dependence between two points on the same series at different times. If the autocovariance $\gamma_x(i,j) = 0$, then there is no linear correlation between X_i and X_j, but there may be some dependence structure between them. From the Equation (3.2), it transpires that the autocovariance is the same as the variance for $i = j$,

$$\gamma_x(i,i) = \mathbb{E}[(X_i - \mu_i)^2] = \text{var}(X_i). \tag{3.3}$$

Covariance for Linear Combinations

If two random variables U, V are a linear combination of random variables of $\{X_j\}$, and $\{U_k\}$, i.e.

[1] Note that in time series analysis the symbol γ is used for the covariance function, while in geostatistics it is used for the variogram function.

$$U = \sum_{j=1}^{m} \alpha_j X_j \tag{3.4}$$

and

$$V = \sum_{k=1}^{r} \beta_k Y_k, \tag{3.5}$$

then the covariance function is defined as

$$\operatorname{cov}(U, V) = \sum_{j=1}^{m} \sum_{k=1}^{r} \alpha_j \beta_k \operatorname{cov}(X_j, Y_j). \tag{3.6}$$

Autocorrelation Function (ACF)

The autocorrelation is a useful tool in time series analysis, for defining repeated patterns. Also autocorrelation assess dependence between the observations, considered that the time series is stationary [59]. The autocorrelation Function (ACF) measures the linear predictability of the series at time t and is estimated by the following equation:

$$\rho(s, t) = \frac{\gamma(s, t)}{\sqrt{\gamma(s, s)\gamma(t, t)}} \tag{3.7}$$

From the above equation it follows that $-1 \leq \rho(s, t) \leq 1$. If X_t can be perfectly predicted from X_s through a linear relationship $X_t = \beta_0 + \beta_z X_s$, then the ACF will be $+1$, when $\beta_1 > 0$, and -1 when $\beta_1 < 0$ [45].

3.2 Trend and Seasonality

Trend

The trend can be considered as a long–term change in the mean value of the time series. It can be increasing, decreasing or constant over a time period. The trend can be determined as a linear trend, an exponential trend, harmonic, etc. It should be mentioned that to observe a trend in a time series, there must be a satisfactory number of the observations. The trend must be removed from the time series to ensure stationarity and to proceed with further stochastic analysis [51].

Seasonality

Some time series present a repeated pattern. These patterns may be directly visible from the time series, or can be observed only after inspecting the periodogram. The periodgram is a useful tool to identify periodicity in time series. In a periodogram plot the spectral density is presented. The spectral density is calculated using the Fourier transformation on the Time Series after the trend is removed. Seasonality shows up as high spectral power in the periodogram [59].

3.3 Stationarity

A time series is defined as stationary if the mean value and variance are constant in the entire time, and the periodic variations have been removed. Also the autocovariance $\gamma(s,t) = \text{cov}(x_s, x_t) = \mathbb{E}[(x_s - \mu_s)(x_t - \mu_t)]$ depends on s and t only through their difference $|s-t|$. There are two forms of stationary: the strict stationarity and the second-order (weak) [14, 59] stationarity (see section 2.4).

The autocovariance function for stationary time series is estimated as follows:

$$\gamma(h) = \text{cov}(X_{t+h}, X_t) = \mathbb{E}[(X_{t+h} - \mu)(X_t - \mu_t)], \tag{3.8}$$

and the autocorrelation (ACF) is expressed as

$$\rho(h) = \frac{\gamma(t+h,t)}{\sqrt{\gamma(t+h,t+h)\gamma(t,t)}} = \frac{\gamma(h)}{\gamma(0)}. \tag{3.9}$$

Testing Stationarity

There are several ways to test the time series stationarity. A simple test is to examine the mean value of the sample. Mean value is calculated in different ranges over the entire time series, and the result must be the same or similar in all the ranges tested. Also the stationarity can be observed from the autocorrelation and partial autocorrelation graph. If the autocorrelation tends to zero after some lags, it is an indication that the time series is stationary.

Also the stationarity can be defined by the Kwiatkowski–Philips–Schmidt–Shin tests (KPSS) [41]. KPSS tests the null hypothesis (i.e. the time series is stationary around a determenistic trend) against the alternative. So small p–values suggest

that the time series is not stationary and other approaches should be considered (i.e differencing, decomposition etc). Finally, to check the stationarity, the Augmented–Dickey–Fuller test (ADF) can be used. For ADF test the null hypothesis is that the time series contains a deterministic component [21].

3.4 Decomposition of Time Series

In time series analysis, it is useful to plot the data set, to check if there is any discontinuities in the time series, such as a sudden change of the value. To analyze these cases it is important to separate the time series into homogeneous segments. If outliers exist in the time series, it should be tested whether they must be removed. Moreover, the time series should be checked for trend and seasonality [45]. So, the typical decomposition model is defined by the following equation:

$$X_t = m_t + s_t + R_t. \tag{3.10}$$

In Equation (3.11), m_t represents the trend component, s_t is the seasonal component with known period d, and R_t is the stationary remainder.

If the trend and the seasonal component change within the time series, transformation of the data can be implemented so that the transformed data are more compatible with the Equation (3.11).

The aim of estimating the deterministic components m_t and s_t is to provide residuals or noise component, which should be stationary [21].

3.4.1 Trend estimation

If the seasonal component is missing, then the model of Equation (3.11), becomes:

$$X_t = m_t + R_t, \tag{3.11}$$

There are two approaches to estimate the trend model. The first is to fit a polynomial trend model, then to subtract the trend from the data, and create a time series of the residuals, which should be stationary. The second approach is to estimate the trend by differencing the time series, and find an appropriate stationary time series [59].

Trend Estimation by Fitting a Polynomial

For the first method a trend polynomial such as $m_t = \alpha_0 + \alpha_1 t + \alpha_2 t^2$ can be fitted to the data $X_1, \ldots, X_t$. The parameters of the trend model $\alpha_0, \alpha_1, \alpha_2$ are estimated by minimizing the sum of error squares $\sum_{t=1}^{n}(X_t - m_t)$ using regression. The least squares method can also for high order polynomial trend [21].

Trend Estimation with Differencing

To estimate the trend by differencing, the lag-1 difference operator ∇ to estimate the trend is defined:

$$\nabla X_t = X_t - X_{t-1} = (1 - B)X_t, \tag{3.12}$$

where B is the backward shift operator, $BX_t = X_{t-1}$.

The powers of the parameters B and ∇ are defined by $B^j(X_t) = X_{t-j}$ and $\nabla^j(X_j) = \nabla\left[\nabla^{j-1}(X_t)\right], j \geq 1$, with $\nabla^0(X_t) = X_t$.

If the operator ∇ is used on a linear trend function like $m_t = c_0 + c_1 t$, then function becomes $\nabla m_t = m_t - m_{t-1} = c_0 + c_1 t - [c_0 + c_1(t-1)] = c_1$ which is a simple constant. In the same way, any polynomial trend of degree k, can be expressed as a constant, with the operator ∇^k.

3.4.2 Estimation of Seasonal Effects

To estimate the seasonal component in the Equation (3.11), it is assumed that the seasonal is equal to $s_t = s_{t+d}$ and $\sum_{j=1}^{d} s_j = 0$, where d is the periodicity. There are several methods to estimate the seasonal component. One of them is to estimate the seasonality by differencing the time series, as in trend component (see Section 3.4.1). Also the smoothing approach is used, estimating the trend component first and then the seasonal component after removal the trend. Another approach is to fit a harmonic model to the time series (parametric approach) [21].

Smoothing Filter

In a set of observations $x_1, \ldots, x_n$, the trend is estimated first. In order to eliminate the seasonal component, the smoothing filter is used. It is obtained a time series $\widehat{m}_t$, representing the trend component as follows:

$$\widehat{m}_t = \sum_{i=-p}^{q} \alpha_i X_{t-i}, \tag{3.13}$$

where $\alpha_i = \frac{1}{2p+1}$ represents the weights, p and q define the window.

In order the sum is defined in the entire time series, the Equation (3.13) is expressed as follows:

$$\widehat{m}_t = \frac{0.5X_{t-q} + X_{t-q+1} + \ldots + X_{t-q} + 0.5X_{t+q}}{d}, \qquad q < t \leq n - q, \tag{3.14}$$

where d is the periodicity.

As such, the the seasonal component s_t is defined by:

$$\widehat{s}_t = X_t - \widehat{m}_t \tag{3.15}$$

Finally, the seasonal component is subtracted from the original data. Afterwards, the residuals used for prediction and estimation are given by $\widehat{R}_t = X_t - \widehat{m}_t - \widehat{s}_t$.

Parsimonious Decomposition

In this method a linear trend model, a harmonic model and a remainder term are used to decompose the time series [21]:

$$X_t = \beta_0 + \beta_1 t + \beta_2 \sin(2\pi t) + \beta_2 \cos(2\pi t) + R_t. \tag{3.16}$$

The above model has four unknowns that can be estimated with the least squares method, or robust fitting methods.

3.5 Time Series Models

Models of time series represent different stochastic methods and can be used to forecast future values in time series. The common models are Autoregressive (AR) model, the Moving Average model (MA), their combination, i.e., the Autoregressive Moving Average model (ARMA), the Autoregressive Integrated Moving Average

(ARIMA) Model and the Seasonal Autoregressive Integrated Moving Average models (SARIMA).

3.5.1 Autoregressive Model (AR)

Autoregressive model (AR) is a stochastic process, in which a value of the series, X_t can expressed as a function of p past values, $X_{t-1}, \ldots, X_{t-p}$. The order of the model is p, i.e the lagged values required to forecast the current value [51]. An autoregressive model of order p, AR(p) is expressed as

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots + \phi_p X_{t-p} + \epsilon_t, \tag{3.17}$$

where X_t is a series, ϵ_t is white noise with variance σ_ϵ^2, and $\phi_1, \phi_2, \ldots, \phi_p$ are constants.

In Equation (3.17) $\mathbb{E}[X_t]$ is considered as $\mathbb{E}[X_t] = 0$, but if the mean value is not zero, the above equation is expressed as:

$$X_t = \alpha + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots + \phi_p X_{t-p} + \epsilon_t, \tag{3.18}$$

where $\alpha = \mu(1 - \phi_1 - \ldots - \phi_p)$. Also the AR(p) can be written using the backshift operator B, so the Equation (3.17) is expressed as

$$(1 - \phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p) X_t = \epsilon_t, \tag{3.19}$$

or concisely $\phi(B)X_t = \epsilon_t$, where the $\phi(B)$ is the autoregressive operator. This polynomial is used to check the stationarity of the time series X_t. If the polynomial's roots lie outside of the unit circle, the AR(p) time series is stationary.

The AR(1) Model

If the order of the autoregression model is equal to one, the Equation 3.17 is expressed as $X_t = \phi X_{t-1} + \epsilon_t$. Also an AR(1) model can be expressed as a linear process as follows

$$X_t = \sum_{j=0}^{\infty} \phi^j \epsilon_{t-j}. \tag{3.20}$$

The AR(1) is stationary with zero mean value, i.e.,

$$\mathbb{E}(X_t) = \sum_{j=0}^{\infty} \phi^j \, \mathbb{E}(\epsilon_{t-j}) = 0. \tag{3.21}$$

The autocovariance function is given by

$$\gamma(h) = \frac{\sigma_\epsilon^2 \phi^h}{1 - \phi^2}, \; h \geq 0, \tag{3.22}$$

where the depedence between the observations decreases when the lag increases, when $|\phi| < 1$.

The autocorrelation the function by means of

$$\rho(h) = \frac{\gamma(h)}{\gamma(0)} = \phi^h, \; h \geq 0. \tag{3.23}$$

3.5.2 Moving Average Model (MA)

Moving Average model(MA), is a linear combination of the current innovation term ϵ, plus the q most recent ones $\epsilon_{t-1}, \ldots, \epsilon_{t-q}$ [45]. The moving average model of order q, i.e., MA(q), is defined as follows:

$$X_t = \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \ldots + \theta_q \epsilon_{t-q}, \tag{3.24}$$

where $\epsilon_t \sim \epsilon_n(0, \sigma_\epsilon^2)$, and $\theta_1, \theta_2, \ldots, \theta_q, (\theta_q \neq 0)$ are parameters.

The moving average model could also be written using the backshift operator B as:

$$\theta(B) = 1 + \theta_1 B + \theta_2 B^2 + \ldots + \theta_q B^q. \tag{3.25}$$

The MA model is always stationary, because it depends only on the parameters θ and the term ϵ_t, which represents the white noise.

The MA(1) model

The MA first order is defined as $X_t = \epsilon_t + \theta \epsilon_{t-1}$, with mean value $\mathbb{E}(X_t) = 0$. The autocovariance is expressed as:

$$\gamma(h) = \begin{cases} (1+\theta^2){\sigma_\epsilon}^2, & h = 0, \\ \theta\sigma_\epsilon^2, & h = 1, \\ 0, & h > 1, \end{cases} \tag{3.26}$$

and the autocorrelation function is expressed as:

$$\rho(h) = \begin{cases} \frac{\theta}{(1+\theta^2)}, & h = 1, \\ 0, & h > 1. \end{cases} \tag{3.27}$$

3.5.3 Autoregression-Moving Average Model (ARMA)

These models are a combination of the autoregressive and the moving average models, for stationary time series [59]. An ARMA (p,q) model is defined as:

$$X_t = \phi_1 X_{t-1} + \ldots + \phi_p X_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} + \ldots + \theta_q \epsilon_{t-q}, \tag{3.28}$$

where $\phi_p \neq 0$, $\theta_q \neq 0$, and $\sigma_w^2 > 0$. The order of model is defined by the order of autoregressive model and moving average model, p and q respectively.

If X_t has $\mathbb{E}[X_t] = \mu \neq 0$, then $\alpha = \mu(1 - \phi_1 - \ldots - \phi_p)$, and the ARMA model is expressed as:

$$X_t = \alpha + \phi_1 X_{t-1} + \ldots + \phi_p X_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} + \ldots + \theta_q \epsilon_{t-q}, \tag{3.29}$$

where $\epsilon_t \sim \epsilon_n(0, \sigma_\epsilon^2)$.

The ARMA models can also be written using the AR operator, and the MA operator, so it is easier to investigate them. So the ARMA(p,q) model is expressed as $\phi(B)X_t = \theta(B)\epsilon_t$.

As the AR models, ARMA models can also expressed as a one–sided linear process, such as:

$$X_t = \sum_{j=0}^{\infty} \psi \epsilon_{t-j} = \psi(B)\epsilon_t, \tag{3.30}$$

where $\psi(B) = \sum_{j=0}^{\infty} \psi_j B^j$, and $\sum_{j=0}^{\infty} |\psi_j| < \infty$.

The ARMA(1,1) Model

The ARMA(1,1) model is defined as: $X_t = \phi X_{t-1} + \theta \epsilon_{t-1} + \epsilon_t$, with mean value, with mean value $\mathbb{E}[X_t] = 0$.

The autocovariance function is expressed as:

$$\gamma(h) = c\phi^h,\ h = 1, 2, \ldots, \tag{3.31}$$

and the autocorrelation function is defined as:

$$\rho(h) = \frac{\gamma(h)}{\gamma(0)} = \frac{(1 + \theta\phi)(\phi + \theta)}{1 + 2\theta\phi + \theta^2}\phi^{h-1},\ h \geqslant 1. \tag{3.32}$$

3.5.4 The Autoregressive Integrated Moving Average (ARIMA) Model

ARIMA models are used to describe series with a trend, which can be removed by differencing. These differences can be described with an ARMA (p,q). So if the dth order difference of a X_t is an ARMA(p,q) model, then it can described by an ARIMA(p,q,d) model [45]. An ARIMA(p,q,d) model is described with the backshift operator B as:

$$\Phi(B)(1 - B)^d X_t = \Theta(B)\epsilon_t. \tag{3.33}$$

3.5.5 Seasonal Autoregressive Model (SAR)

A Seasonal Autoregressive model (SAR), of order p is defined as:

$$X_t = \Phi_1 X_{t-1} - \Phi_2 X_{t-2} - \ldots - \Phi_p X_{t-p} + \epsilon_t, \tag{3.34}$$

where $\epsilon_t \sim \epsilon_n(0, \sigma_\epsilon^2)$, and $\Phi_1, \Phi_2, \ldots, \Phi_{t-p}$ are constants.

The SAR model is stationary process, if the solutions of the seasonal characteristic equation is equal to 1 in absolute value. SAR is a specific case of an AR model of order $p = Ps$. In this case all ϕ coefficients are equal to zero, except at the seasonal lags $s, 2s, \ldots, Ps$.

The SAR(1) model

The first order seasonal autoregressive with annual seasonal period (s=12 months) is defined as:

$$X_t = \Phi X_{t-12} + \epsilon_t, \text{ or } (1 - \Phi B^{12})X_t = \epsilon_t. \tag{3.35}$$

For the SAR(1) model, using the techniques of seasonal AR(1), so the autocovariance is defined as:

$$\begin{cases} \gamma(0) = \frac{\sigma_\epsilon}{1-\Phi^2} \\ \gamma(\pm 12h) = \frac{\sigma_\epsilon^2 \Phi^h}{1-\Phi^2}, \quad h = 0, 1, 2, \ldots, \\ \gamma(h) = 0, \qquad\qquad \text{otherwise.} \end{cases} \tag{3.36}$$

The autocorrelation is estimated as:

$$\rho(\pm 12h) = \Phi^h. \tag{3.37}$$

3.5.6 Seasonal Moving Average Model (SMA)

In general a Seasonal Moving Average model (SMA), of order Q with seasonal period is defined as:

$$X_t = \epsilon_t + \Theta_1 \epsilon_{t-1} - \Theta_2 \epsilon_{t-2} - \ldots - \Theta_q \epsilon_{t-q}, \tag{3.38}$$

where $\epsilon_t \sim \epsilon_n(0, \sigma_\epsilon^2)$, and $\Theta_1, \Theta_2, \ldots, \Theta_{t-q}$ are constants.

SMA is a stationary process with an autocorrelation function, that is not nonzero only at the seasonal lags $s, 2s, \ldots, Qs$. The SMA is a specific case of an MA model of order $q = Qs$, in which all θ coefficients are equal to zero, except at the seasonal lags $s, 2s, \ldots, QS$.

The SMA(1) model

The first order seasonal moving average model with annual seasonal period (s=12 months) is defined as:

$$X_t = \epsilon_t + \Theta\epsilon_{t-12}. \tag{3.39}$$

For the SMA(1) model, using the techniques of seasonal MA(1), so the autocovariance is defined as:

$$\begin{cases} \gamma(0) = (1+\Theta^2)\sigma_\epsilon^2 \\ \gamma(\pm 12h) = \Theta\sigma_\epsilon^2, \qquad h = 0, 1, 2, \ldots, \\ \gamma(h) = 0, \qquad\qquad\quad \text{otherwise.} \end{cases} \tag{3.40}$$

The autocorrelation is estimated as:

$$\rho(\pm 12h) = \frac{\Theta}{1+\Theta^2}. \tag{3.41}$$

3.5.7 Seasonal ARIMA models (SARIMA)

In time series like environmental data seasonal fluctuations are often encountered. As such, it is necessary to introduce autoregressive and moving average models which determined with seasonal reoccurence (seasonal lags) [51, 59]. These models are known as pure seasonal autoregressive and moving average model ARMA(P,Q)$_S$, are expressed as:

$$\Phi_P(B^S)X_t = \Theta_Q(B^S)\epsilon_t, \tag{3.42}$$

where the operators $\Phi_P(B^S)$ and $\Theta_Q(B^S)$ are the seasonal autoregressive and seasonal moving average operators for orders P, Q, respectively, with seasonal period S, and they are expressed as:

$$\Phi_P(B^S) = 1 - \Phi_1(B^S) - \Phi_2(B^{2S}) - \ldots - \Phi_P(B^{PS}), \tag{3.43a}$$

$$\Theta_Q(B^S) = 1 - \Theta_1(B^S) - \Theta_2(B^{2S}) - \ldots - \Theta_Q(B^{QS}). \tag{3.43b}$$

Usually the seasonal and non–seasonal operators combined in a multiplicative seasonal autoregressive moving average model defined as ARMA $(p, q) \times (P, Q)_S$ and expressed as:

$$\Phi_P(B^S)\phi(B)X_t = \theta_Q(B^S)\theta(B)\epsilon_t. \tag{3.44}$$

The multiplicative seasonal autorgressive integrated moving average model (SARIMA) is defined as:

$$\Phi_P(B^S)\phi(B)\nabla^D{}_S\nabla^d X_t = \delta + \Theta_Q(B^S)\theta(B)\epsilon_t, \tag{3.45}$$

where ϵ_t is the Gaussian white noise. The general form is defined as ARIMA $(p,d,q)\times(P,D,Q)_S$. The polynomials $\phi(B)$, and $\theta(B)$ represent the autoregressive and moving average components of orders p,q respectively. The polynomials $\Phi(B)$, and $\Theta(B)$ are the seasonal autoregressive and seasonal moving average component with orders P,Q respectively. The ordinary difference component is $\nabla^d = (1-B)^d$, and the seasonal difference component is given by $\nabla_S^D = (1-B^S)^D$.

The SARIMA $(0,0,1)\times(1,0,0)_{12}$ model

A SARIMA model $(0,0,1)\times(1,0,0)_{12}$ is denoted as:

$$X_t = \epsilon_t + \epsilon_{t-1}\theta_{t-1} + \Phi X_{t-12}, \tag{3.46}$$

where $\mid\theta\mid< 1$, and $\mid\Phi\mid< 1$.

The autocovariance is calculated as:

$$\gamma(0) = \frac{1+\theta^2}{1-\Phi^2}\sigma_\epsilon^2, \tag{3.47}$$

and the autocorrelation as:

$$\begin{cases} \rho(12h) = \Phi^h, & h = 1,2,\ldots, \\ \rho(12h-1) = \rho(12h+1) = \frac{\theta}{1+\theta^2}\Phi^h, & h = 0,1,2,\ldots, \\ \rho(h) = 0, & \text{otherwise.} \end{cases} \tag{3.48}$$

3.6 Model Selection

The optimal time series model is usually chosen based on Akaike's Information Criterion, Akaike's Bias Corrected Information Criterion, and Bayesian Information

Criterion. These criteria measure the adequacy of the fit by balancing against the number of the parameters in the model [59].

Akaike's Information Criterion (AIC)

The Akaike's Information Criterion is estimated as follows:

$$AIC = 2k - 2\log(\widehat{L}), \tag{3.49}$$

where $\widehat{L}$ is the maximum likelihood estimator, and k is the number of the model's parameters [3].

The minimum AIC identifies the optimal model, yielding by the k. The selection of the model is implemented by minimizing the $\widehat{L}$. The $\widehat{L}$ decreases as the k increases.

Akaike's Bias Corrected Information Criterion (AICc)

The Akaike's Bias Corrected Information Criterion (AICc) is estimated as follows :

$$AICc = AIC + \frac{2k^2 + 2k}{n - k - 1} \tag{3.50}$$

where k is the number of the model's parameters, and n is the number of the observations [13].

The AICc is a corrected form of Equation (3.49), based on small distributional results or the linear regression.

Bayesian Information Criterion (BIC)

The Bayesian Information Criterion (BIC) is estimated as:

$$BIC = k\log(n) - 2\log(\widehat{L}), \tag{3.51}$$

where n is the number of the observations, k is the number of the parameters, and $\widehat{L}$ is the maximum likelihood [56]. The BIC is usually larger than AIC, therefore tends to choose models with small orders. In large samples the BIC choose the optimal order, while AICc choose the highest order in small samples.

3.7 Forecasting

The main goal of time series analysis is to predict the future evolution of the data, i.e the values X_{n+m}. The forecasting is implemented based on the known values of a time series $X_{1:n} = X_1, X_2, \ldots, X_n$. The above models are using to implemented the forecasts. For this section is though that the time series is stationary, and the parameters of the above models are known [51]. The conditional expectation $\mathbb{E}[X_{n+m} - g(X_{1:n})]^2$, where $g(X_{1:n})$ is a function of the observations, minimizes the mean square error predictor of X_{n+m}, which is expressed as:

$$X^t_{n+m} = \mathbb{E}[X_{n+m}|X_{1:n}]. \tag{3.52}$$

3.7.1 Forecasting AR(p)

Considering a AR(p) as in Equation (3.17) the expectation for forecast is expressed as follows:

$$\mathbb{E}[X_{n+k}|X_1, \ldots, X_n], \tag{3.53}$$

and the variance as:

$$\mathrm{Var}[X_{n+k}|X_1, \ldots, X_n], \tag{3.54}$$

where k is the step forecast.

The first step forecast is estimated as:

$$\widehat{X}_{n+1;n} = \phi_1 x_n + \ldots + \phi_p x_{n+1-p}, \tag{3.55}$$

and the k–step forecast is estimated as:

$$\widehat{X}_{n+k;n} = \phi_1 \widehat{X}_{n+k-1;1:n} + \ldots + \phi_p \widehat{X}_{n+k-p;1:n}. \tag{3.56}$$

Forecasting AR(1)

It is assumed a stationary autoregressive model first order AR(1), $X_t = \phi X_{t-1} + \epsilon_t$, where $\epsilon_t \sim \epsilon_n(0, \sigma_t^2)$. The conditional expectation at time $n+1$ is given by:

$$\mathbb{E}[X_{n+1}|X_1, \ldots, X_n] = \phi_1 x_n. \tag{3.57}$$

So the predictor is denoted as:

$$\begin{aligned} \widehat{X}_{n+1;1:n} &= \mathbb{E}[X_{n+1}|X_1, \ldots, X_n] \\ &= \mathbb{E}[\phi X_{n+k-1} + \epsilon_{n+k}|X_1, \ldots, X_n] \\ &= \mathbb{E}[\phi X_{n+k-1}|X_1, \ldots, X_n] \\ &= \ldots \\ &= \phi_1^k x_n. \end{aligned} \tag{3.58}$$

As it is observed the predictor is depended on the last observation and it tend to zero exponentially.

The confidence interval is based on:

$$Var[X_{n+k}|X_1, \ldots, X_n] = \left(1 + \sum_{j=1}^{k-1} \phi_1^{2j}\right) \sigma_\epsilon^2. \tag{3.59}$$

It should be noticed that as k tends to infinity, the forecast converges to zero and the conditional variance to σ_X^2.

3.7.2 Forecasting MA(q)

For simplicity is considered a MA(1), i.e. $X_t = \epsilon_t + \theta\epsilon_{t-1}$. The conditional expectation is $\mathbb{E}[X_{n+k;1:n}|X_1, \ldots, X_n]$ [21]. From the expectation is noticed that for $k > 2$ the predictor is equal to zero. So it is necessary the MA model to be expressed as AR(∞), i.e.:

$$\mathbb{E}_n = \sum_{j=0}^{\infty} (-\beta_1)^j X_{n-j}. \tag{3.60}$$

The predictor is estimated as:

$$\widehat{X}_{n+1;1:n} = \sum_{j=0}^{\infty} \widehat{\beta}_1(-\widehat{\beta}_1)^j X_{n-j}. \tag{3.61}$$

In general in MA(q) models, all the forecast for $k > q$, except the case $k \leqslant q$, will be equal to zero.

3.7.3 Forecasting ARMA(p,q)

In ARMA(p,q) models there is the same issue as in MA(q), so the predictor is estimated as:

$$\widehat{X}_{n+k;n} \sum_{i=1}^{p} \phi_i \, \mathbb{E}[X_{n+k-i}|X_{-\infty}^{n}] + \sum_{j=1}^{q} \theta_i \, \mathbb{E}[\epsilon_{n+k-j}|X_{-\infty}^{n}], \tag{3.62}$$

where the AR and MA conditional expectations are:

$$\mathbb{E}[X_t|X_{-\infty}^{n}] = \begin{cases} x_t & t \leqslant n \\ \widehat{X}_{t;1:n}, & t > n, \end{cases} \tag{3.63}$$

$$\mathbb{E}[\epsilon_t|X_{-\infty}^{n}] = \begin{cases} \epsilon_t & 0 < t \leqslant n \\ 0, & t > n. \end{cases} \tag{3.64}$$

3.7.4 Forecasting ARIMA(p,d,q)

It is assumed a times series X_t which is fitted with an ARIMA(p,1,q). The differences which are taken is first order, and the remainder is $Y_t = X_t - X_{t-1}$. The remainder follows as ARMA(p,q). Hence, the predictors are obtained as $\widehat{Y}_{n+1;1:n}, \ldots, \widehat{Y}_{n+k;1:n}$. [14]. The k–step forecast for the initial time series has a trend, based on $\widehat{X}_{n+1;1:n} = \widehat{Y}_{n+1;1:n} + X_n$. The predictors $\widehat{X}_{n+1;1:n}$ has to be integrated as:

$$\begin{aligned} &\widehat{X}_{n+1;1:n} = \widehat{Y}_{n+1;1:n} + X_n, \\ &\widehat{X}_{n+1;1:n} = \widehat{Y}_{n+2;1:n} + \widehat{X}_{n+1;1:n} = X_n + \widehat{Y}_{n+1;1:n} + \widehat{Y}_{n+2;1:n}, \\ &\vdots \\ &\widehat{X}_{n+k;1:n} = X_n + \widehat{Y}_{n+1;1:n} + \ldots + \widehat{Y}_{n+k;1:n}. \end{aligned} \tag{3.65}$$

As it noticed from the Equation 3.65, the k–step forecast for the initial data is the cumulative sum of the predicted terms of the differenced data. The prediction

interval is increasing indefinitely in respect of increasing horizon k.

3.7.5 Forecasting SARIMA(p,d,q)(P,D,Q)^S

In order to forecast a SARIMA model is necessary for the trend and the seasonal component to be removed. The trend and the seasonal components can be forecasted with the methods described above, based on the last observations. The remainder of the time series can be fitted with an ARMA model, and the forecast is implemented, as mentioned above. In forecasts must be added the predictors of the trend, and the seasonal component.

Chapter 4

Data Analysis

4.1 Study Area

The study area is the country of Netherlands, which is located in Northwestern Europe. The country comprises 12 provinces that border with Germany to the east, to Belgium to the south and the North Sea to the northwest, with maritime borders in the North Sea with Belgium, Germany, and the United Kingdom. The three islands of Caribbean Sea (Saint Eustatius, Saba, and Bonaire), along with Netherlands, consists a constituent country of the Kingdom of the Netherlands. The Caribbean islands are not considered in this study. The largest cities are Amsterdam (the capital city), Rotterdam, The Hague, Utrecht, and Eindhoven.

Netherlands is also known as one of the "Low Countries", because of the low elevation and flat topography. Only 50% of its land is higher than 1 meter above the sea level while 17% is below the sea level. The country extends over a surface area of 41543 km^2, with a population 17.30 million people as of November 2019. Netherlands is one of the most densely populated countries in the world, ranked as the 30th most densely populated country and the one of the largest exporters of food and agricultural products, due to fertile soil, mild climate, and intensive agriculture [48]. The land covers 33893 km^2, of the total area of the country while the rest (i.e. 7650 km^2) is covered by water. The lowest point of the Netherlands is Zuidplaspolder (at -7 m below sea level), and the highest point on European mainland is Vaalserberg (322.7 m above the sea level); including the Caribbean sea colonies, the highest point is the Mount Scenery on Saba (887 m above the sea level).

The data set implied measurements of potential wind generating capacity from 46 locations around the Netherlands. That is how a 1KW wind turbine woul produce over an hourly time interval. Raw data were wind speeds at 80-m elevation equal to the height of the GE wind turbine model. From wind speeds, the equivalent hourly wind power generation assuming a sigmoid power curve is estimated [61]. The geographical distribution of the stations is shown in Figure 4.1. Some of the stations are located offshore. The available data span the six-year period 2001–2006, ignoring the leap days. There are no gaps (missing points) in the dataset [61]. This study focuses on the analysis of the spatial and temporal features the average monthly wind power based on the installed power data.

For the temporal analysis, the monthly average power production in each location is calculated. Thus, a coarse-grained time series with 72 time instants (6 years $\times$ 12 months) is generated. In this study two locations for study are presented, specifically: the onshore station 1 , and offshore station 31. For the spatial analysis, the mean annual power production in each station is calculated.

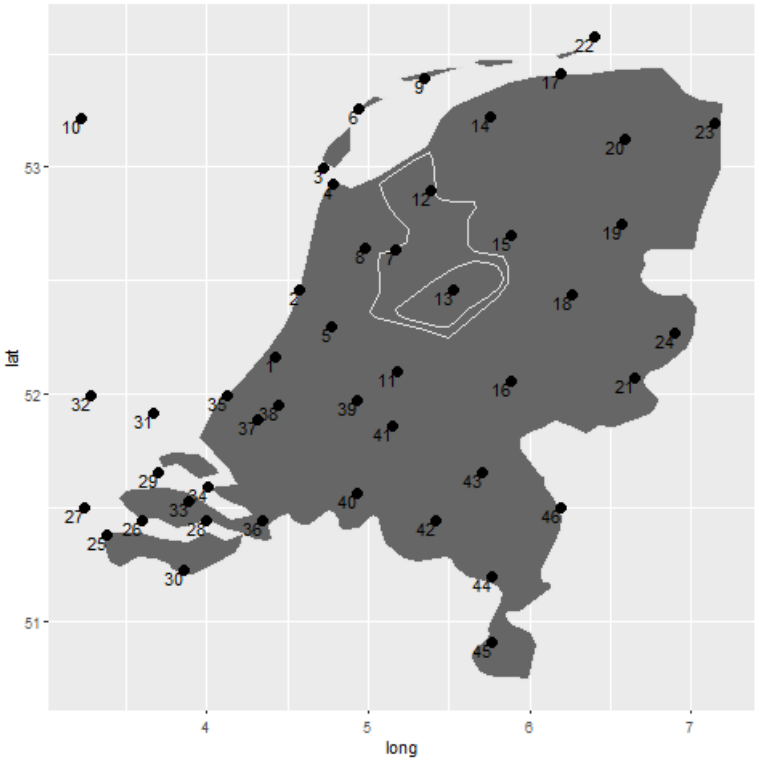


Figure 4.1: Map of the Netherlands showing the locations of the 46 wind power stations both onshore and offshore (black circles).

4.2 Temporal Analysis of Wind Power at Onshore Station

The time series of onshore station is shown in Figure 4.2, and its moments in Table 4.1. In Figure 4.2 it can be seen that the highest installed power production is observed in the latter months of the year. Hence the time series present an annual seasonality, which is described below.

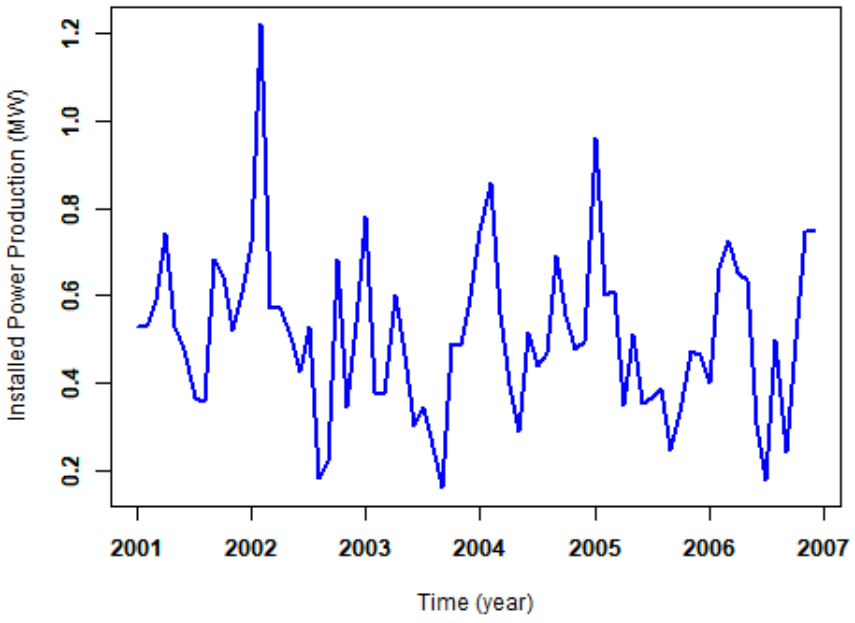


Figure 4.2: Time series of average monthly power production. The horizontal axis represents time (years: 2001–2006) and the vertical axis shows the installed power in MW.

Table 4.1: Summary statistics for wind power production at Onshore Station. "St. dev." stands for "standard deviation." All statistics are measured in MW except for "skewness" which is dimensionless.

Onshore Station	Mean	Min	Max	Median	St.dev.	Skeweness
Power produced	0.51	0.16	1.22	0.50	0.19	0.74

In order to determine the best probability distribution function to model the time series, the Weibull, lognormal, and normal models are tested. The best model is chosen using the maximum likelihood, which consists of finding the parameters that

maximize the log-likelihood, Akaike's information criterion (AIC), and Bayesian information criterion (BIC). The results for onshore station are shown in Table 4.2.

The Weibull distribution fits well the time series of 31 stations. The lognormal distribution provides a better fit for 11 stations, while the normal distribution is the best fit for 4 stations. The time series which are fitted with the lognormal distribution are transformed by calculating the logarithm of the values. For the time series that follow the Weibull distribution, we decided to not transform the data because the empirical distribution was already close to the normal distribution.

Table 4.2: Values of different information criteria for three probability distribution models: Weibull, lognormal and normal. AIC: Akaike's Information Criterion; LL: logarithm of the likelihood; BIC: Bayesian Information Criterion. The optimal model (Weibull) has the lowest values of AIC and BIC and the highest value of LL.

Distribution	**AIC**	**LL**	**BIC**
Weibull	**−34.29**	**19.14**	**−29.73**
lognormal	−33.51	18.75	−28.96
Normal	−33.01	18.51	−28.46

Figure 4.3 displays the probability density histogram with the theoretical Weibull pdf (top left), the Q-Q plot between the empirical data and the model (top right), the theoretical and empirical cumulative distribution functions (bottom left), and the respective probability (P-P) plot (bottom right). As is evidenced in the plots, the density plot of Weibull is close to the Normal distribution. Hence, we decided to not implement a transformation of the wind power data. The same approach is applied to the other stations of our study as well. The parameters of the Weibull distribution are presented in the Table 4.3.

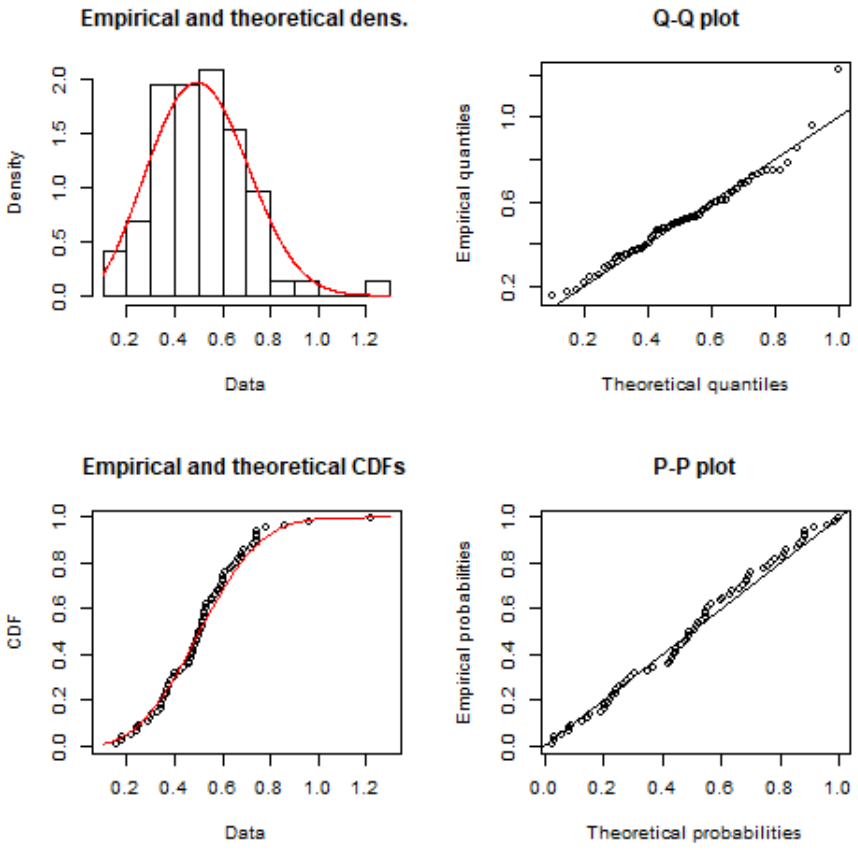


Figure 4.3: Top left: empirical probability density histogram fitted to the theoretical Weibull distribution. Top right: Q-Q plot of the the theoretical versus the empirical values. Lower left: empirical and theoretical cumulative distribution functions. Lower right: Probability (P-P) plot.

Table 4.3: Weibull distribution parameters (shape and scale) and their error estimates at onshore station based on maximum likelihood estimates.

	Shape	**Scale**
Onshore Station	2.86 ± 0.26	0.57 ± 0.02

The wind power time series at Onshore Station seems to be stationary, because the mean value does not vary in the entire time period. This was tested with the Augmented Dickey-Fuller test (see Section 3.3). The ADF test for the data admits the alternative hypothesis, i.e., that the time series is stationary. In Figure 4.4, it is shown that the wind power generation is autoregressive, due to the spike at order 1 and a cycling evolution of autocorrelations.

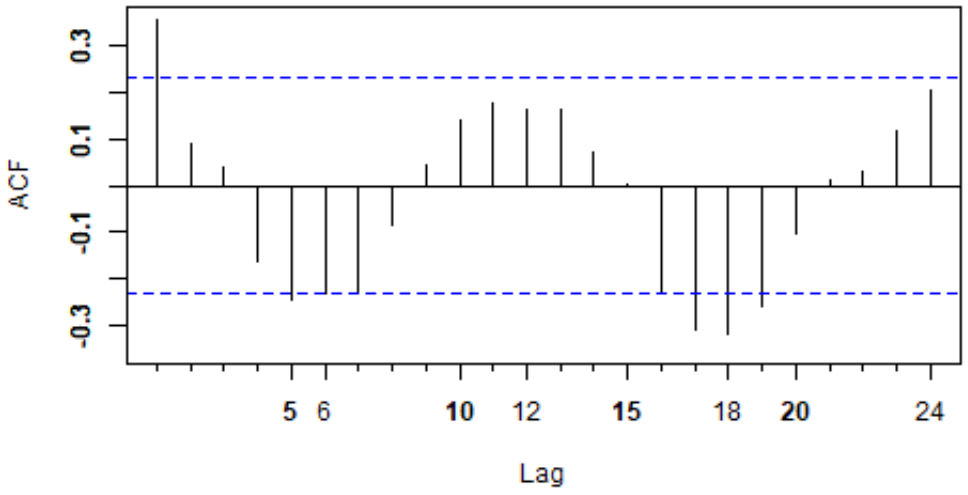


Figure 4.4: The autocorrelation function (ACF) for the average monthly wind power production at Onshore Station. The horizontal axis represents the time lag, while the vertical axis measures the autocorrelations.

4.2.1 Seasonal Decomposition

In order to find if the time series contains a seasonality component, the periodogram is calculated. Each step of the time series represents a month, so an annual periodicity corresponds to a period of 12 in the current data set. The frequency and the time period are reciprocals of each other, so a period of 12 months corresponds a frequency of 1/12 (or 0.083). As evidenced in the periodogram plot in Figure 4.5, the data exhibit annual seasonality.

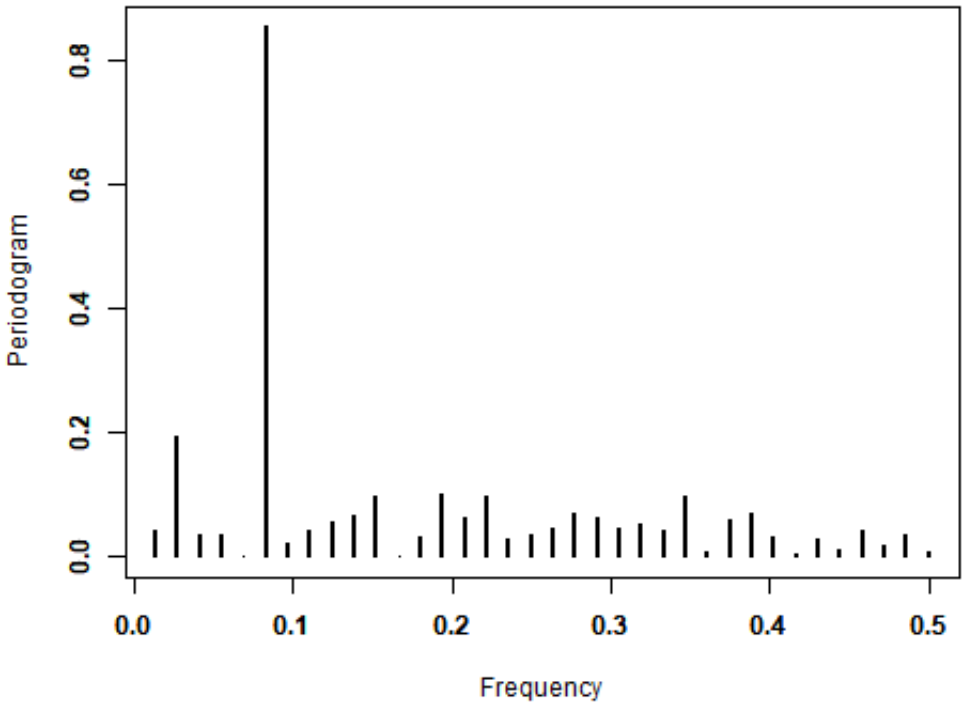


Figure 4.5: Periodogram of monthly average wind power at Onshore Station. The horizontal axis represents the frequency and the vertical axis the value of the periodogram.

The volatility of the data for seasonal variation is defined by estimating the box-plot of the squared instant wind power production in Onshore production. In Figure 4.6 each box represents the squared wind power for each month for the six years. As it is shown in Figure 4.6, some outliers are existed, especially in the winter months. Hence, we can conclude that some volatility is indeed present. The same pattern is observed in the other stations.

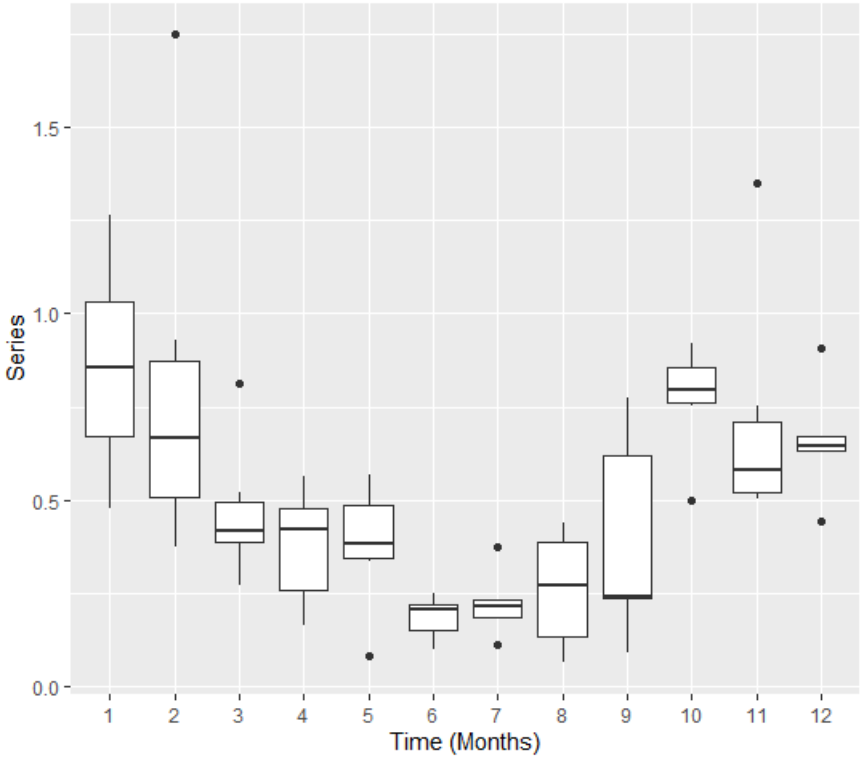


Figure 4.6: Box-plot of the squared instant wind power production. The horizontal axis represents the time in months. The vertical axis represents the squared instant wind power production.

For the prediction of average monthly power production in subsequent times it is necessary to model the seasonality inherent in the data. It is important to make a distinction between two conceptually- different types of seasonality, i.e deterministic and stochastic. In this study two approaches are used for this purpose. In the first approach, a SARIMA model is fitted to the original data (or to their logarithms for the lognormally distributed data) and is used to implement the prediction. In the second approach, a seasonal harmonic model [see Equation (4.1)] is estimated and extracted from the data, in order to remove the determinstc seasonality and then the residuals are fitted to a SARIMA model, for the sthochastic seasonality. This approach is implemented because SARIMA models cannot explicitly identify the deterministic process of seasonality. The second approach is used for onshore station.

$$s_t = \mu + A \sin\left(\frac{2\pi t}{T}\right) + B \cos\left(\frac{2\pi t}{T}\right), \tag{4.1}$$

where μ, A, B are constants, T is the period (one year), and t is the time. The

estimates for the seasonal model parameters of the average monthly power of onshore station are presented in Table 4.4. Based on these results, the cosine term is statistically significant at the 5% level while the sine term is not.

Table 4.4: Seasonal model parameters of average monthly wind power at Onshore Station. The Standard Error (SE) for a given variable is given by the Residual Standard Error divided by the square root of the sum of squares for the particular variable. The p-value is used to test the null hypothesis that the respective coefficient is zero.

coefficient	**estimate**	**p-value**	**SE**
μ	0.51	2×10^{-16}	0.02
A	0.04	0.09	0.03
B	0.14	2.81×10^{-7}	0.03

4.2.2 Estimation of SARIMA Model

In order to predict the average monthly wind power in the 12 months (Jan-2007 to Dec-2007) following the study period, a SARIMA model (p,d,q)(P,D,Q)(S) and the harmonic model determined by Equation (4.1) are used. Several SARIMA models were tested, including models with d=1, or D=1, but it is observed that there were remaining autocorrelations, and in some cases the Normal distribution didn't fit the residuals. Also the best model was chosen based on AIC. SARIMA models with d=1 or D=1 had higher AIC than the chosen models. As such, our model selection is confined within the SARMA family.

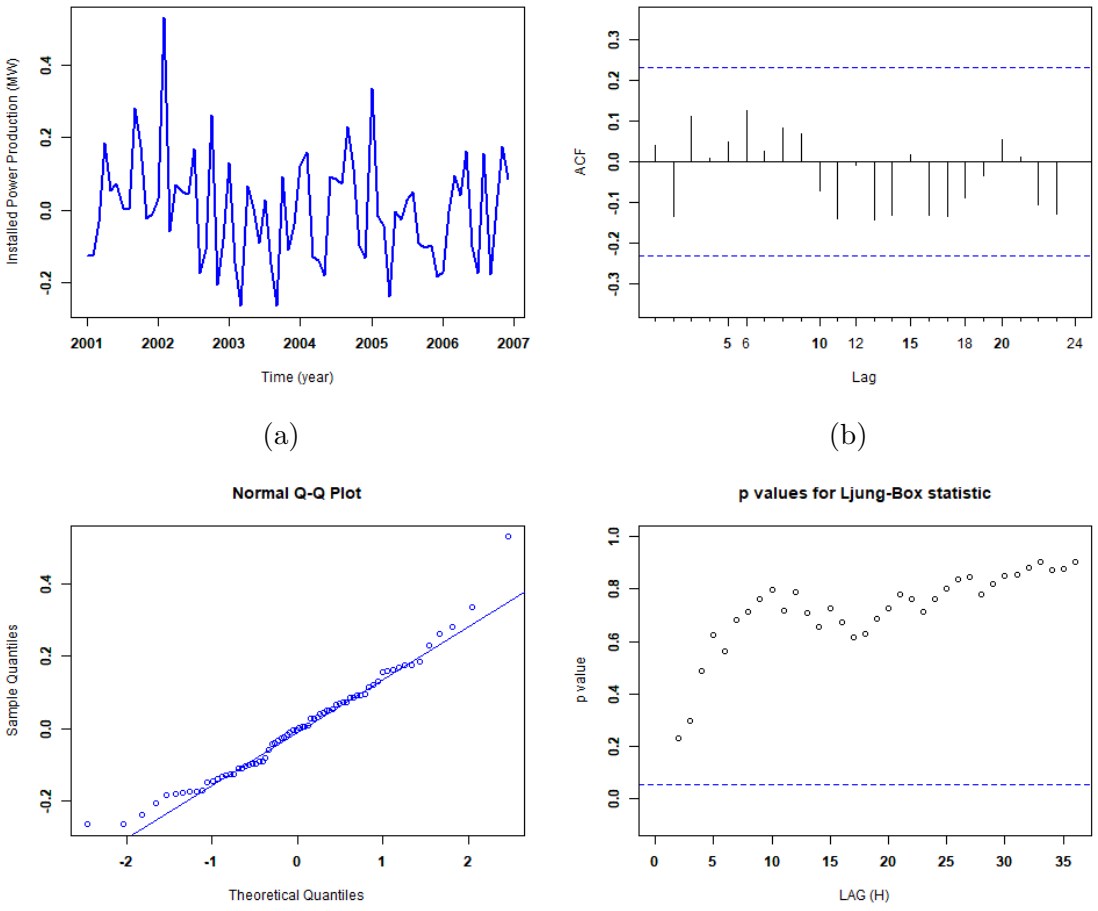
First, the harmonic model is subtracted from the data. Then, several SARIMA models are tested on the residuals. The best SARIMA model was chosen based on the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). The values of the information criteria for models of different orders are shown in Table 4.5. The optimal model is a SARIMA (0,0,0)(0,0,1)(12). The combination of the SARIMA model for the residuals and the harmonic model for the periodicity will be used for forecasting.

As shown in Figure 4.7a, there is no remaining correlation in the residuals, after the harmonic and the SARIMA model are fitted. In Figure 4.7b the values of autocorrelations are within the boundaries (blue lines), i.e., in the region where the autocorrelation is considered negligible (statisticallly insignificant). The absence of autocorrelation in the residuals is also shown in Figure 4.7d with p-values in Ljung-

Table 4.5: Results of information criteria for several SARIMA models (p,d,q)(P,D,Q)(S), where p is the AR order, d is the difference order, q is the MA order, P is the Seasonal AR order, D is the seasonal difference, and Q is the Seasonal MA order. AIC is the Akaike information criterion, AICc is the AIC with a correction for finite sample sizes, BIC is the Bayesian information criterion and the value. The best model is the one with the lowest values for the information criteria.

model	**AIC**	**AICc**	**BIC**
(0,0,0) (0,0,1) (12)	**−0.87**	**−0.87**	**−0.74**
(0,0,1)(1,0,2) (12)	−0.82	−0.81	−0.63
(1,0,1)(1,0,2)(12)	−0.80	−0.79	−0.58
(1,0,2)(2,0,2)(12)	−0.75	−0.75	−0.50

Box, in which they are above the critical threshold of 0.05. Also the residuals are close to the Gaussian distribution, according to the normal distribution plot. The estimated parameters for the SARIMA model are presented in Table 4.6.



(a) (b)

(c) (d)

Figure 4.7: SARIMA fitted model for installed power production. In Figure 4.7a is the time series of residuals of installed power production, in Figure 4.7b is the autocorreltation function, in Figure 4.7c is the normal distribution plot, and in Figure 4.7d are the p-values for the Ljung-Box statistic for the autocorrelation test.

Table 4.6: SARIMA model parameters for the residuals of installed power production of Onshore Station. The SE is the standard error of the estimates and the p-value is used in the context of null hypothesis testing of zero correlation in order to quantify the idea of statistical significance of evidence.

model	Estimated coefficient	SE	p-value
SMA(1)	-0.25	0.13	0.05

4.2.3 Power Production Forecasting

Using the fitted SARIMA model, the monthly average wind power production for the year (2007) following the study's period is predicted. The harmonic model,

which is removed from the initial time series, is added to the SARIMA predictions. In Figure 4.8 the original data (blue line) and the predictions (red line) are shown. Generally, in SARIMA models, the confidence intervals are estimated based on the normal distribution. In this thesis, although most of the stations are fitted to the Weibull distribution, the distributions are close to the normal. So for this thesis, the confidence intervals are calculated from the data's distribution. Thus, in Figure 4.8 the green lines represent the interval of two standard deviations around the prediction (95.45% confidence interval), while the black lines represent the 68.27% confidence interval (based on one standard deviation and the normal probability assumption).

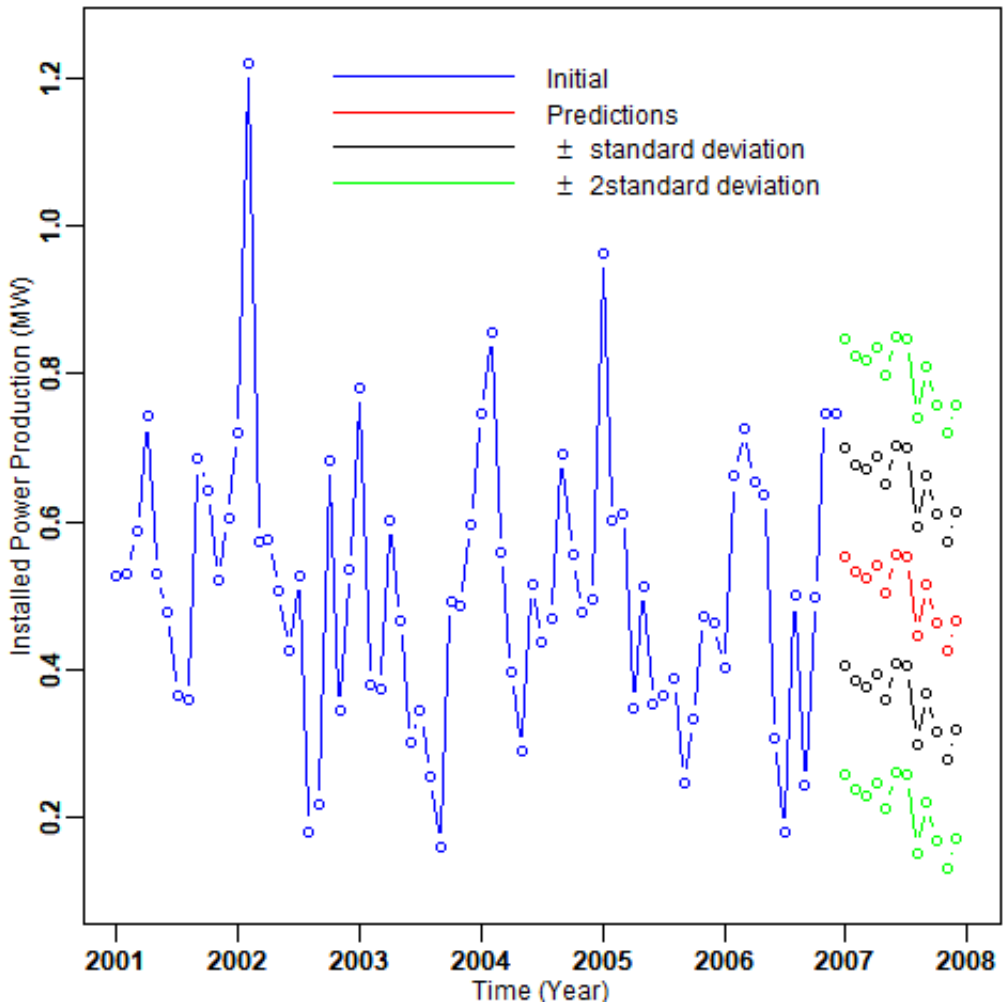


Figure 4.8: Predictions of monthly average wind power production for 2007 based on the data for the period 2001–2006. The blue line represents the original data, and the red line represents the SARIMA predictions. The green lines represent the interval of two standard deviations around the prediction (95.45% confidence interval), while the black lines represent the 68.27% confidence interval (based on one standard deviation and the normal probability assumption).

In Figure 4.8 is observed that after the three first predictions, the rest of the predictions tend to the harmonic model of the installed power production. This behavior is expected, due to the fitted model being a seasonal moving average model. The first prediction (January 2007) is much lower than the last data value (December 2006). Such behavior is observed for the entire time period (2001-2006), i.e for each year, the monthly power production increases during the later months of the year and sharply decreases in the first months of the year. In the last few months of the predictions this change is not captured sufficiently because the model trends to the mean.

4.2.4 Assessment of Model Performance

The method of cross-validation is used for the assessment of the temporal model's performance. One-step ahead forecast is used to validate the model's performance. Specifically, the first half of the time series (36 values) is chosen for the initial training set and then one-step ahead forecasting is used to predict the power production in the following month. This prediction uses the seasonal harmonic model and the SARIMA model chosen before. The coefficients of the SARIMA model are estimated from the training set. The prediction error is estimated as the deviation of the original series and the predictions from the cross-validation. The validation measures are presented in Table 4.7.

Table 4.7: Cross validation performance measures calculated through leave-one-out cross validation for the monthly average wind power of Onshore Station. ME: mean error; MAE: mean absolute error; RMSE: root mean squared error; ErrMin: minimum error; ErrMax: maximum error.

ME (MW)	**MAE (MW)**	**RMSE (MW)**	**ErrMin(MW)**	**ErrMax(MW)**
-0.006	0.12	0.14	-0.24	0.34

The low value of the mean error indicates the absence of bias. The root mean square error is 27% of the average monthly wind power in Onshore Station.

The prediction error for the 36 values (January 2004–December 2006) was calculated and the histogram is shown in Figure 4.9. In this station the errors follows the bimodal distribution, because there are two peaks in the histogram. This indicates that there are two groups of errors, which could mean that some predictions are overestimated or underestimated. As it shown in Figure 4.9 errors in the range between -0.2 MW and -0.1 MW have the highest frequency, thus we have underestimation.

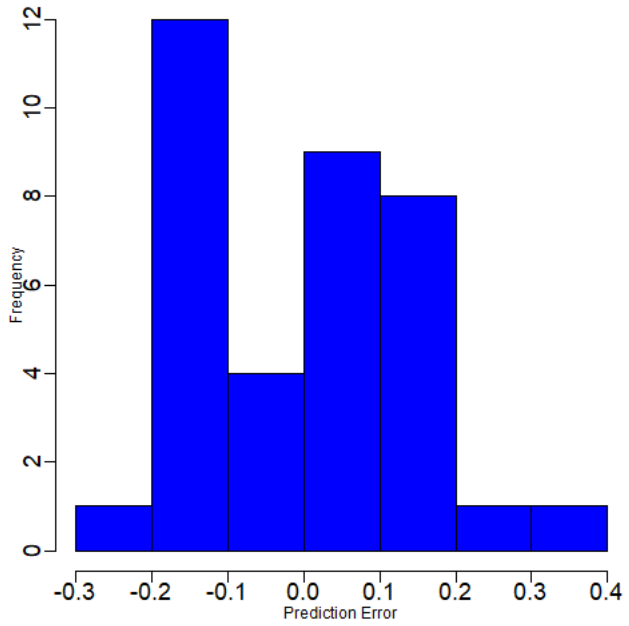


Figure 4.9: Histogram of the prediction error based on one-step ahead forecast cross-validation for Onshore Station.

4.3 Temporal Analysis of Wind Power at Offshore Station

The time series of station 31 (offshore station) is shown in Figure 4.10, and its moments in Table 4.8. In Figure 4.10 it can be seen that the highest installed power production is observed in the latter months of the year. Hence the time series present an annual seasonality, which is described below.

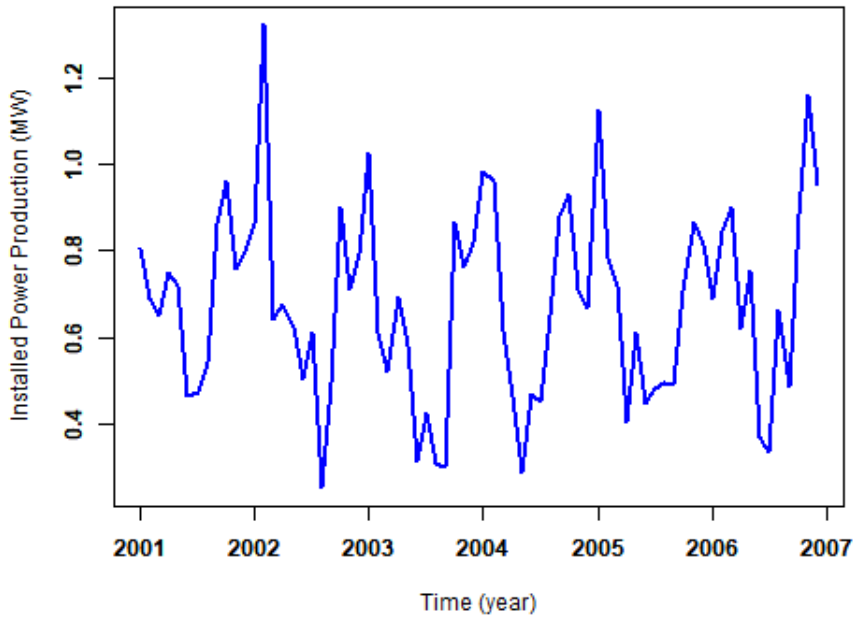


Figure 4.10: Time series of average monthly power production. The horizontal axis represents time (years: 2001–2006), and the vertical axis shows the installed power in MW.

Table 4.8: Summary statistics for the installed power production of Offshore Station. "St. dev." stands for "standard deviation." All the statistics are measured in MW except for skewness which is dimensionless.

Offshore Station	**Mean**	**Min**	**Max**	**Median**	**St.dev.**	**Skewness**
Installed Power Production	0.68	0.25	1.32	0.68	0.22	0.27

In order to determine the best probability distribution function to model the the time series, the same processes as in Onshore Station is implemented. The Weibull distribution fits well the time series of Offshore Station. The results of of information criteria for distribution fitting for Offshore Station are presented in Table 4.9.

Figure 4.11 displays the probability density histogram with the theoretical Weibull pdf (top left), the Q-Q plot between the empirical data and the model (top right), the theoretical and empirical cumulative distribution functions (bottom left), and the respective probability (P-P) plot (bottom right). As is evidenced in the plots, the density plot of Weibull is close to the normal distribution. Hence we decide to not implement a transformation of the wind power data. The same routine is

Table 4.9: Values of different information criteria for the three probability distribution models: Weibull, lognormal, and normal. AIC is Akaike's Information Criterion. LL is the logarithm of the likelihood. BIC is the Bayesian Information Criterion. The optimal model (Weibull) has the lowest values of AIC and BIC and the highest value of LL.

Distribution	AIC	LL	BIC
Weibull	**−10.21**	**7.10**	**−5.66**
LL	−6.21	5.10	−1.65
Normal	−9.45	6.72	−4.9

applied to the other years as well. The parameters for the Weibull distribution are presented in Table 4.10.

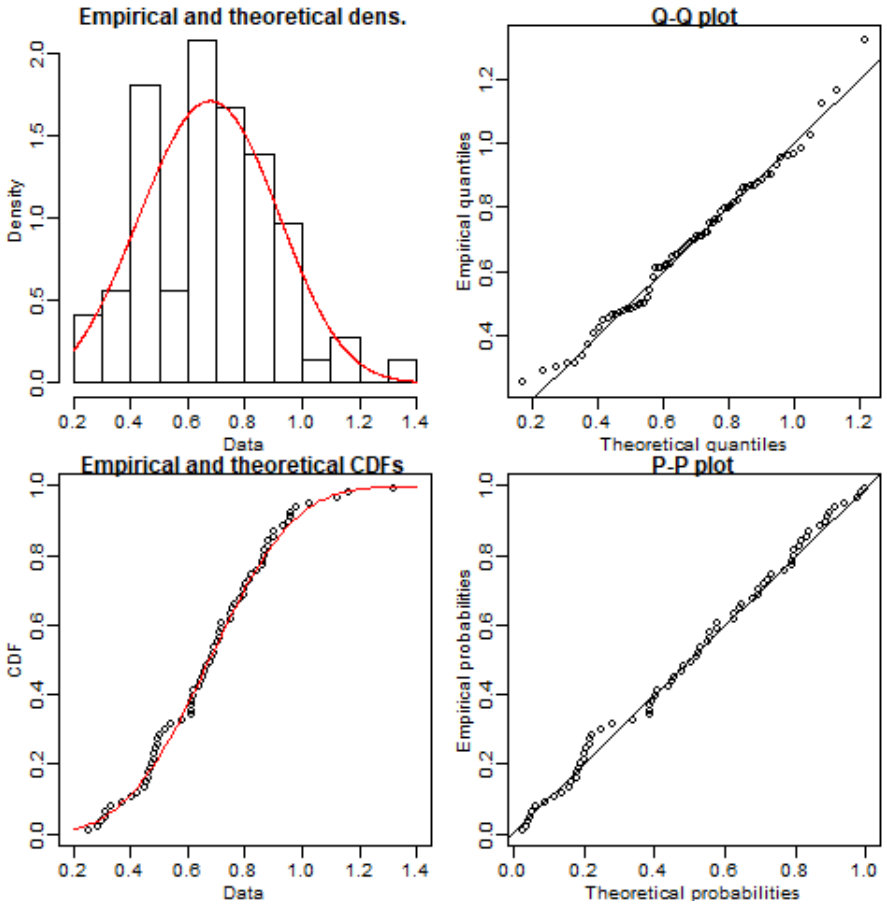


Figure 4.11: Top left: empirical probability density histogram fitted to the theoretical Weibull distribution. Top right: Q-Q plot of the the theoretical versus the empirical values. Lower left: empirical and theoretical cumulative distribution functions. Lower right: Probability (P-P) plot.

The wind power time series at Offshore Station seems to be stationary, because the mean value does not vary in the entire time period. This was tested with the ADF

Table 4.10: Weibull distribution parameters (shape and scale) and their error estimates at Offshore Station based on maximum likelihood estimates.

	Shape	**Scale**
Offshore Station	3.34 ± 0.3	0.76 ± 0.03

test (see section 3.3). The ADF test for the data admits the alternative hypothesis, i.e., that the time series is stationary. despite to the autocorrelation figure, which it seems that the autocorrelation is dependent from the lag, as it shown in Figure 4.12 .

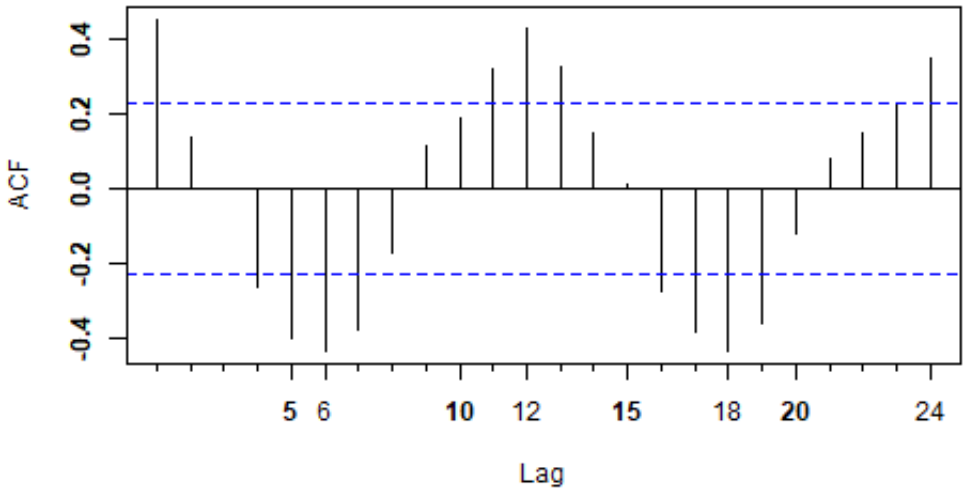


Figure 4.12: The autocorrelation function for the average monthly wind power production of Offshore Station. The horizontal axis represents the time lag, while the vertical axis measures the autocorrelations.

4.3.1 Seasonal Decomposition

In order to find if the time series contains a seasonality component, the periodogram is calculated. Each step of the time series represents a month, so an annual periodicity corresponds to a period of 12 in the current data set. The frequency and the time period are reciprocals of each other, so a period of 12 months correspond a frequency of 1/12 (or 0.083). As evidenced in the periodogram plot in Figure 4.13, the data exhibit annual seasonality.

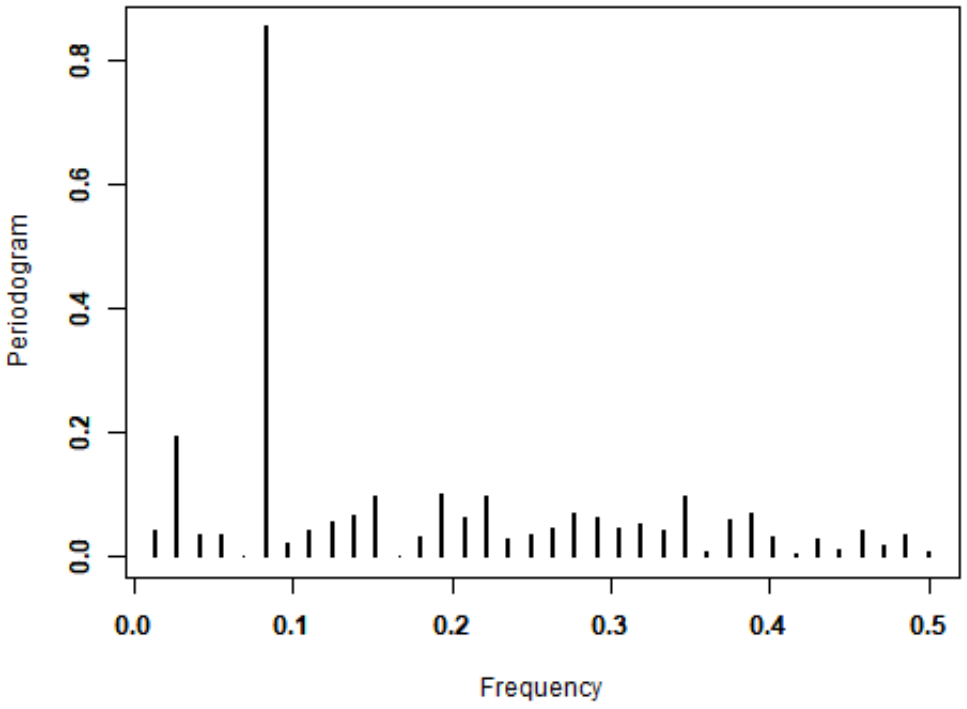


Figure 4.13: Periodogram of monthly average wind power at Offshore Station. The horizontal axis represents the frequency and the vertical axis the value of the periodogram.

The volatility of the data for seasonal variation is defined by estimating the box-plot of the squared instant wind power production in Offshore production. In Figure 4.14 each box represents the squared wind power for each month for the six years. As it is shown in Figure 4.14, some outliers are existed, especially in the winter months. Hence, we can conclude that some volatility is indeed present. The same pattern is observed in the other stations.

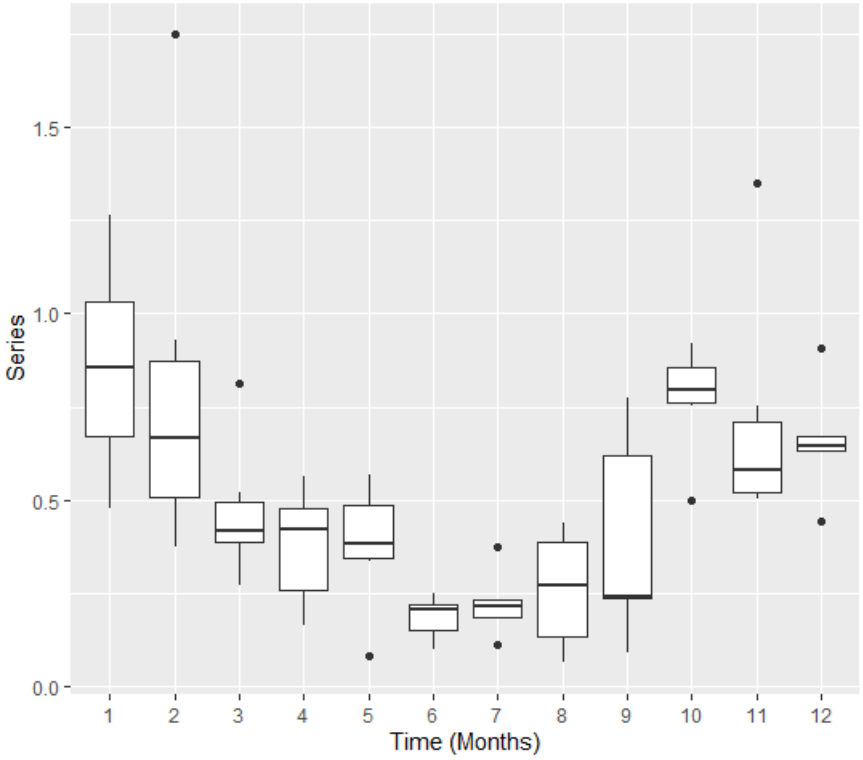


Figure 4.14: Box-plot of the squared instant wind power production. The horizontal axis represents the time in months. The vertical axis represents the squared instant wind power production.

In order to model the seasonality inhernt in the data at Offshore Station, the first decomposition method is applied as described in Section 4.2.1. As such, the time series is decomposed simply by fitting a SARIMA model.

4.3.2 Estimation of SARIMA Model

In order to predict the average monthly power in the 12 months (Jan 2007 to Dec 2007), following the study period, a SARIMA model (p,d,q)(P,D,Q)(S) is fitted to the original data. Several SARIMA models are tested, and the best was chosen based on the Akaike information criterion, and the Bayesian information criterion. The values of the information citeria for the models of different orders are shown in Table 4.11. The optimal model is a SARIMA (0,0,1)(1,0,1)(12). The plot of the model is shown in Figure 4.15.

As it shown in Figure 4.15a there is no remaining correlation in residuals, after the SARIMA model is fitted. In Figure 4.15b the values of autocorrelations are within the boundaries (blue lines), i.e in the region where the autocorrelation is considered

Table 4.11: Results of information criteria for several SARIMA model (p,d,q)(P,D,Q)(S), where p is the AR order, d is the difference, q is the MA order, P is the Seasoanl AR order, D is the seasonal difference, and Q is the Seasonal MA order. The AIC is the Akaike information criterion, the AICc is the AIC with a correction for finite sample sizes, and BIC is the Bayesian information criterion. The best model is the one with the lowest values.

Model	**AIC**	**AICc**	**BIC**
(1,0,0) (1,0,1) (12)	−0.37	−0.37	−0.25
(0,0,1)(1,0,1) (12)	**−0.47**	**−0.46**	**−0.31**
(1,0,0)(2,0,0)(12)	−0.46	−0.46	−0.31
(1,0,2)(2,0,2)(12)	−0.41	−0.38	−0.18

negligible. The absence of autocorrelations in residuals is also shown in Figure 4.15d with p-values in Ljung-Box, in which they are above the critical threshold of 0.05. Also the residuals are close to the normal distribution, according to the normal distribution plot. Hence is thought that the chosen SARIMA model is appropriate for forecasting. The estimated parameters for the SARIMA model are presented in Table 4.12.

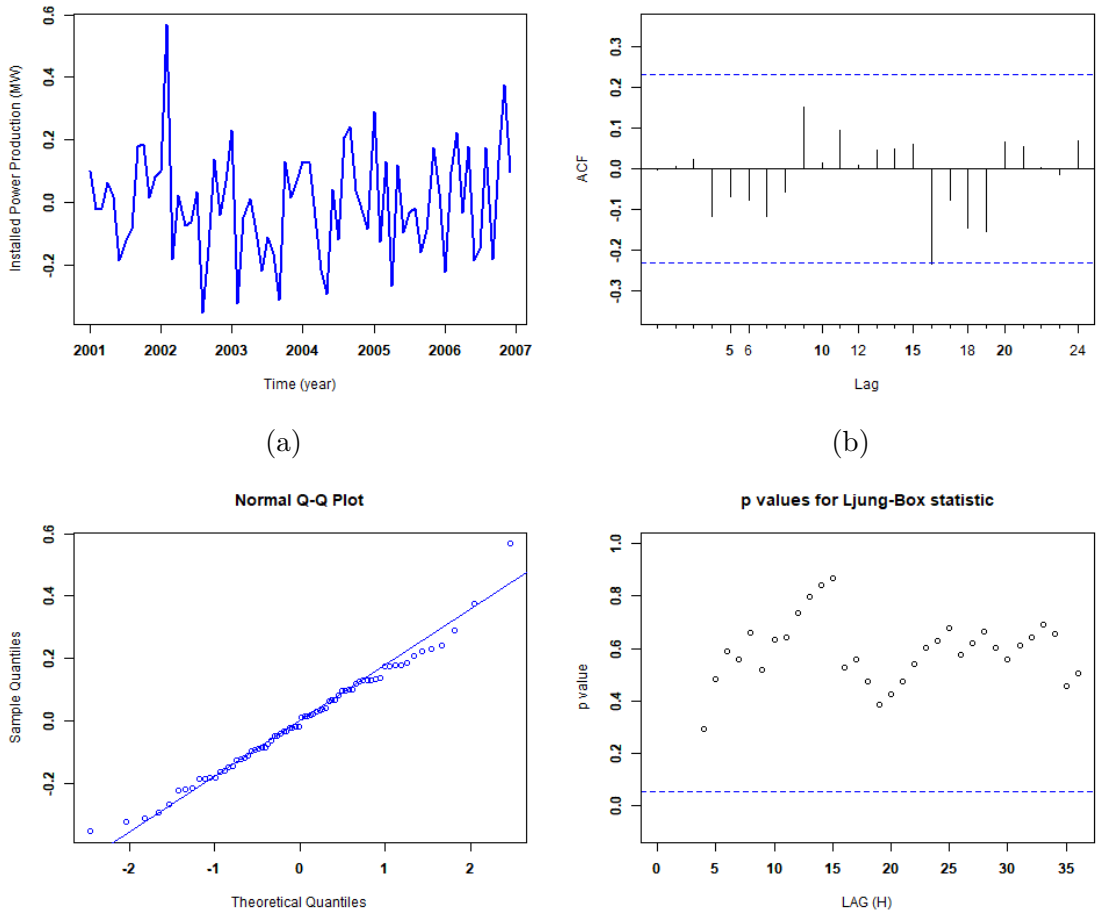


Figure 4.15: SARIMA fitted model for average monthly wind power. In Figure 4.15a is the time series of residuals of installed power production, in Figure 4.15b is the autocorreltation function, in Figure 4.15c is the normal distribution plot, and in Figure 4.15d are the p-values for the Ljung-Box statistic for the autocorrelation test.

Table 4.12: SARIMA model parameters for the residuals of installed power production of Offshore Station. The SE is the standard error of the estimates and the p-value is used in the context of null hypothesis testing of zero correlation in order to quantify the idea of statistical significance of evidence.

Model	Estimated coefficient	SE	p-value
MA(1)	0.28	0.13	0.03
SAR(1)	0.94	0.13	0.00
SMA(1)	−0.73	0.30	0.02

4.3.3 Power Production Forecasting

Using the fitted SARIMA model, the monthly average wind power production for the year (2007) following the study's period is predicted. In Figure 4.16 the original data (blue line) and the predictions (red line) are shown. The green and black lines in the graph represents the confidence intervals which correspond to a range of one and two standard deviations respectively. Generally, in SARIMA models, the confidence intervals are estimated based on the normal distribution. In this thesis, although most of the stations are fitted to the Weibull distribution, the distributions are close to the normal. So, the confidence intervals are calculated from the data's distribution.

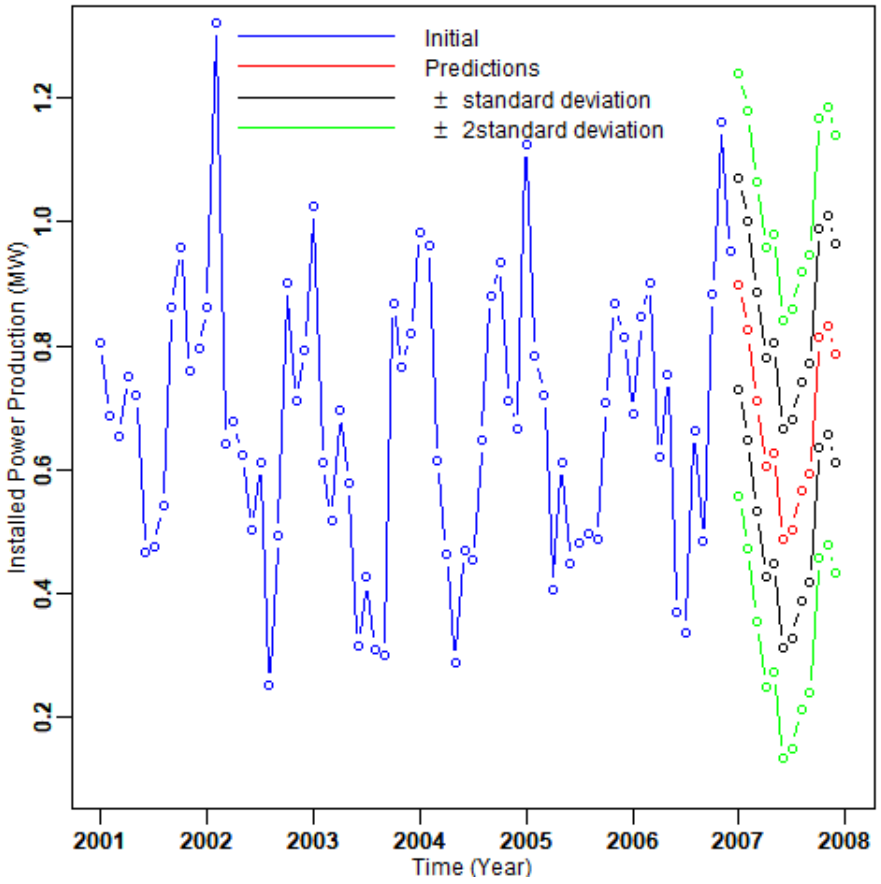


Figure 4.16: Predictions of monthly average wind power production for 2007 based on the data for the period 2001–2006. The blue line represents the original data, and the red line represents the SARIMA predictions. The green lines represent the interval of two standard deviations around the prediction (95.45% confidence interval), while the black lines represent the 68.27% confidence interval (based on one standard deviation and the normal probability assumption).

In Figure 4.16, it can be observed that the periodicity of the original data, also

exists in the predictions. Same as in the data (2001–2006), in the middle of the year (2007) the predictions are lower than in early and later months. The estimated confidence intervals have a high range, so is assumed that some of the prediction could be underestimating or overestimating the production. This pattern could be due to the low order of the SARIMA model.

4.3.4 Assessment of Model Performance

The method of cross-validation is used for the assessment of the temporal model's performance. One-step ahead forecast is used to validate the model's performance. As in Onshore Station the first half of the time series (36 values) is chosen for the initial training set and then one-step ahead forecasting is used to predict the power production in the following month (January 2007). This prediction uses the SARIMA model chosen before. The coefficients of the SARIMA model are estimated from the training set. The prediction error is estimated as the deviation of the original series and the predictions from the cross-validation. The validation measures are presented in Table 4.13.

Table 4.13: Cross validation performance measures calculated through the leave-one-out cross validation for the monthly average installed power production of the Offshore Station. ME: mean error; MAE: mean absolute error; RMSE: root mean squared error; ErrMin: minimum error; ErrMax: maximum error.

ME (MW)	**MAE (MW)**	**RMSE (MW)**	**ErrMin(MW)**	**ErrMax(MW)**
0.014	0.15	0.17	− 0.29	0.40

The low value of mean error ensures the absence of bias. The root mean square is 26% of the average monthly wind power in Offshore Station. The accuracy of the model is comparable to similar work for short-term predictions using either wind speed forecasts and the power curve, or wind power forecasts directly. In Figure 4.17 the histogram of the prediction errors of the mean monthly power production for Offshore Station is presented. Errors in the range between 0.1 MW and 0.2 MW have the highest frequency, while the highest error values (in the range 0.3–0.4 MW) are the least frequent.

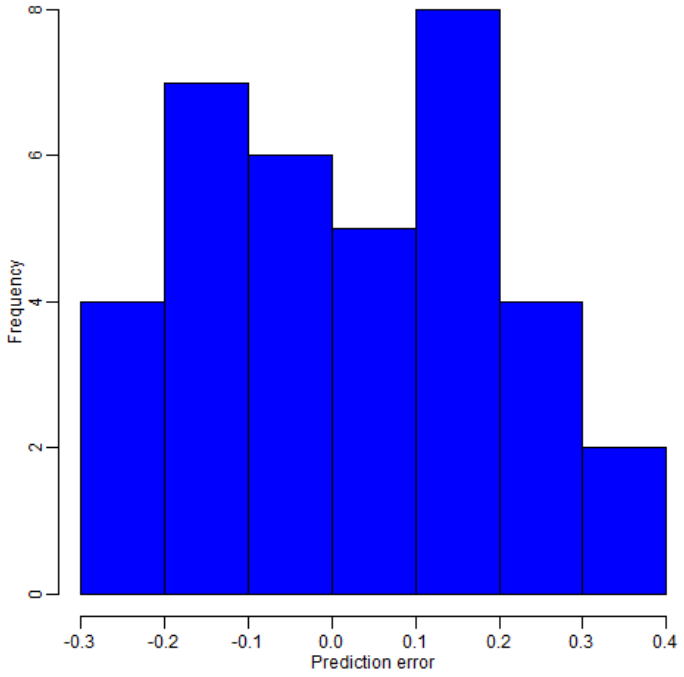


Figure 4.17: Histogram of the prediction errors based on one-step-ahead forecast cross-validation for Offshore Station.

4.4 Comparison of the two Stations

Studying the two stations (Onshore Station , Offshore station) some differences are readily apparent. In Onshore station it is necessary to remove a harmonic model first and then fit the SARIMA model, in order to extract the seasonality. In contrast, in Offshore Station the SARIMA model is fitted directly to the data which proves to be enough to remove the seasonality. Furthermore, the SARIMA model of Offshore Station returns better p-values than Onshore Station. The validation measures inOnshore Station are better than those of Offshore Station, except the RMSE. In Onshore Station the RMSE is 0.14 MW or 27% of the mean value and in Offshore Station is 0.17 MW or 25% of the mean value. Also the prediction in Onshore Station tends to mean value very fast, due to the simple seasonal moving average model. Offshore Station is less smooth because the SARIMA model involves a seasonal autoregresssion. In conclusion the differences may be due to the offshore station being unaffected by the terrain or the topography of the area. Also, the implementation of the method (removing a harmonic model before SARIMA for the onshore station) may also affect the results. The accuracy of the model is comparable to similar work for short-term predictions using either wind speed forecasts and the power curve or wind power forecasts directly. Jing Shi et.al. use data from

the supervisory control and data acquisition (SCADA) system of an offshore wind turibine, rated at 2 MW. The data are collected during the period of December 2009–February 2010, with time lag of 10 min. Their estimated RMSE is equal to 0.22 MW [36]. David Barbosa de Alencear, et.al use historical series of the meteorological variables from national organization system of environmental data (SONDA) of the National Institute of Space Research (INPE). The measured data began from 1 January 2004 and ends to 31 May 2017, with time lag 1 minute. They calculate the wind power from the power curve, in different time scales [4].

4.5 Spatial Analysis

In this section, spatial analysis is used for the 46 stations. The data used for this analysis and subsequent interpolation are the annual mean installed power production. Only the first year (2001) is presented in this section and the rest are in the Appendix. Summary statistics of the annual installed power for year 2001 are presented in Table 4.14.

Table 4.14: Summary statistics for annual mean of the installed power production for the year 2001.

2001	Mean	Min	Max	Median	St.dev.	Skewness
Installed Power Production (MW)	0.50	0.16	0.93	0.47	0.19	0.35

Although data fits well the normal distribution, the best fit is the Weibull distribution. To choose the best model of the distribution, three models were tested: normal, lognormal, and Weibull. The best fit was chosen using the maximum likelihood, which consists of finding the parameters that maximize the log-likelihood (LL). So, according to the AIC, log maximum likelihood and BIC the best model is chosen, and the results for each fit are shown in Table 4.15.

Table 4.15: Values of information criteria for the three distributions (Weibull, lognormal, and normal). The AIC is the Akaike's Information Criterion, the LL is the logarithm of the likelihood, and the BIC is the Bayesian Information Criterion. The optimal model is the one that has the lowest value of AIC or BIC. Low AIC and BIC values correspond to LL values.

Distribution	AIC	LL	BIC
Weibull	**−19.84**	**11.92**	**−16.19**
lognormal	−18.70	11.34	−15.04
Normal	−17.79	10.90	−14.13

In Figure 4.18 the pdf is shown for the empirical values and theoretical Weibull distribution, the Q-Q plot with the theoretical and empirical values, the theoretical and empirical values of the cumulative density function, and the probability plot. As can be seen from the figures, the density plot of Weibull is close to the Normal distribution, so it is not necessary to implement a transformation to the data. The same routine is applied to the other years as well. In Table 4.16 the parameters of the Weibull distribution are presented for each year.

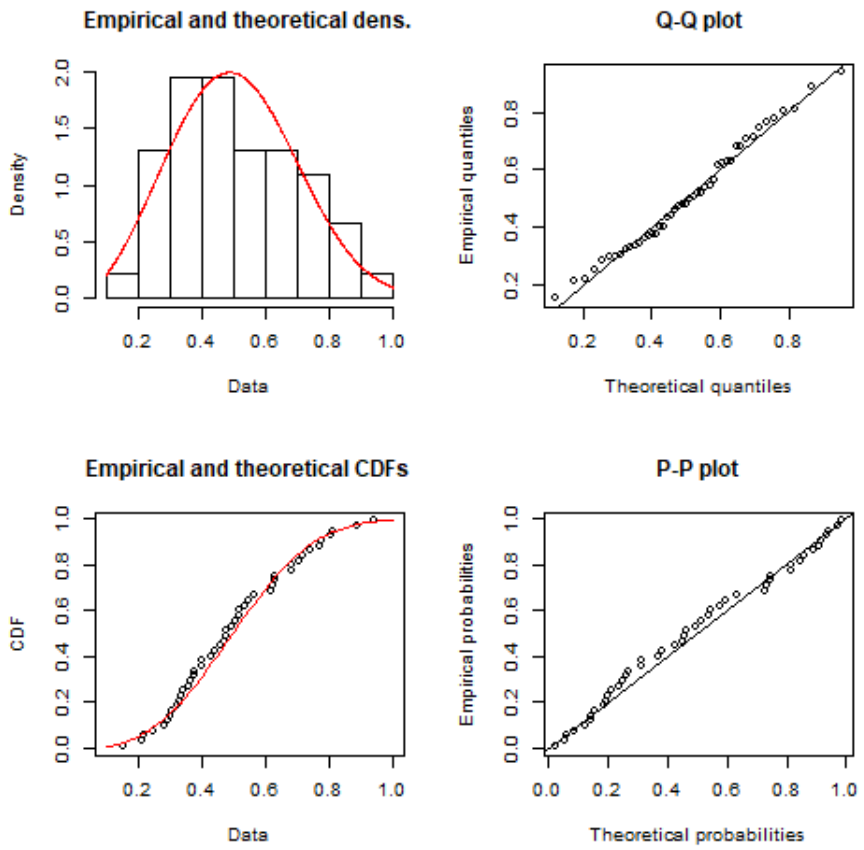


Figure 4.18: Top left: Histogram of empirical values and the theoretical Weibull probability density function are shown. Top right: Q-Q plot of the theoretical and the empirical values. Bottom left: empirical cumulative distribution function (cdf) and the theoretical Weibull cdf. Lower right: Weibull probability plot.

Table 4.16: Parameters (shape and scale) of the Weibull distribution, and their standard error for the time period 2001–2006. The estimation method is the maximum likelihood.

Year	Shape	Stand. Error	Scale	Stand.Error
2001	2.87	0.33	0.57	0.03
2002	3.38	0.39	0.58	0.03
2003	2.78	0.32	0.48	0.03
2004	3.15	0.36	0.56	0.02
2005	2.66	0.31	0.52	0.03
2006	3.02	0.35	0.57	0.03

4.5.1 Variogram Analysis

In order to construct a spatial model of the installed power production, it is necessary to model a suitable variogram model to the data. First, the experimental variogram is estimated, using the Equation (2.15). Afterwards, a suitable theoretical variogram is fitted to the experimental. Three models were tested: the spherical, exponential and gaussian models (see section 2.6.3). In order to define the best fit the minimum sum of weighted squared error was used (Equation (4.2)). As shown in Table 4.17, the best fit is the spherical model. The spherical model is often preferred in most studies, in order to implement Ordinary Kriging. The spherical variogram model is compact ($\gamma(r > h) = \sigma^2$), so it is less computationally intensive. Also the spherical variogram model is used for wind data across a surface, because it accounts for a progressive decrease in spatial autocorrelation, as characteristic in wind storms [20].

$$\epsilon^2 = \sum_{i=1}^{L} \{\gamma(\mathbf{r_i}) - \widehat{\gamma}(\mathbf{r_i})\}^2 w_i, \tag{4.2a}$$

$$w_i = \frac{N_i}{r_i^2}. \tag{4.2b}$$

In Equation (4.2) L is the number of the lags of the experimental variogram, w_i is the weight for the lag $i = i, \ldots, L$, $\mathbf{r_i}$ is the lag vector, $r_i = \mid \mathbf{r_i} \mid$ is the Euclidean distance, and N_i is the number of the point pairs.

Table 4.17: Sum of squared errors between the empirical and the theoretical variogram models. The total error for each model is equal to the sum of the squared differences between the values of empirical and the respective theoretical variogram model. The best fit is the one with the lowest error.

Model	Squared Error
Exponential	0.007
Spherical	**0.0002**
Gaussian	0.03

In Figure 4.19 the experimental variogram with the theoretical spherical model are presented. The model parameters estimated by minimizing the error in Equation (4.2) are presented in Table 4.18. As shown in Figure 4.19 the model has good agreement with the experimental variogram in distances less than 100km. Since the average minimum distance between stations is 34.5 km, the fit is adequate.

Table 4.18: Parameters of the optimal spherical variogram model for the installed power production. σ^2 is th variance, ξ is the correlation length, and c_0 is the nugget effect. The parameters are estimated by minimizing the error function given by the Equation 4.2.

$\sigma^2(\mathbf{MW^2})$	$\xi(\mathbf{km})$	$\mathbf{c_0}(\mathbf{MW^2})$
0.046	216.02	0.004

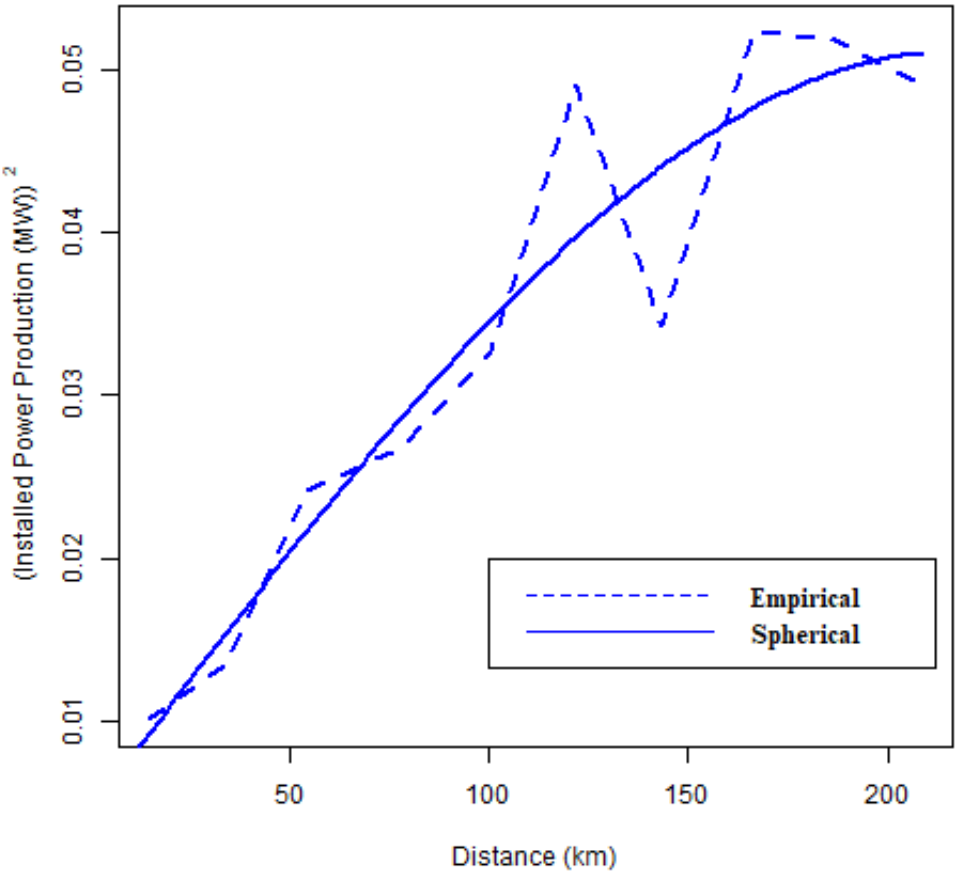


Figure 4.19: Experimental variogram (dashed line), and theoretical Spherical model (continuous line), using the Equation (2.20). The horizontal axis is the lag distance r in km, and the vertical distance represents the variogram values for installed power production, for n each lag. The estimated parameters are nugget $c_0 = 0.004$ MW^2, variance $\sigma^2 = 0.046$ MW^2, and correlation length ξ=216.02 km. The extent of the distance shown in this figure is equal to the correlation length.

4.5.2 Ordinary Kriging

For the visualization of the distribution of the annual mean installed power production, ordinary kriging (OK) is used. As shown above, the data are following the Weibull distribution which is close to the Gaussian distribution in this case. The interpolation is performed on a grid of 300×300 cells. The dimensions of each cell are 1.0 km $\times$1.1 km. Afterwards, a mask with the boundaries of Netherlands is applied on the grid for both the kriging predictions map and the kriging variance maps. The final kriging prediction map is presented in Figure 4.20. For four of the offshore stations (stations 10, 27, 31, 32), a box of 7×7 is used so that the prediction around the stations is more easy to discern visually.

As shown in the map, the highest values are observed in the West area of the map, due to the stations that are near to the sea. In those areas the wind is stronger than in the Eastern area as such wind power generation is higher. Finally, the mean value of the predicted installed power production does not differ significantly through the years, i.e the range varies from 0.43 MW to 0.53 MW. So it is assumed that the wind's annual mean velocity is constant over the years, with the resulting constant production for each station. This is beneficial for a wind park.

In Figure 4.21, the kriging variance is shown. The square root of the variance for each cell gives the standard deviation around the prediction value which is a measure for the prediction's uncertainty. The nugget effect is 0.004 MW^2. The variance near each station is equal to nugget, and is increasing as the distance increases. Since the configuration of the stations is more sparse further from the shore, the kriging variance is higher in the East.

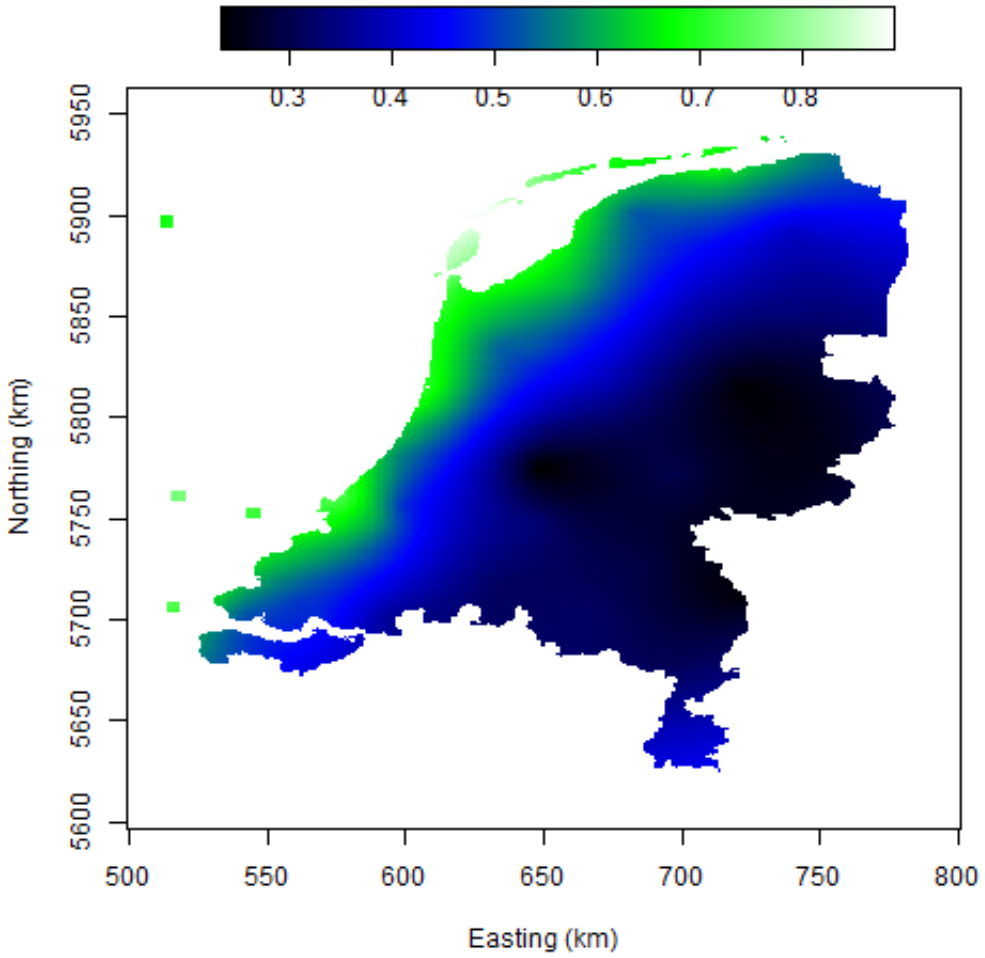


Figure 4.20: Map of the estimated annual mean installed power production for 2001, based on the spherical variogram model. The horizontal axis represents the Easting (km), and the vertical axis represents the Northing (km) coordinates.

In Figure 4.21, the kriging variance is shown. The highest variance is observed in West area, wherein the installed power production is higher than the East. The highest variance is observed in the west area of map, due to the stations proximity to the sea. As such, there is higher estimation of the installed power production there. In close proximity to the stations, kriging variance is very low.

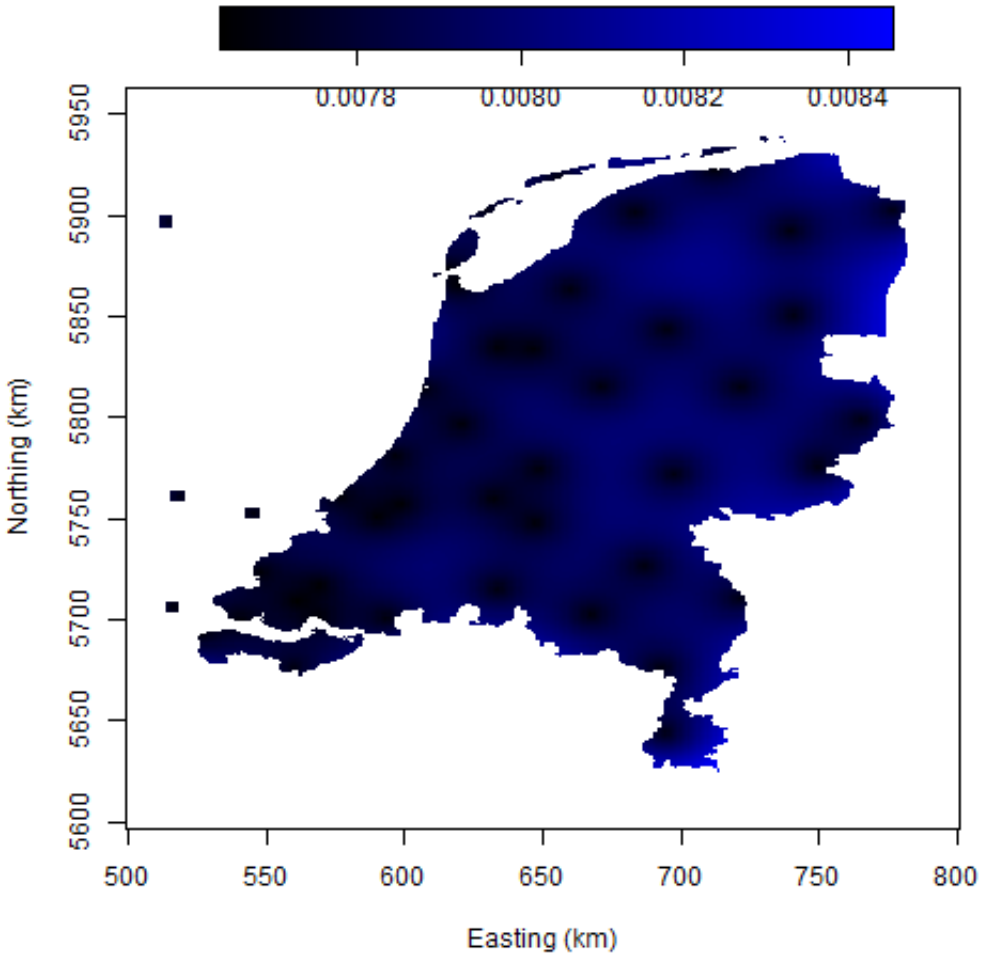


Figure 4.21: Map of the estimated variance of the mean installed power production for 2001, based on Spherical variogram model. The horizontal axis represents the Easting measured in kilometers, and the vertical axis represents the Northing measured in kilometers. The values are in MW^2.

4.5.3 Cross-Validation Analysis

The leave-one-out cross-validation method is used for the assessment of the model's performance. The results of the sample and predicted mean power production values for year 2001 are presented in the bar-plot of Figure 4.22 for each station. From the bar-plot, there's evidence that the model slightly underestimates the wind power production of stations in the North sea, or on the North Sea Coast. This is further investigated in the in the bar-plot of Figure 4.23 (North Sea stations: 2–4, 6, 9, 10, 17, 22, 25, 27, 29, 31, 32, 35). The mean error between the predictions and the sample for the north sea stations is -0.055 MW representing a slight underestimation by the model. This underestimation could be the result of kriging smoothing.

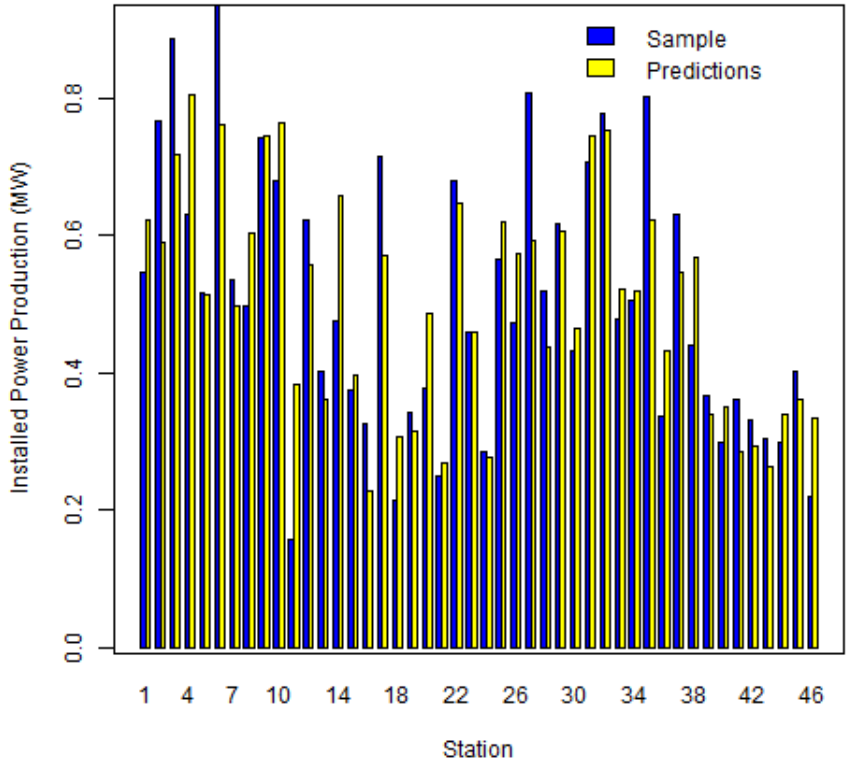


Figure 4.22: Estimated (yellow) and sample (blue) values for the year 2001, using leave-one-out cross-validation. The horizontal axis shows the number of station and the vertical axis represents the power production (MW) for both the original sample values (blue), and the predicted values (yellow).

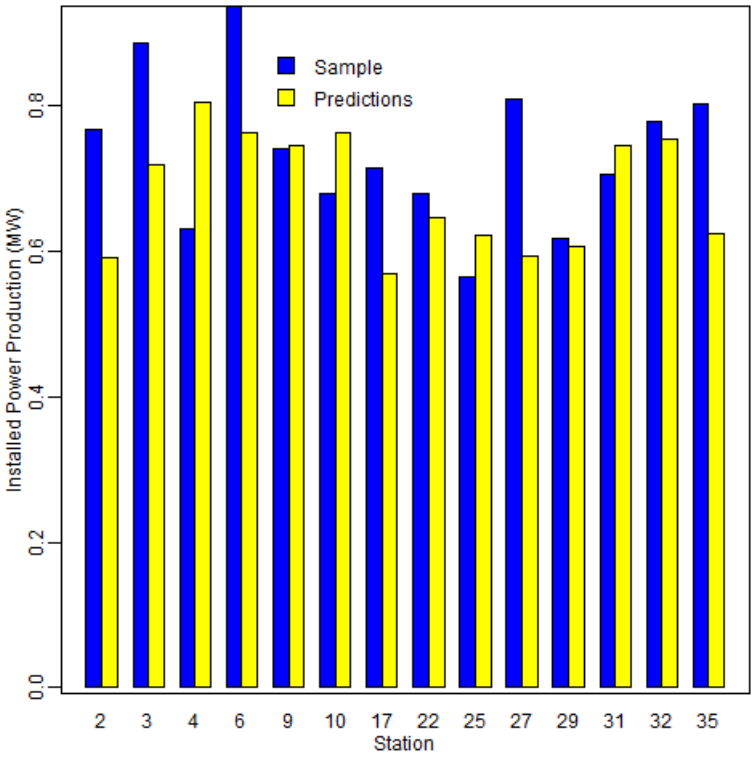


Figure 4.23: Estimated (yellow) and sample (blue) values of coastal stations for year 2001, using one leave-one-out cross-validation. The horizontal axis shows the number of station and the vertical axis represents the power production (MW) for both the original sample values (blue), and the predicted values (yellow).

The validation measures are presented in Table 4.19. The low value of mean error indicates the absence of bias. The linear correlation between the predicted values and the sample values (calculated with Pearson's ρ) is high and the RMSE and MAE are low. As such the model has a good performance despite the slight underestimation of the North Sea stations. The accuracy of the model is comparable to similar work using different methods of kriging such as Augmented Kriging (which is based on Universal Kriging) [31]. In the study of Jin Hur, et.al. the McCamey area CREZ data set is used. This data consists of wind farm outputs with 1 minute time resolution during January to September 2009. They predict weekly outputs for wind power, and the estimated percentage of average MAE is equal to 5%, which is lower but comparable to ours. Presence of more stations would be beneficial for the spatial interpolation. As such, it could help the detection and evaluation of suitable locations for future wind farm sites [61].

Table 4.19: Cross validation performance measures calculated through the leave-one-out cross validation for the mean installed power production of the year 2001. ME: mean error. MAE: mean absolute error. RMSE: root mean squared error. ρ: Pearson's correlation coefficient. ErrMin: minimum error between the prediction and the sample value. ErrMax: maximum error between the prediction and sample value. The validation measures are in MW.

ME	**MAE**	**RMSE**	ρ	**ErrMin**	**ErrMax**
$1.83\ 10^{-4}$	0.08	0.10	0.85	-0.22	0.22

Chapter 5

Conclusions

In this study a spatial and a temporal model for installed power production in Netherlands (2001–2006) is estimated using geostatistical and forecasting methods. The goal is to investigate the spatial and temporal variability of wind in this country and time period and validate the forecast.

The data set consists of average daily measurements of installed power production, for 46 stations distributed within Netherlands. The available period is for 6 years 2001–2006, ignoring the leap days. For both cases (spatial and temporal analysis) different temporal discretization of the data were tested, i.e daily, weekly, monthly, and annual average of installed power production. For the spatial analysis the annual average was chosen and for temporal analysis the monthly average was used. The Weibull distribution is fit on the data for spatial analysis. For the temporal analysis, depending on the data-set, normal, log–normal and Weibull distribution were chosen as the optimal fit.

In spatial analysis the annual average of installed wind power production is used. Due to Netherlands being a relatively flat country without rough terrain or significant altitude changes, it is not necessary to remove a topographic trend from the data. Also, because the Weibull distribution is close to the normal distribution for this data set, a transformation from Weibull to normal distribution was not implemented in this study. To investigate the spatial variability, the empirical variogram was calculated. The Spherical variogram model was found to be the best fit on the empirical variogram from all models tested, based on the sum of weighted squared errors. Ordinary Kriging was subsequently used to create an interpolation map that details the spatial estimation of the installed wind power production in each area of

the country.

Studying the maps of predictions for all six years (2002 to 2006 in the appendix), the highest predictions are located in the West area of the map, specifically in stations which are in the sea. The lowest estimators are in the south-east area of map. The highest values of the predicted power production are explained by the strong winds which come from the North Sea. This pattern is observed every year, i.e the highest values are predicted in the west area of the map, and the lowest in the east. Also it is observed that the mean value of the predicted annual installed power production does not differ significantly between the years, i.e. the range varies from 0.43 to 0.53. Hence it can be assumed that the velocity and the frequency of winds are constant over the years, so each station has a constant production every year, which is beneficial for a wind park.

In comparison between the predicted and the original installed power production, the model slightly underestimates wind power production in the North Sea, or on the North Sea Coast. The mean error between the predictions and the sample for the North Sea stations is -0.055 MW, representing a slight underestimation by the model. This underestimation could be the result of kriging smoothing. The results are comparable with similar works, which they use different kriging methods [29, 31].

In temporal analysis the monthly average of installed power production for each station is calculated. As such, the time series pertaining to each station have 72 steps. The Weibull distribution fits well the time series of 31 stations. log–normal distribution was fit on 11 time series and the normal distribution was fit in the last 4 time series. The time series which are fit with log–normal distribution, are transformed by calculating the logarithm of the values. For the time series more closely following Weibull distribution there was no need for any transformation because the distribution was already close to the normal distribution.

To estimate the annual seasonality, SARIMA models were used. For seven time series, a harmonic model was subtracted from the data first and SARIMA was used on the residuals instead of the original data. An autoregressive–moving–average model is fitted to the data (or the residuals) with annual seasonality. The best SARIMA model was chosen based on AIC, AICc, and BIC. After estimating the parameters of the model, a prediction for the following year (2007) was implemented. To validate the accuracy of the model, cross–validation was used and good performance was observed in all data sets. Further validation was performed by ensuring that there

is no autocorrelation remaining after applying the optimal SARIMA model.

Studying the two stations (Station 1 onshore, station 31 offshore) some differences are readily apparent. In Onshore Station it is necessary to remove a harmonic model first and the fit the SARIMA model, in order to extract the seasonality. In contrast, in Offshore Station the SARIMA model is fitted directly to the data which proves to be enough to remove the seasonality. Furthermore, in Onshore Station the RMSE is 27% of the mean value, and in Offshore Station is 25% of the mean value. In conclusion the differences may be due to the offshore station being unaffected by the terrain or the topography of the area. The accuracy of the model is comparable to similar work for short-term predictions using either wind speed forecasts and the power curve or wind power forecasts directly [4, 36].

In a future study, different time scales could be used to compare the results. In a similar study, using different time scales to predict the power production, the RMSE is increasing as the time scale is increasing (time lags are becoming larger) [18]. Also, it is necessary to compare current results with other methods, such as using the power curve. Furthermore, the deterministc seasonal effects(i.e. predictable periodic changes in the levels of the time series) could be incorporated by augmenting the ARIMA model with seasonal dummy variables [44]. Finally, an interesting evolution of the present study is to connect the results with the industrial standards for prediction and risk assessment, by estimating the percentages P50, and P90. Indicatively, these percentages were estimated for the two stations. For Onshore Station and Offshore Station, 1000 simulations are implemented (using the command arima.sim [52] in R from stats package), and the respective percentiles (P50 and P90) are estimated. For Onshore Station, P50 is equal to 13.5 GWh (Comparable to the actual 13.4 GWh produced), and P90 is equal to 12.5 GWh for the years 2004–2006. For Offshore Station , P50 is equal to 18.0 GWh (equal to the actual 18.0 GWh), and P90 is equal to 16.8 GWh for the years 2004–2006.

Appendices

Appendix A

Figures for Spatial Analysis

In Appendix A the figures from 2002 until 2006 for the spatial analysis, i.e the annual power production are presented. The figures include:

1. Experimental variogram (dashed line), and theoretical Spherical model (continuous line), using the Equation 2.20. The horizontal axis is the lag distance r in km, and the vertical distance represents the variogram values for installed power production, for n each lag. The estimated parameters for the theoretical model are presented separately on every figure's caption
2. Kriging-based on leave-one-out cross-validation predictions versus sample values for power production
3. Map of estimated annual mean installed power production for 2001, based on Spherical variogram model. The horizontal axis represents the Easting measured in kilometers, and the vertical axis represents the Northing measured in kilometers.
4. Map of estimated variance of mean installed power production for 2001, based on Spherical variogram model. The horizontal axis represents the Easting measured in kilometers, and the vertical axis represents the Northing measured in kilometers.

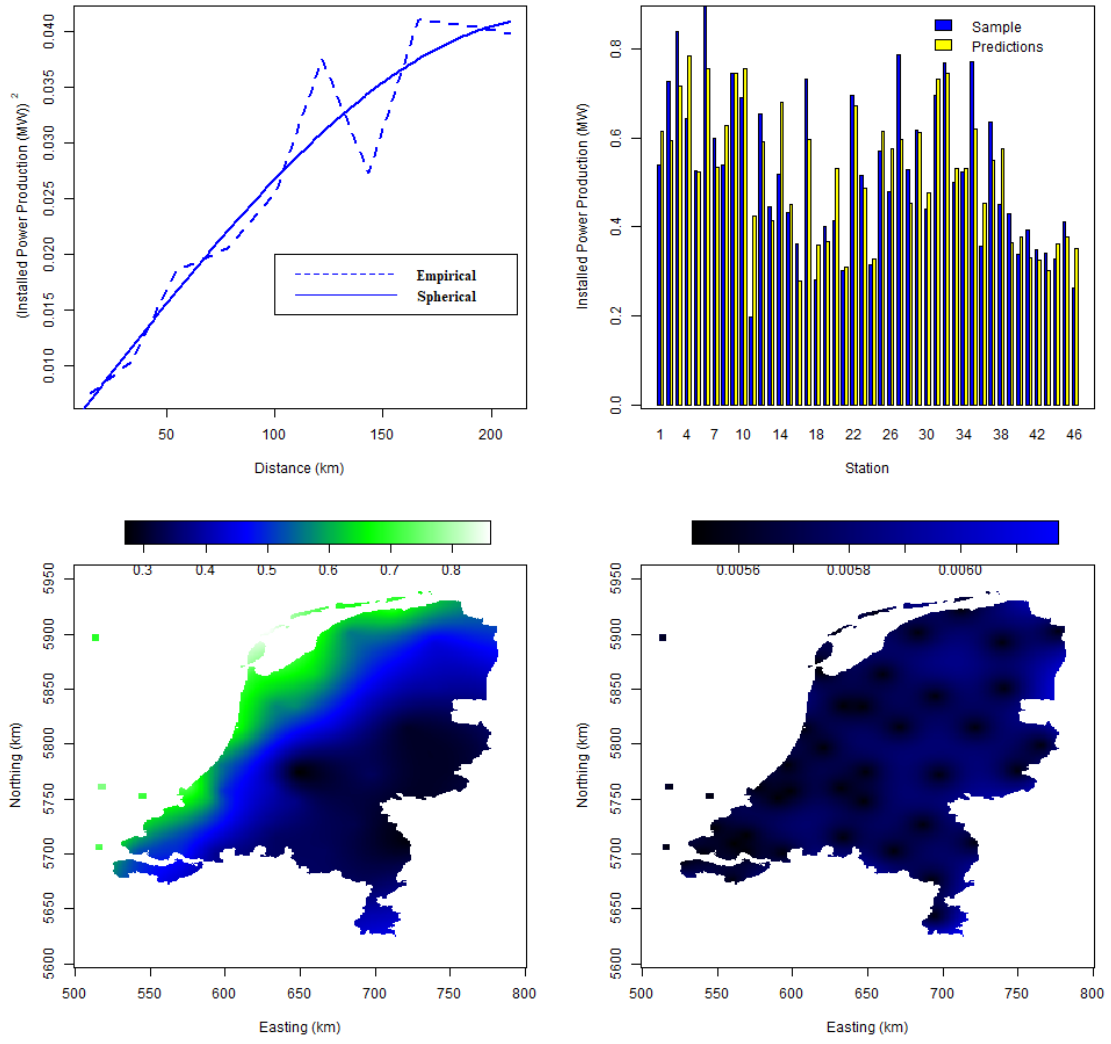


Figure A.1: Year 2002 annual power production. The Spherical variogram parameters are: nugget=0.0032 (MW^2), variance $\sigma^2 = 0.0381$ (MW^2), and range = 226.36311 km.

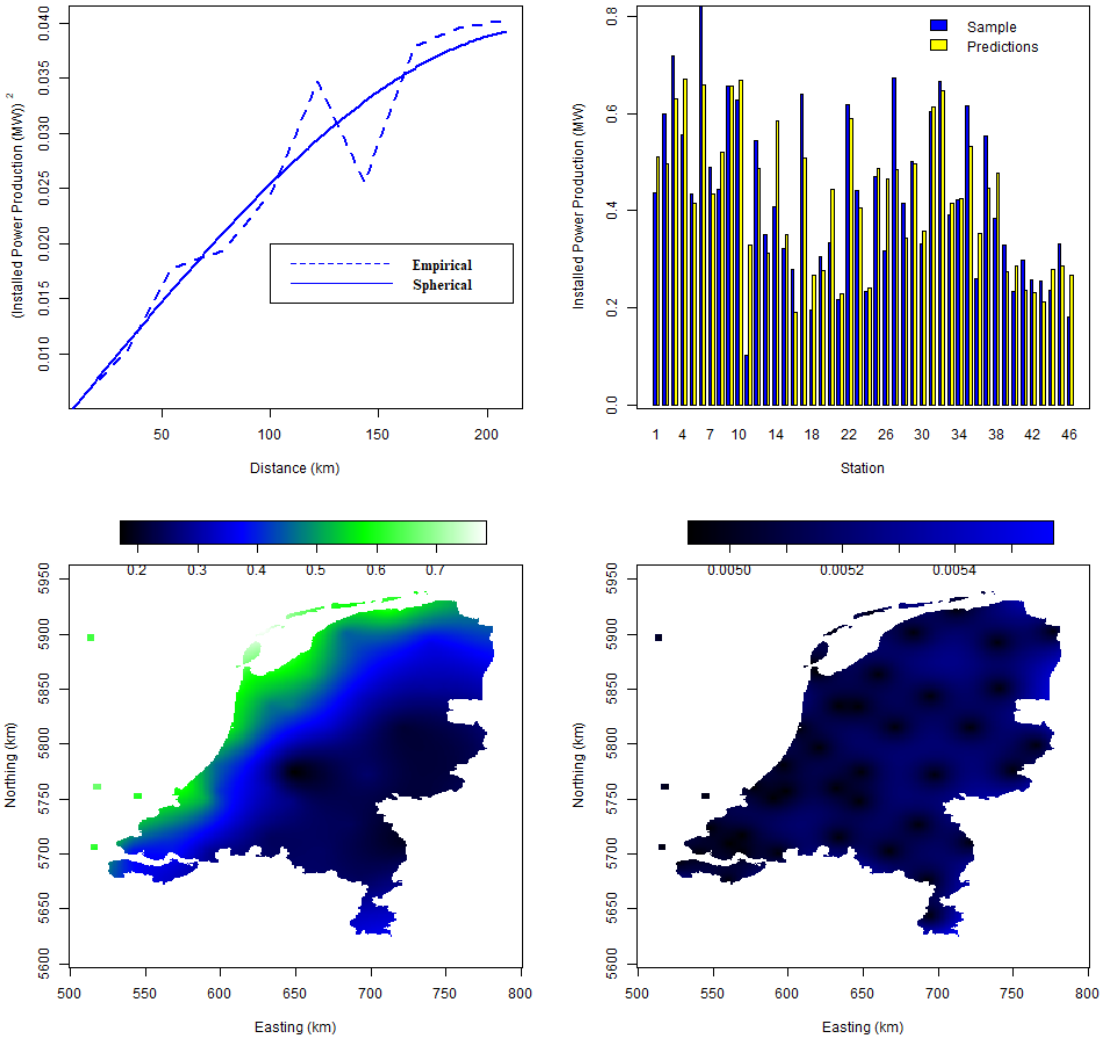


Figure A.2: Year 2003 annual power production. The Spherical variogram parameters are: nugget=0.0028 (MW^2), variance σ^2 = 0.0369 (MW^2), and range = 229.2011 km.

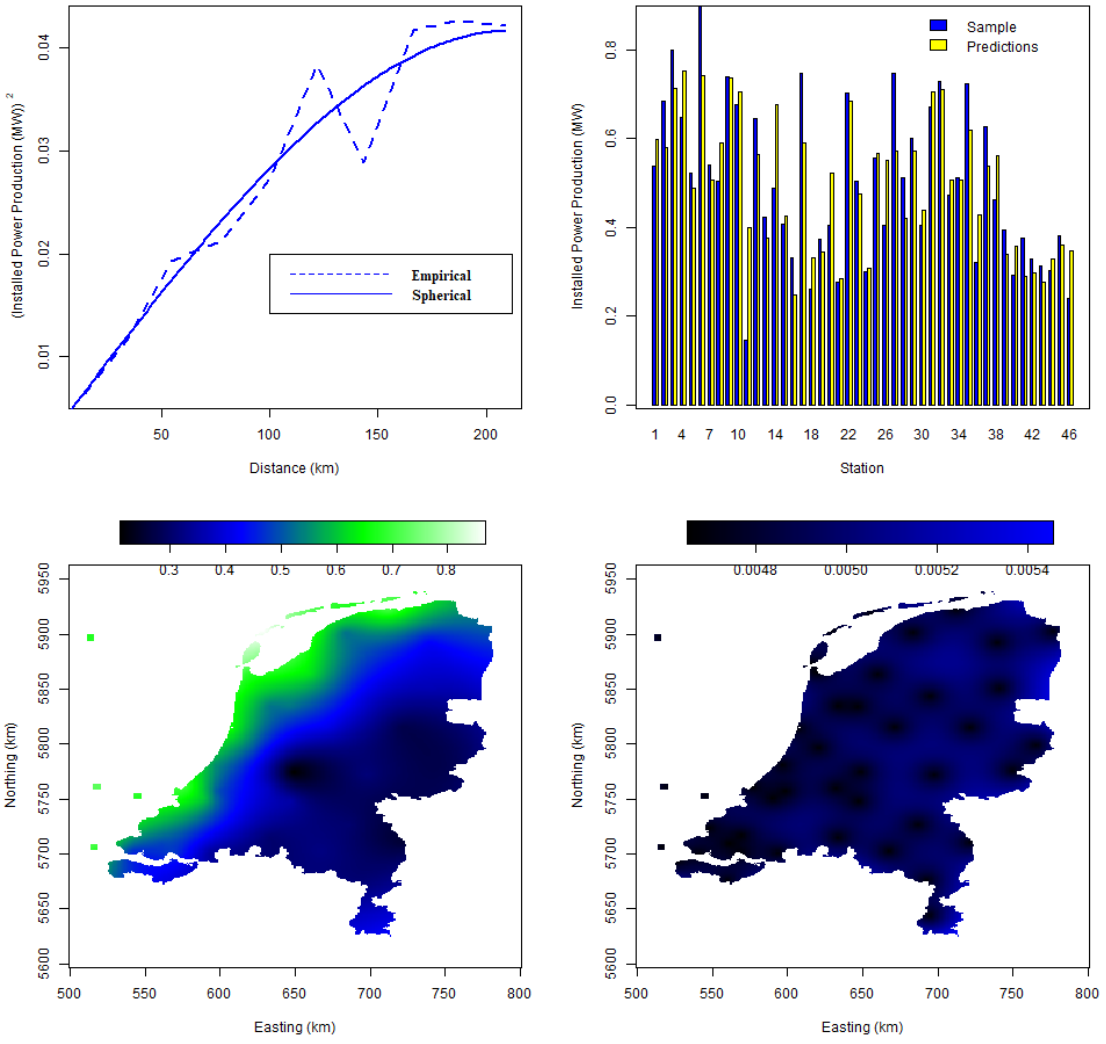


Figure A.3: Year 2004 annual power production. The Spherical variogram parameters are: nugget=0.0025 (MW^2), variance σ^2 = 0.0391 (MW^2), and range = 210.5247 km.

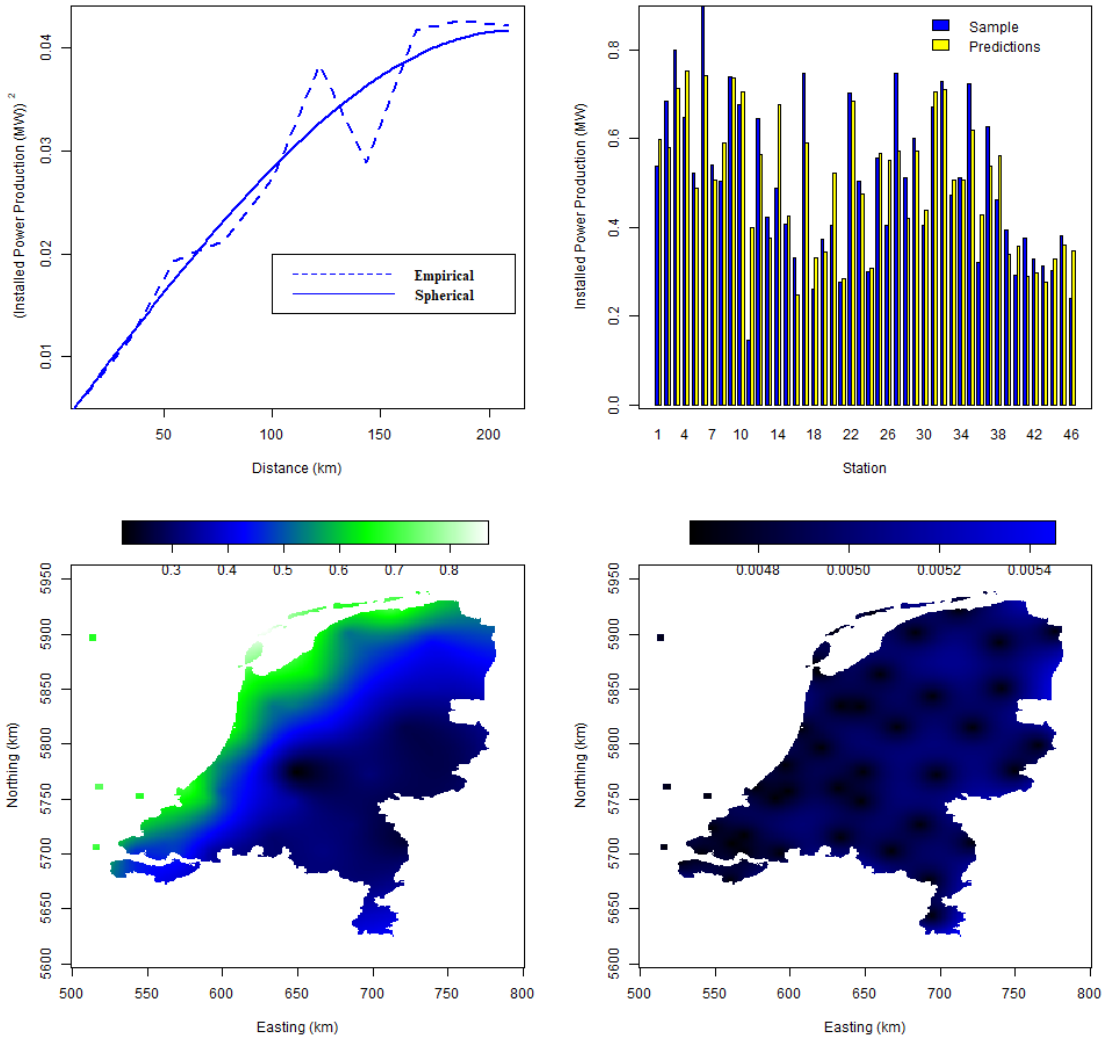


Figure A.4: Year 2005 annual power production. The Spherical variogram parameters are: nugget=0.0040 (MW^2), variance σ^2 = 0.0493 (MW^2), and range = 247.6334 km.

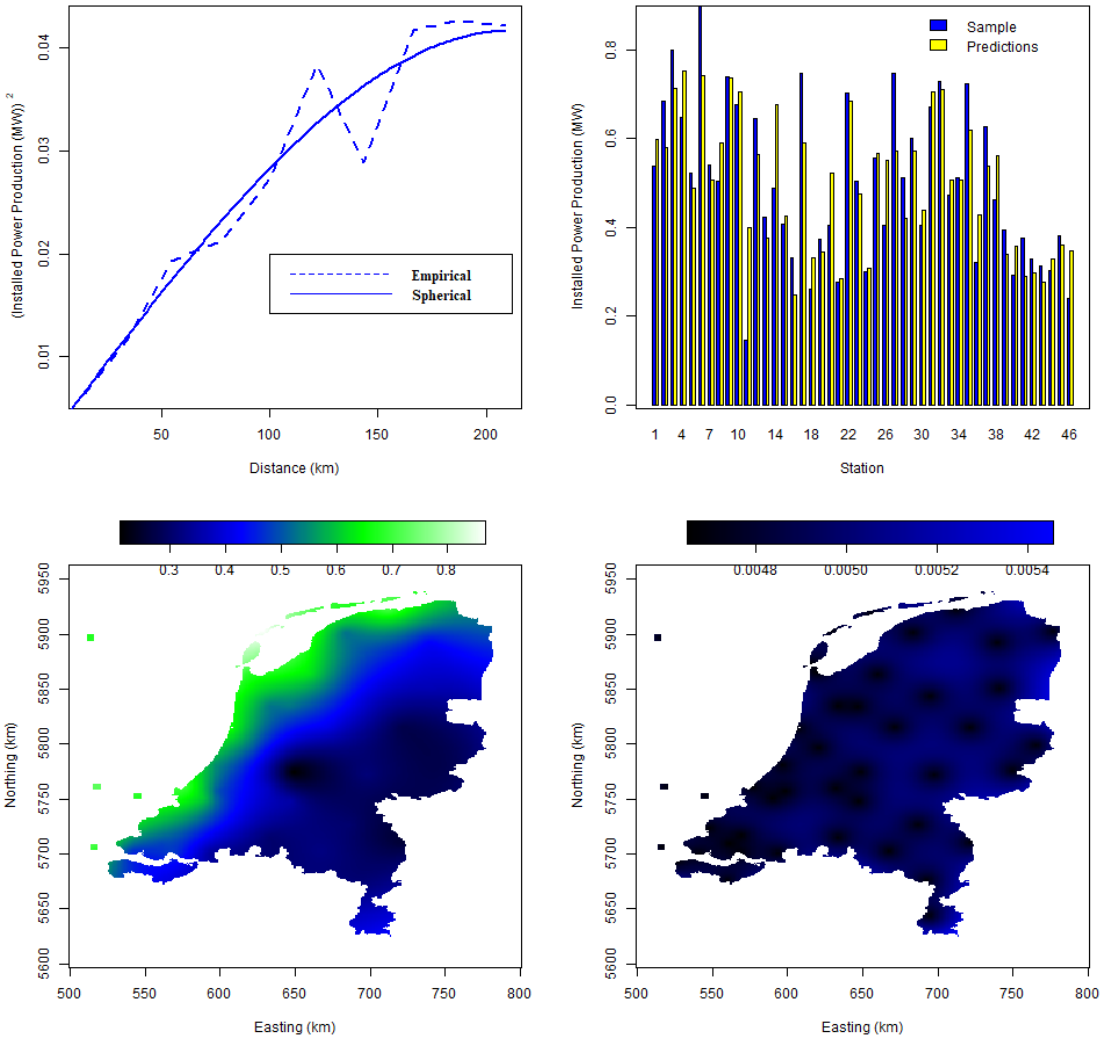


Figure A.5: Year 2006 annual power production. The Spherical variogram parameters are: nugget=0.0034 (MW^2), variance σ^2 = 0.0047 (MW^2), and range = 215.5592 km.

Appendix B

Figures for Temporal Analysis

In Appendix B the figures for the temporal analysis for the monthly average installed power production are presented. The figures include:

1. Time series of average monthly power production. The horizontal axis represents time (years: 2001–2006) and the vertical axis shows the installed power in MW.
2. Periodogram of monthly average wind power at Station 1. The horizontal axis represents the frequency and the vertical axis the value of the periodogram.
3. Time series of residuals of installed power production.The horizontal axis represents time (years: 2001–2006) and the vertical axis shows the residuals in MW.
4. The autocorrelation function (ACF) for the residuals . The horizontal axis represents the time lag, while the vertical axis measures the autocorrelations.
5. The normal distribution plot.
6. The p-values for the Ljung-Box statistic for the autocorrelation test.

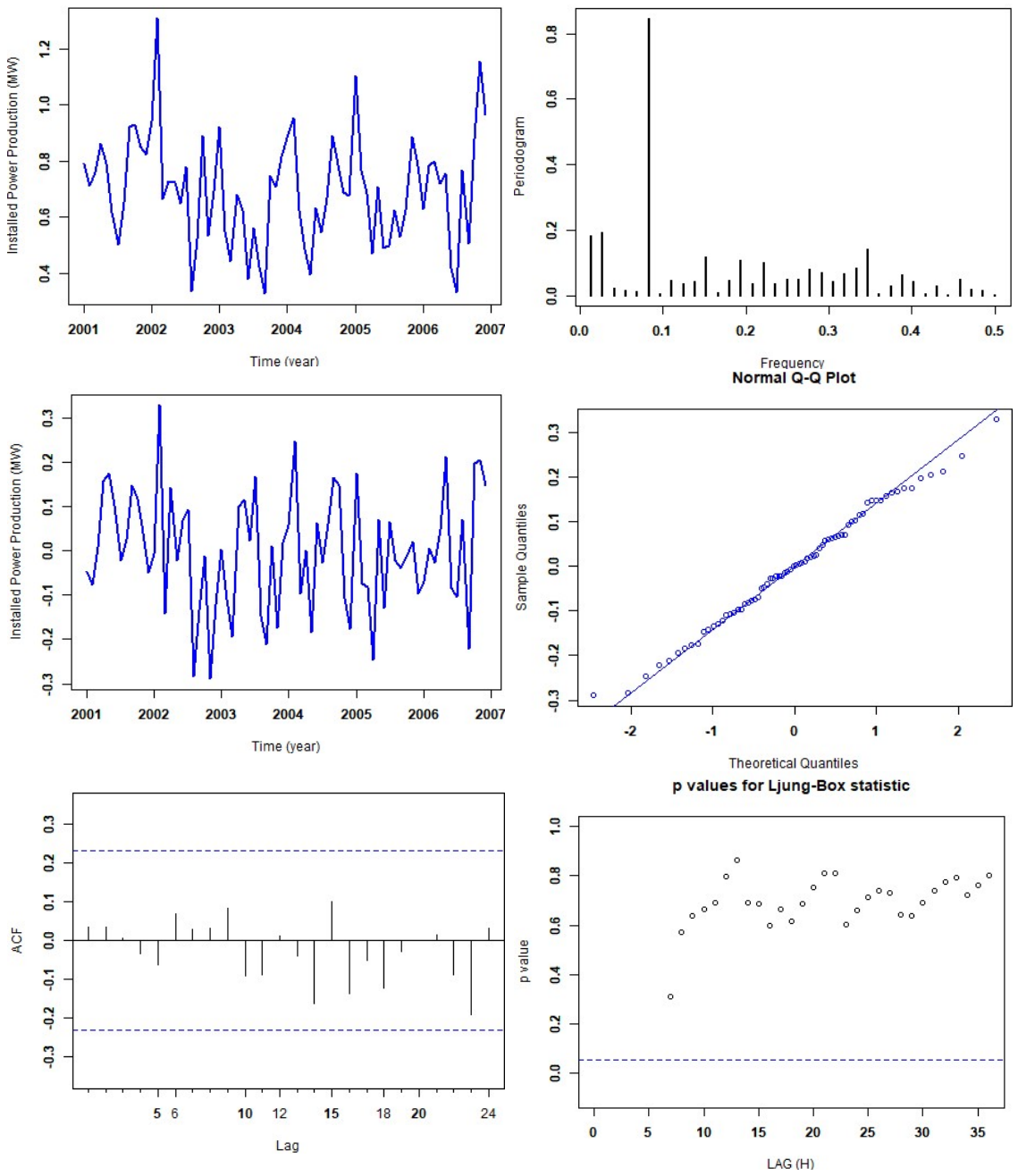


Figure B.1: Station 2: The fitted SARIMA model for installed power production is a SARIMA (1,0,3)(1,0,1)(12).

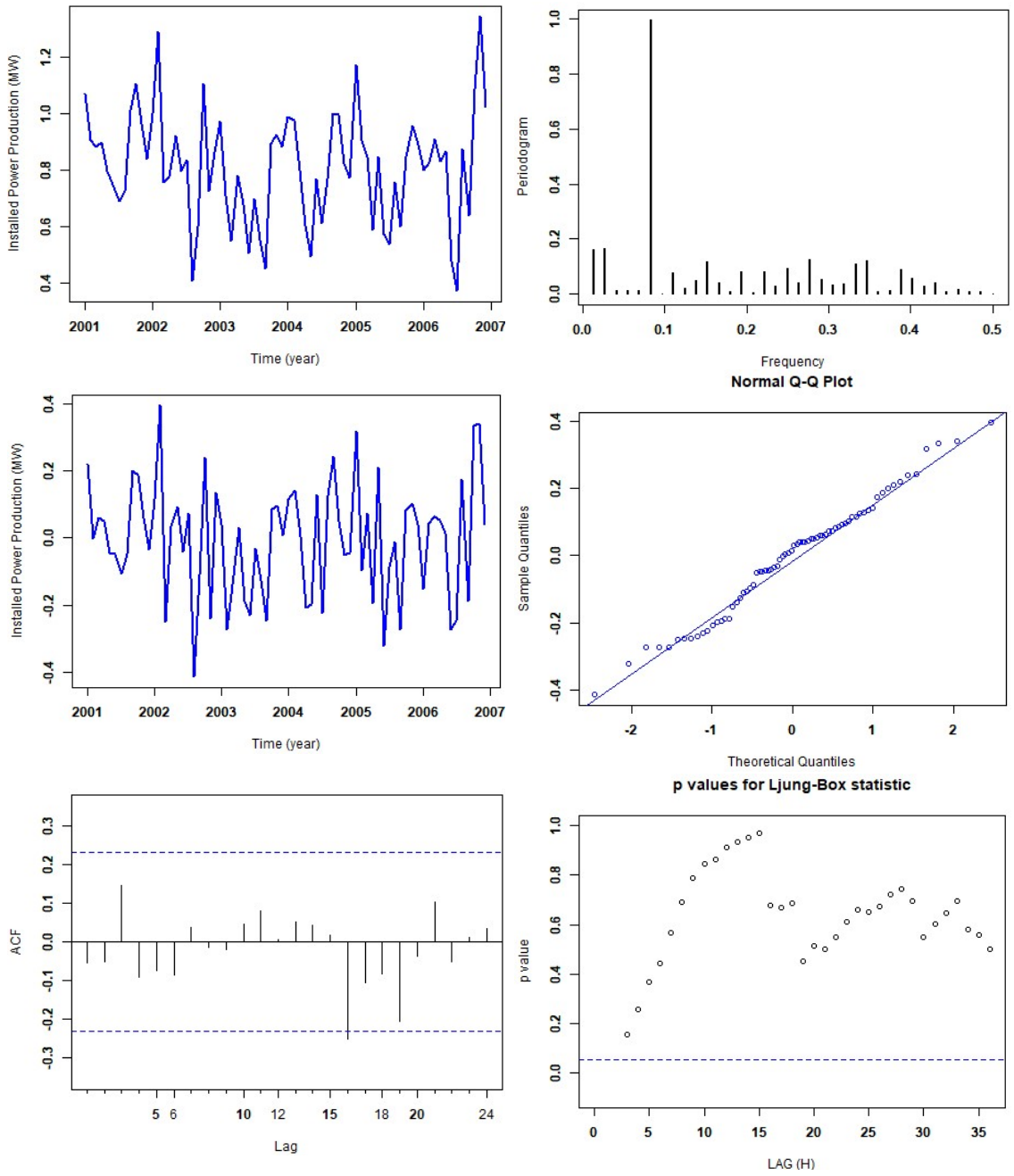


Figure B.2: Station 3 : The fitted SARIMA model for installed power production is a SARIMA (0,0,1)(1,0,0)(12).

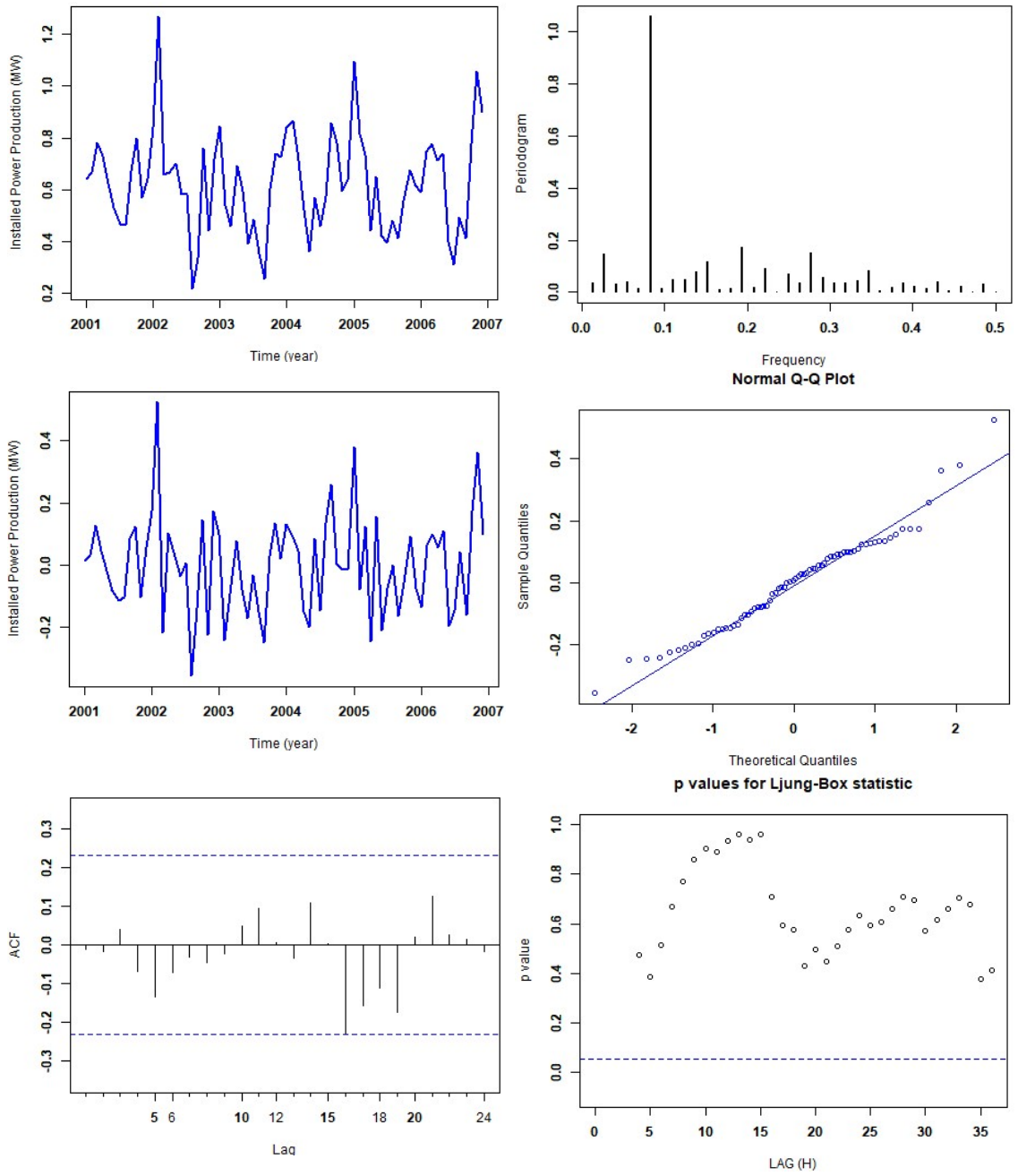


Figure B.3: Station 4: The fitted SARIMA model for installed power production is a SARIMA (0,0,1)(1,0,1)(12).

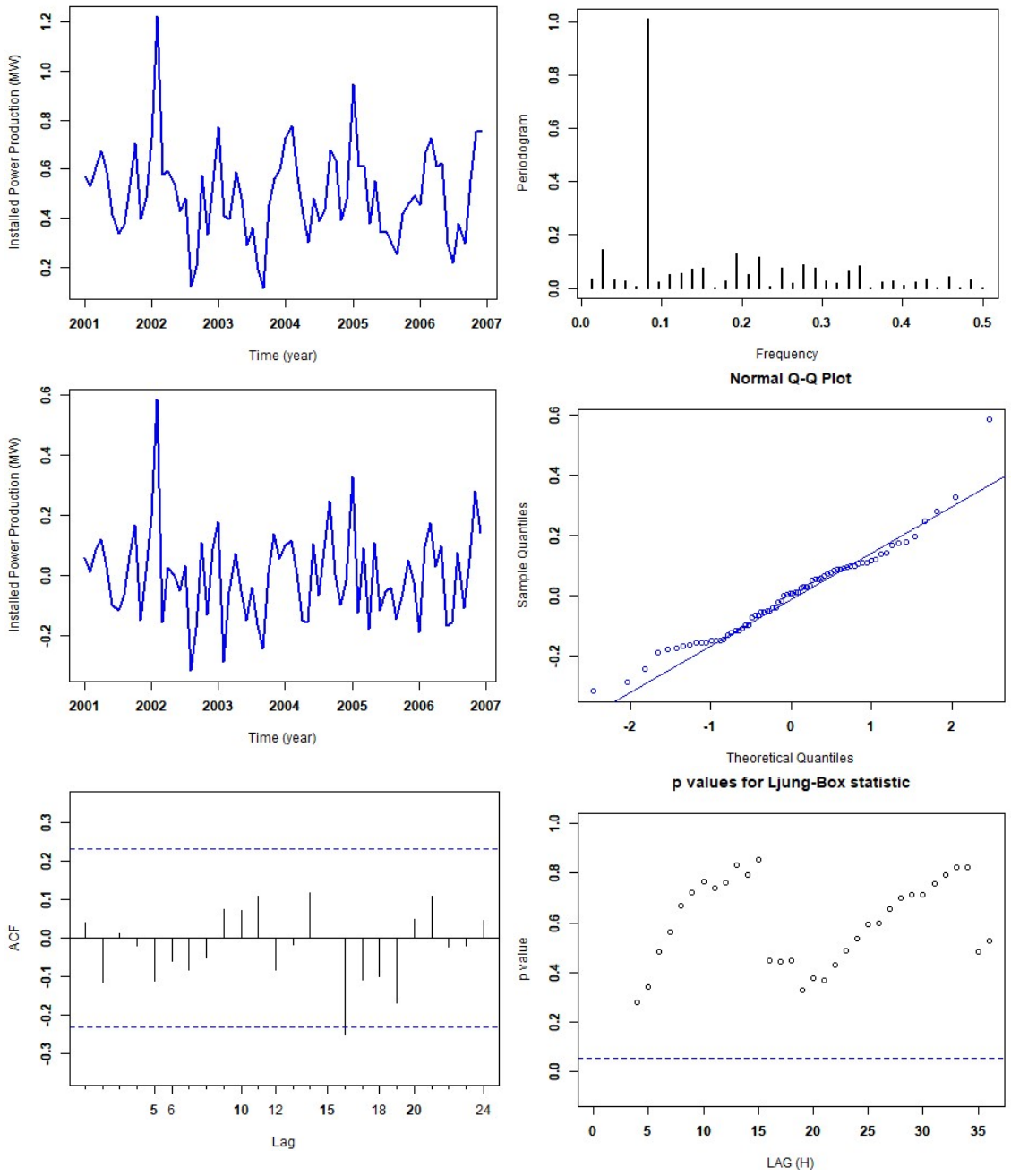


Figure B.4: Station 5 : The fitted SARIMA model for installed power production is a SARIMA (1,0,0)(1,0,1)(12).

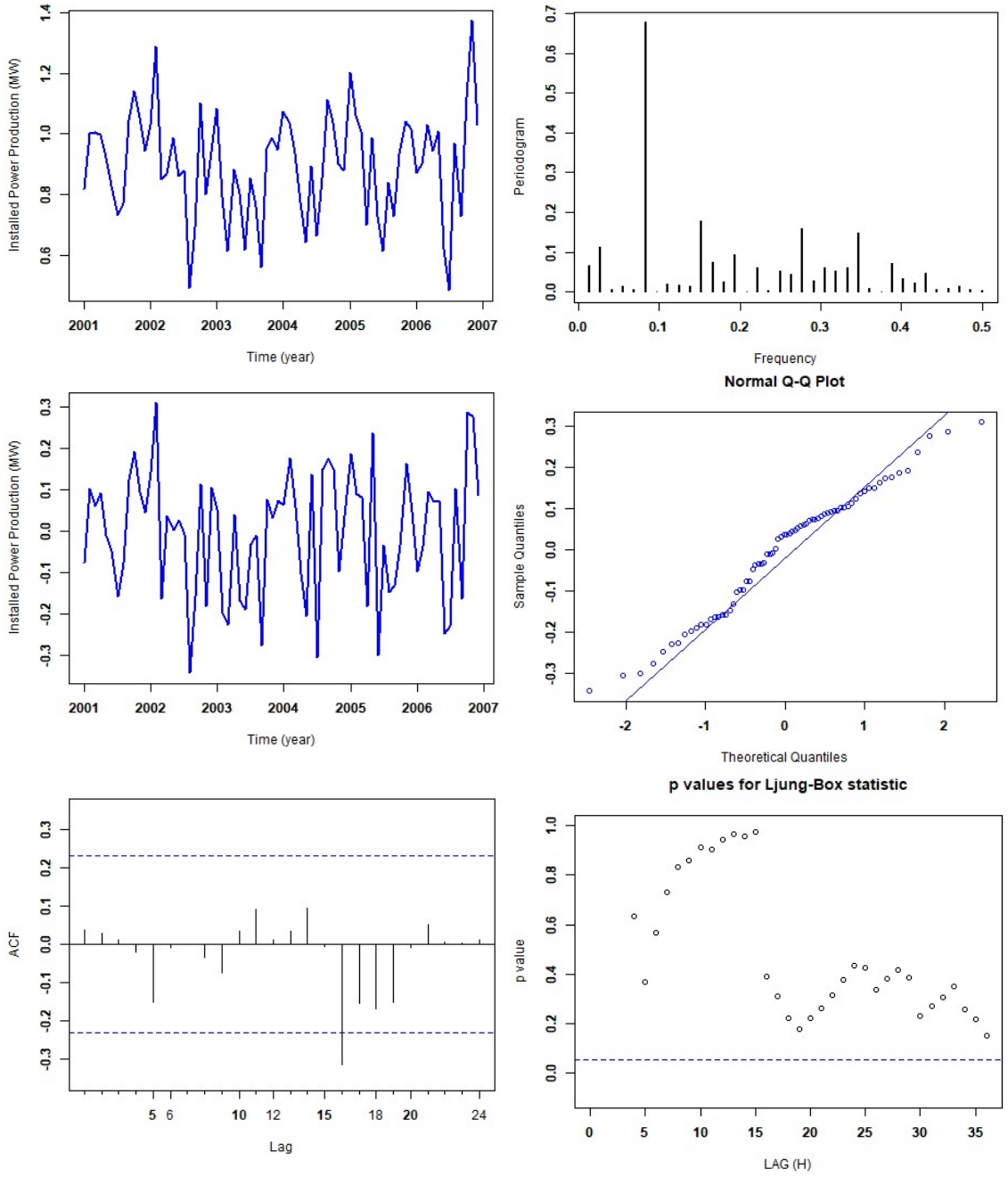


Figure B.5: Station 6 : The fitted SARIMA model for installed power production is a SARIMA (1,0,1)(1,0,0)(12).

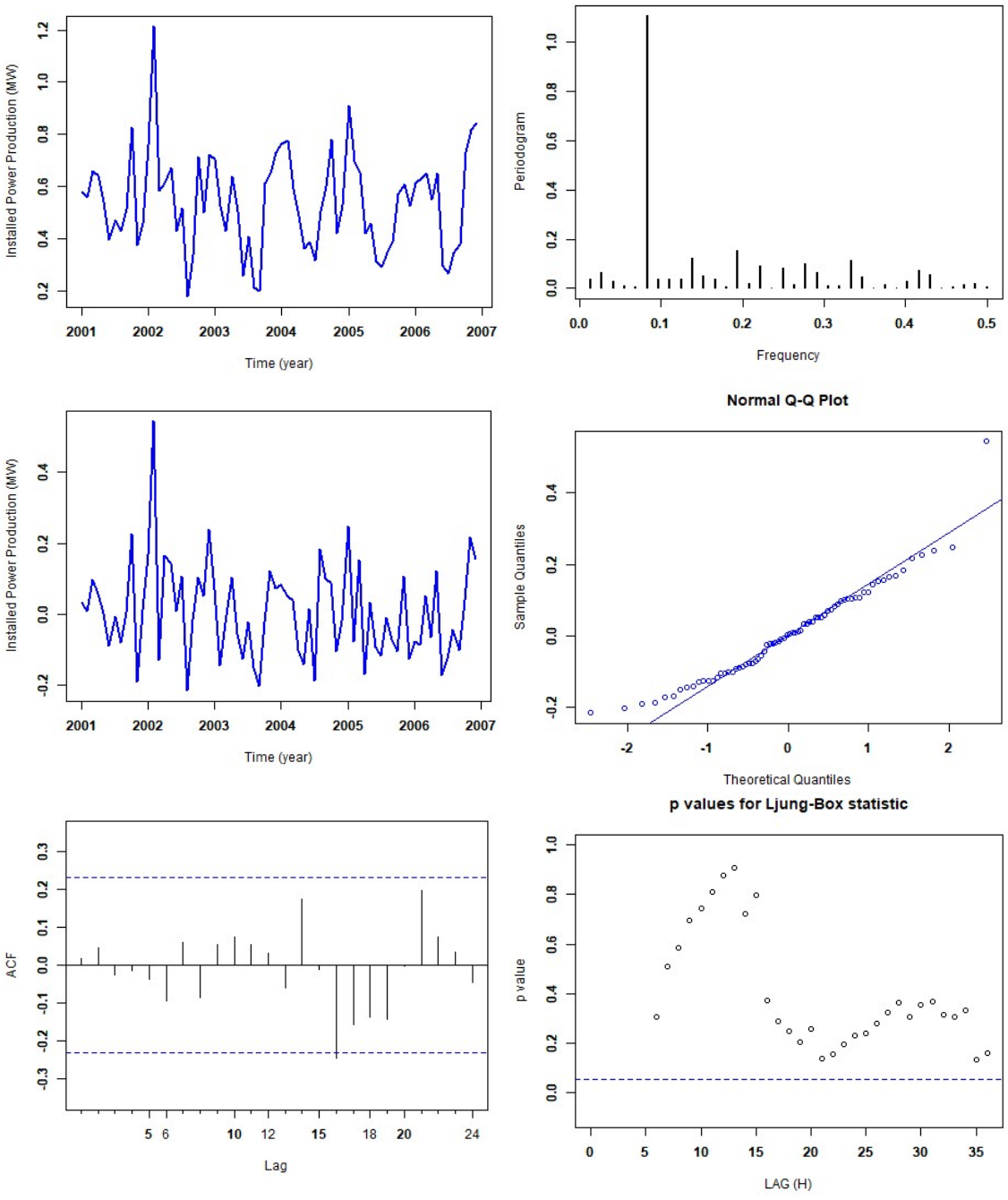


Figure B.6: Station 7: The fitted SARIMA model for installed power production is a SARIMA (1,0,2)(1,0,1)(12).

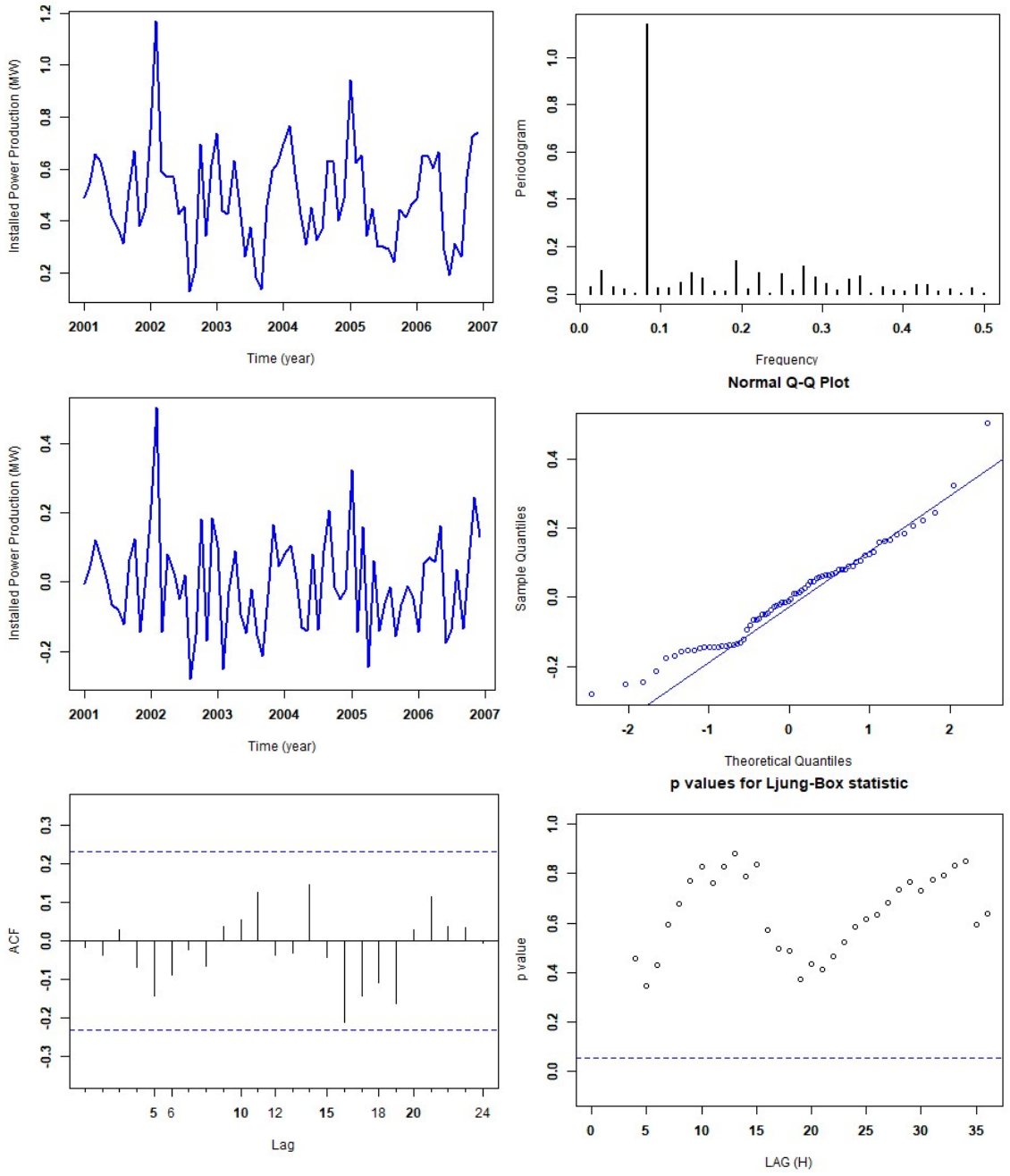


Figure B.7: Station 8: The fitted SARIMA model for installed power production is a SARIMA (0,0,1)(1,0,1)(12).

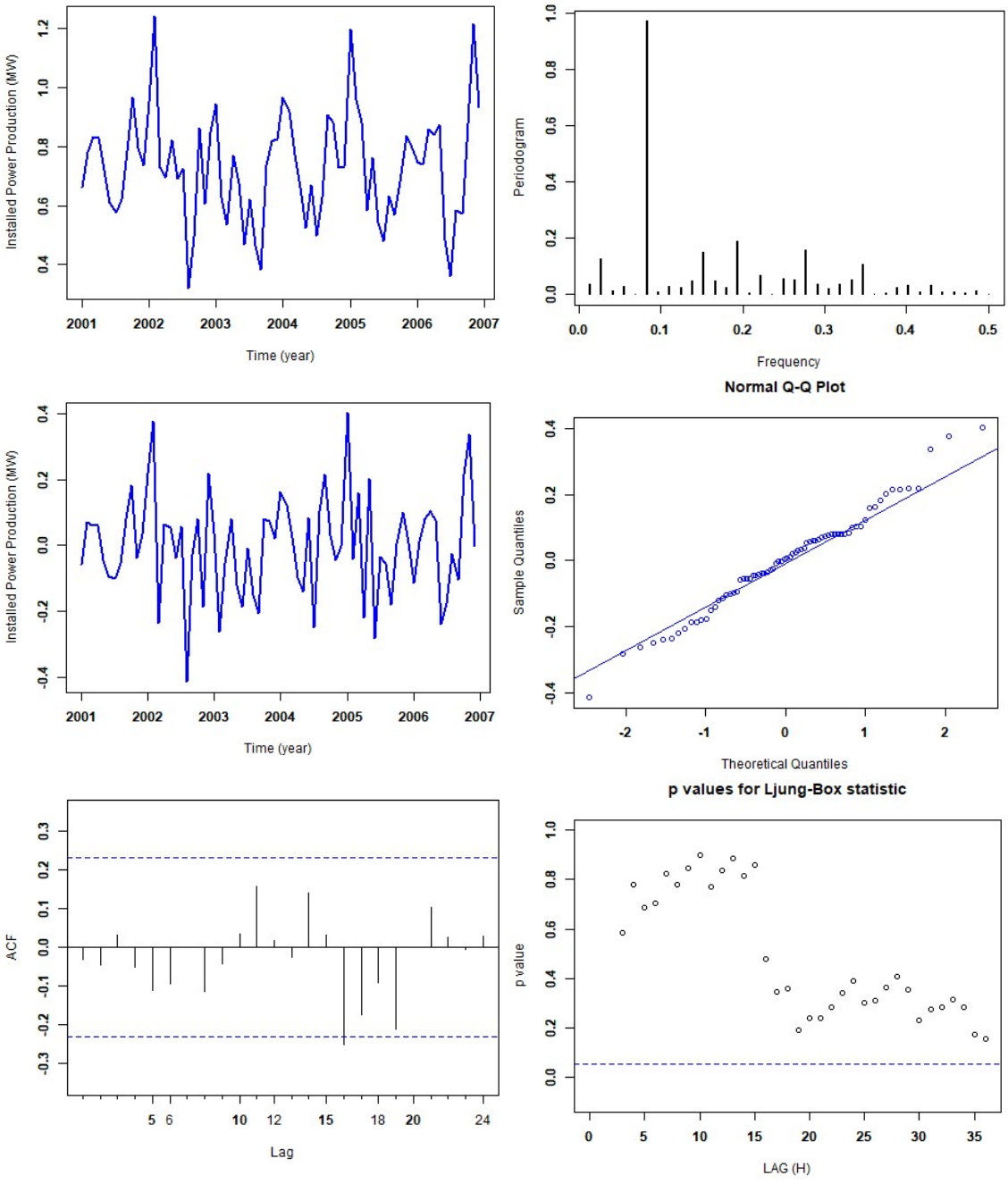


Figure B.8: Station 9: The fitted SARIMA model for installed power production is a SARIMA (0,0,1)(1,0,0)(12).

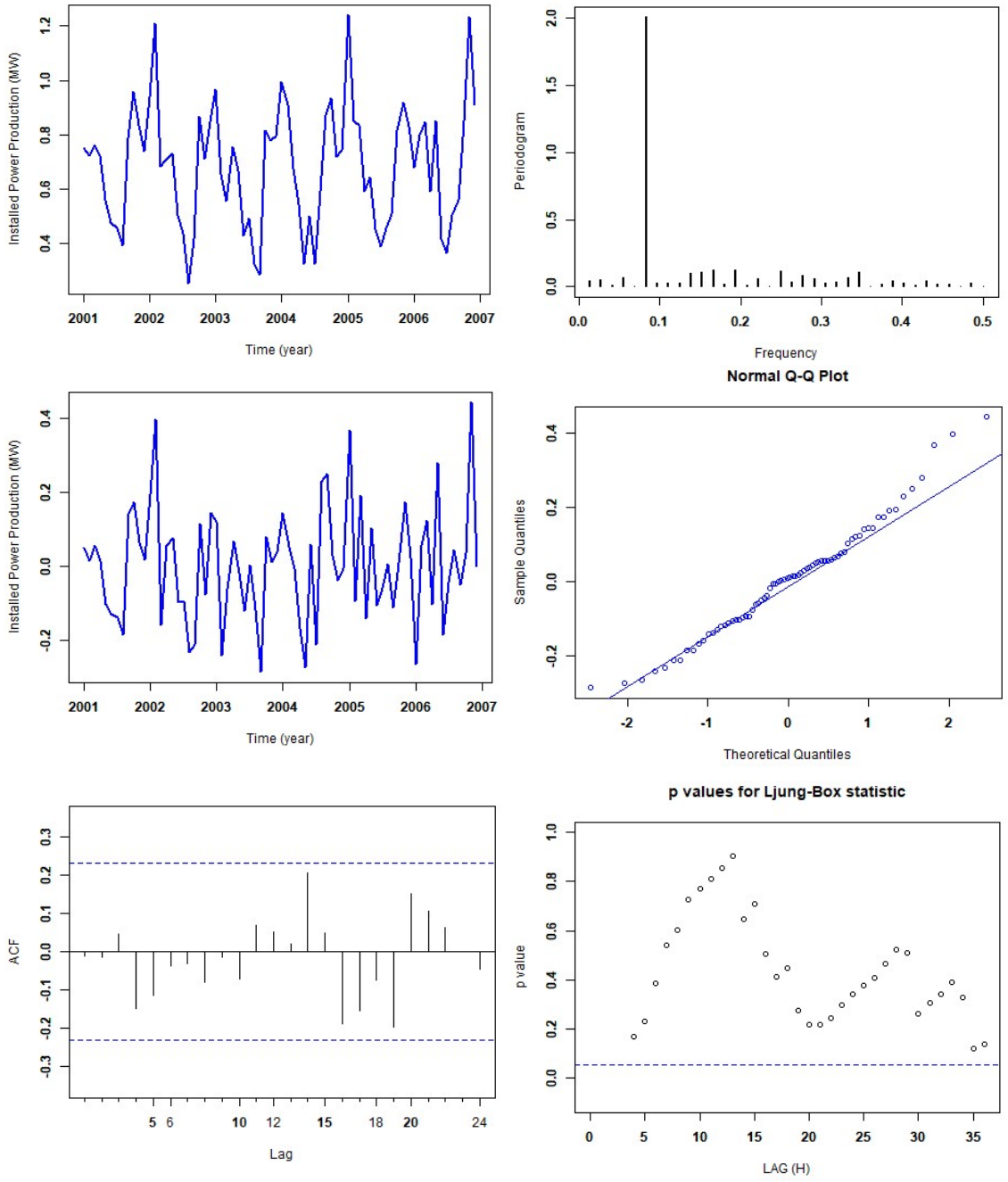


Figure B.9: Station 10: The fitted SARIMA model for installed power production is a SARIMA (0,0,1)(1,0,1)(12).

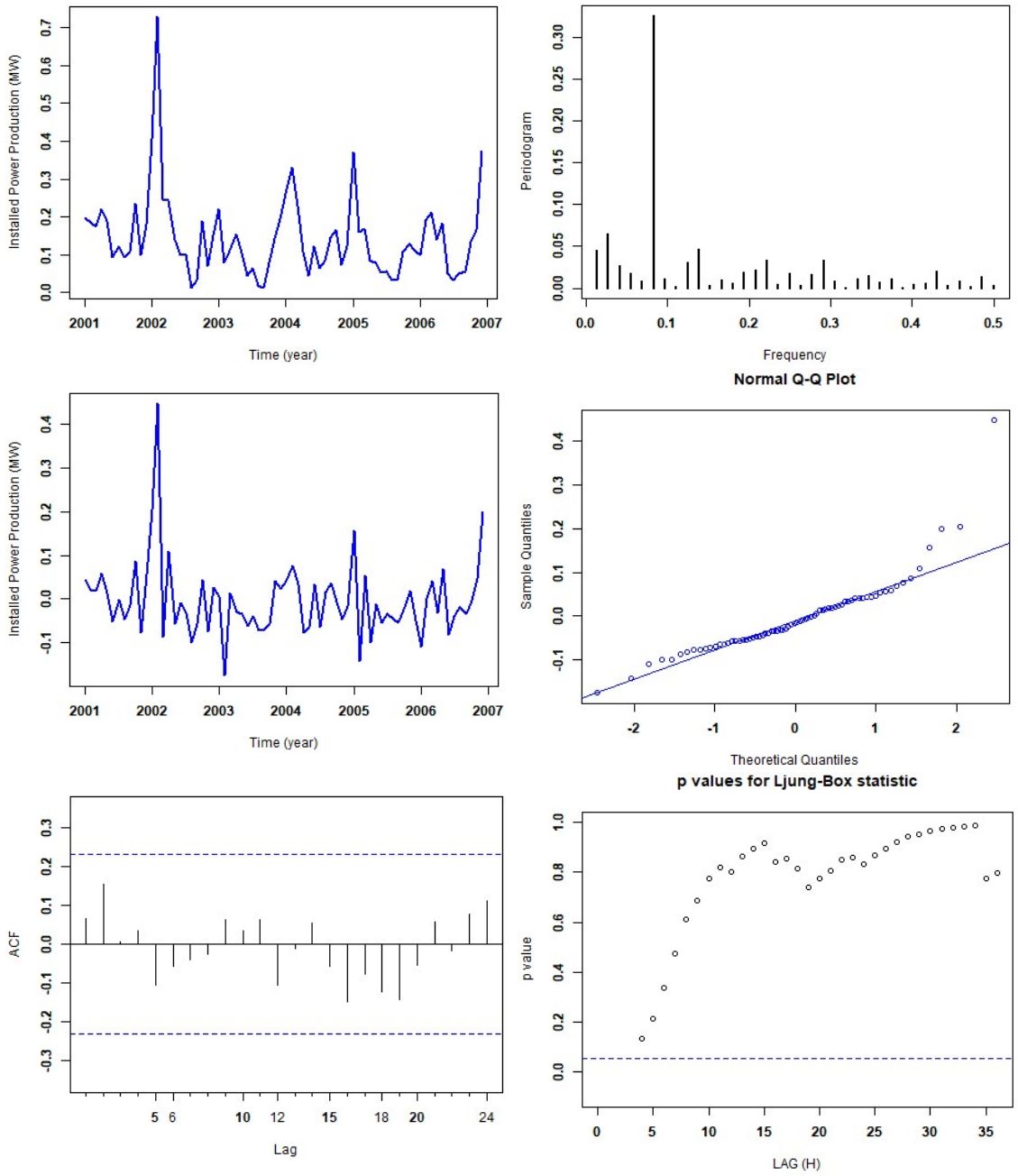


Figure B.10: Station 11: The fitted SARIMA model for installed power production is a SARIMA (0,0,1)(1,0,1)(12).

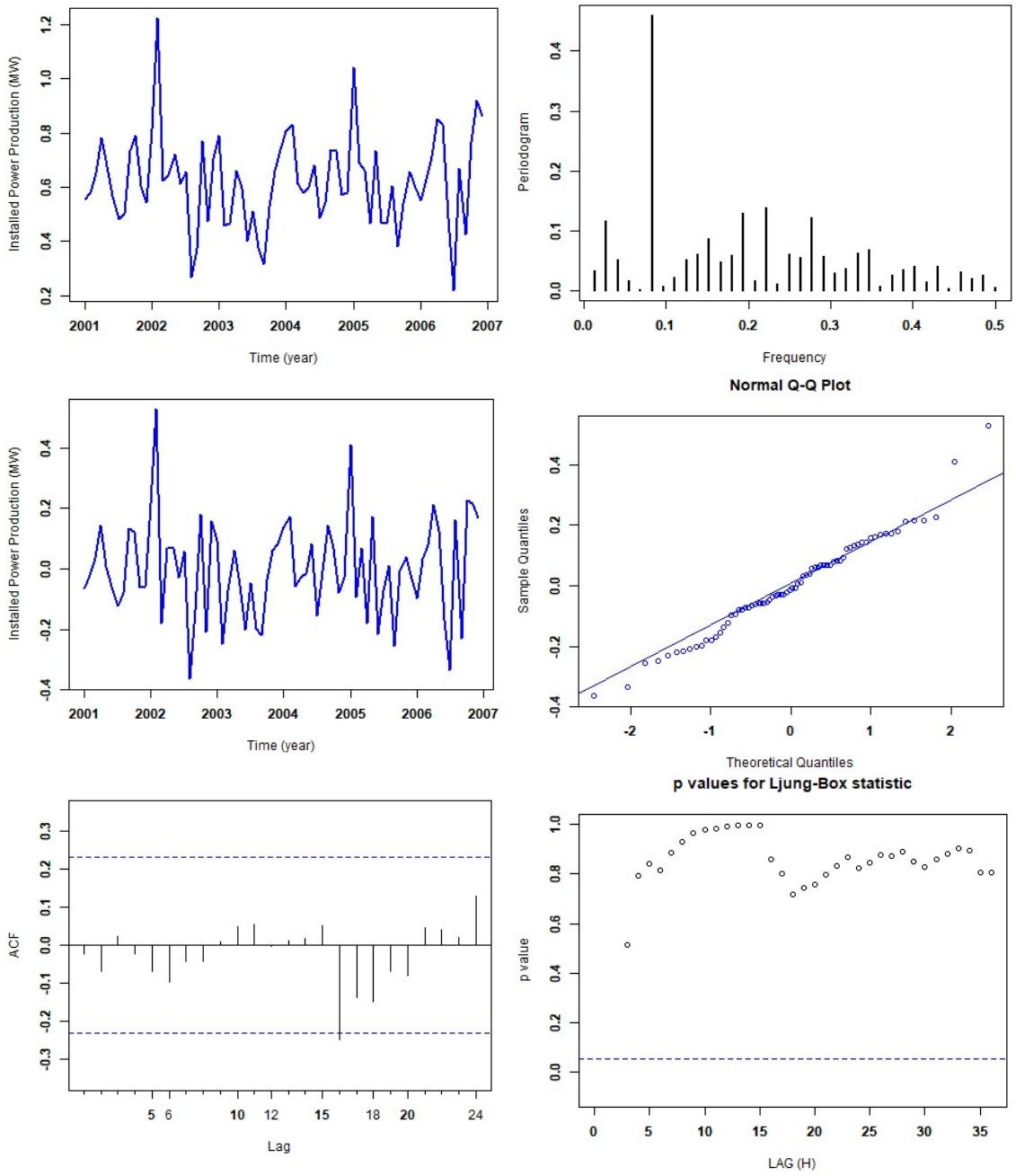


Figure B.11: Station 12: The fitted SARIMA model for installed power production is a SARIMA (0,0,1)(1,0,0)(12).

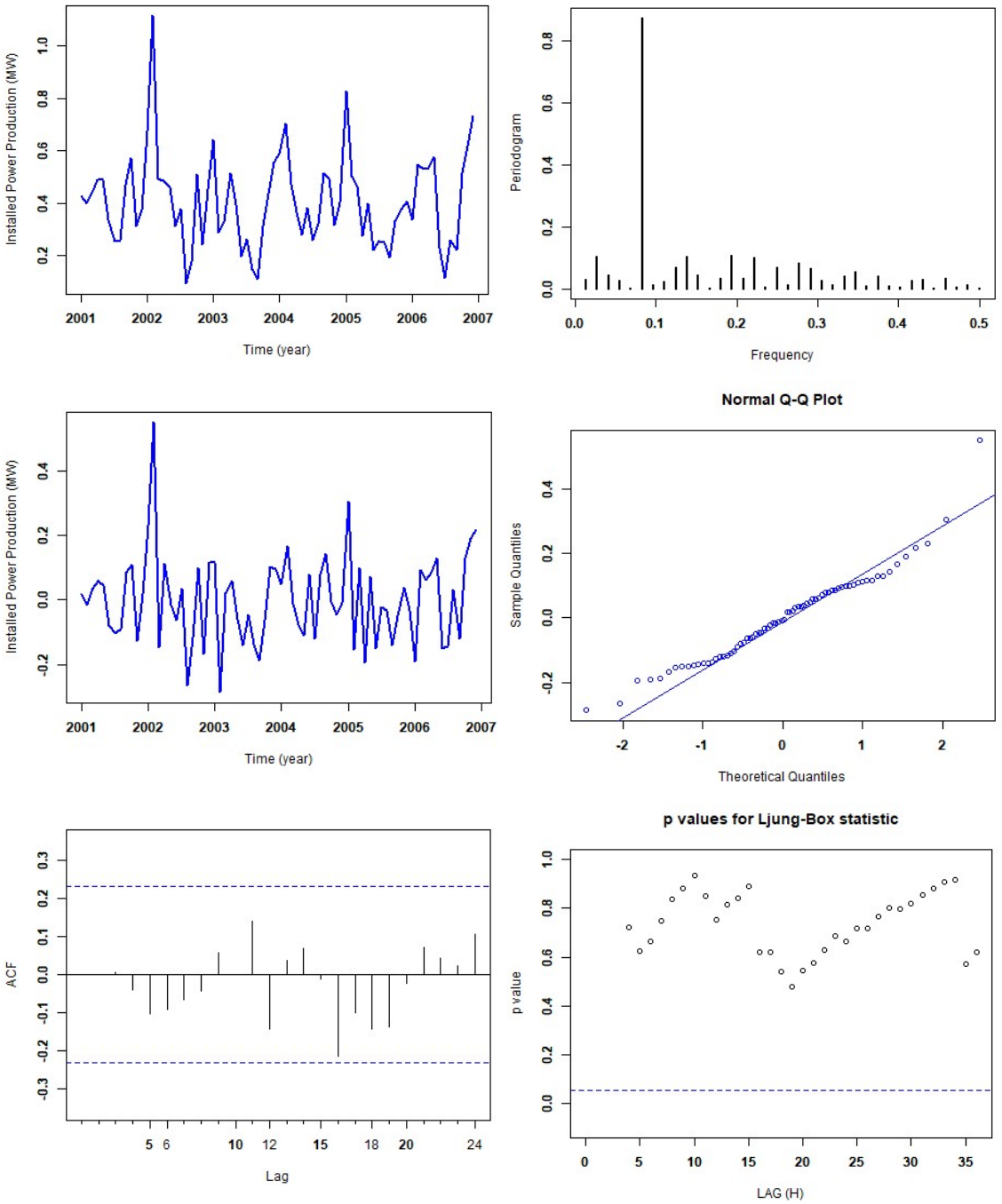


Figure B.12: Station 13: The fitted SARIMA model for installed power production is a SARIMA (0,0,1)(1,0,1)(12).

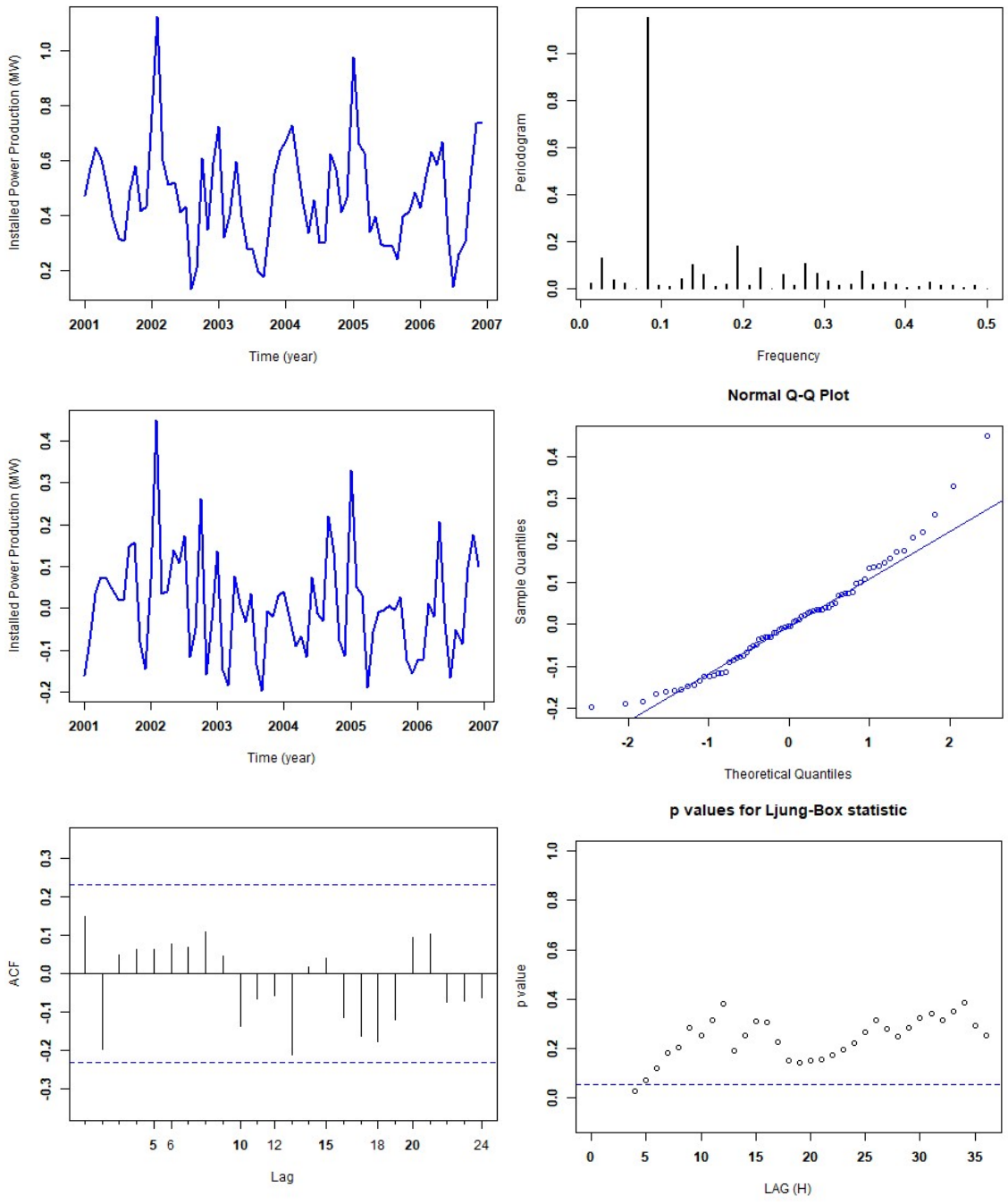


Figure B.13: Station 14: The fitted SARIMA model for installed power production is a SARIMA (1,0,1)(1,0,0)(12).

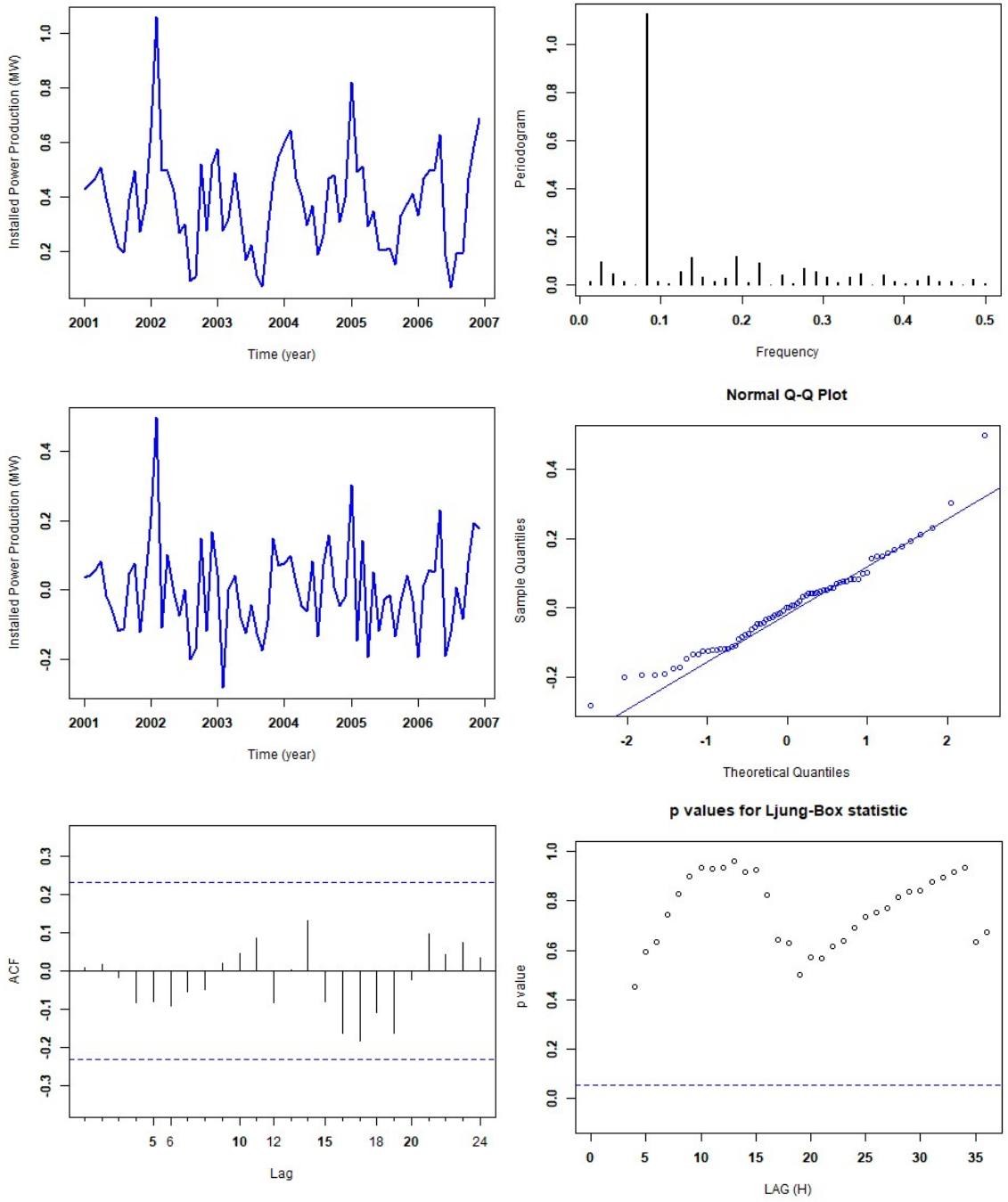


Figure B.14: Station 15: The fitted SARIMA model for installed power production is a SARIMA (0,0,1)(1,0,1)(12).

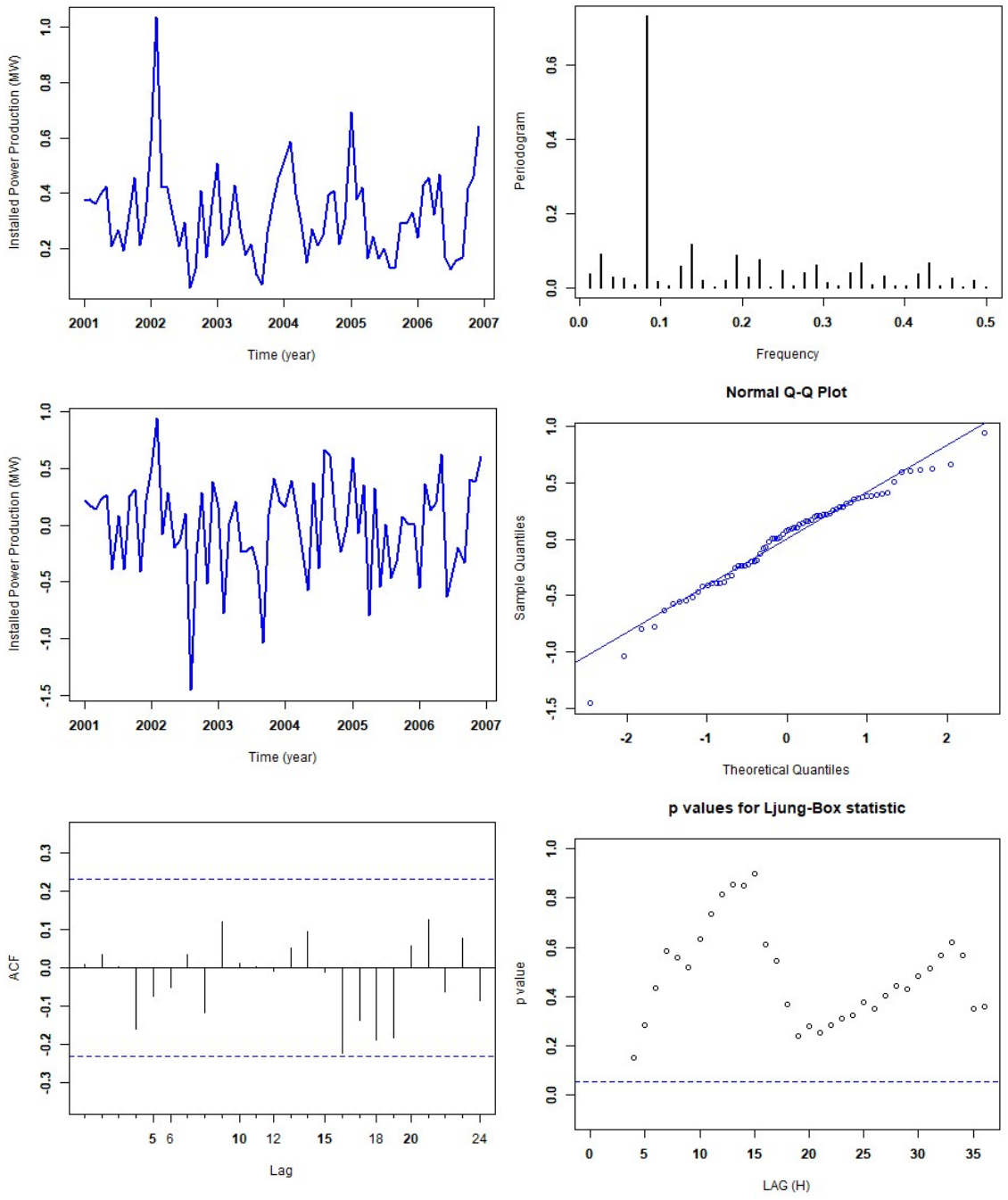


Figure B.15: Station 16: The fitted SARIMA model for installed power production is a SARIMA (0,0,1)(1,0,1)(12).

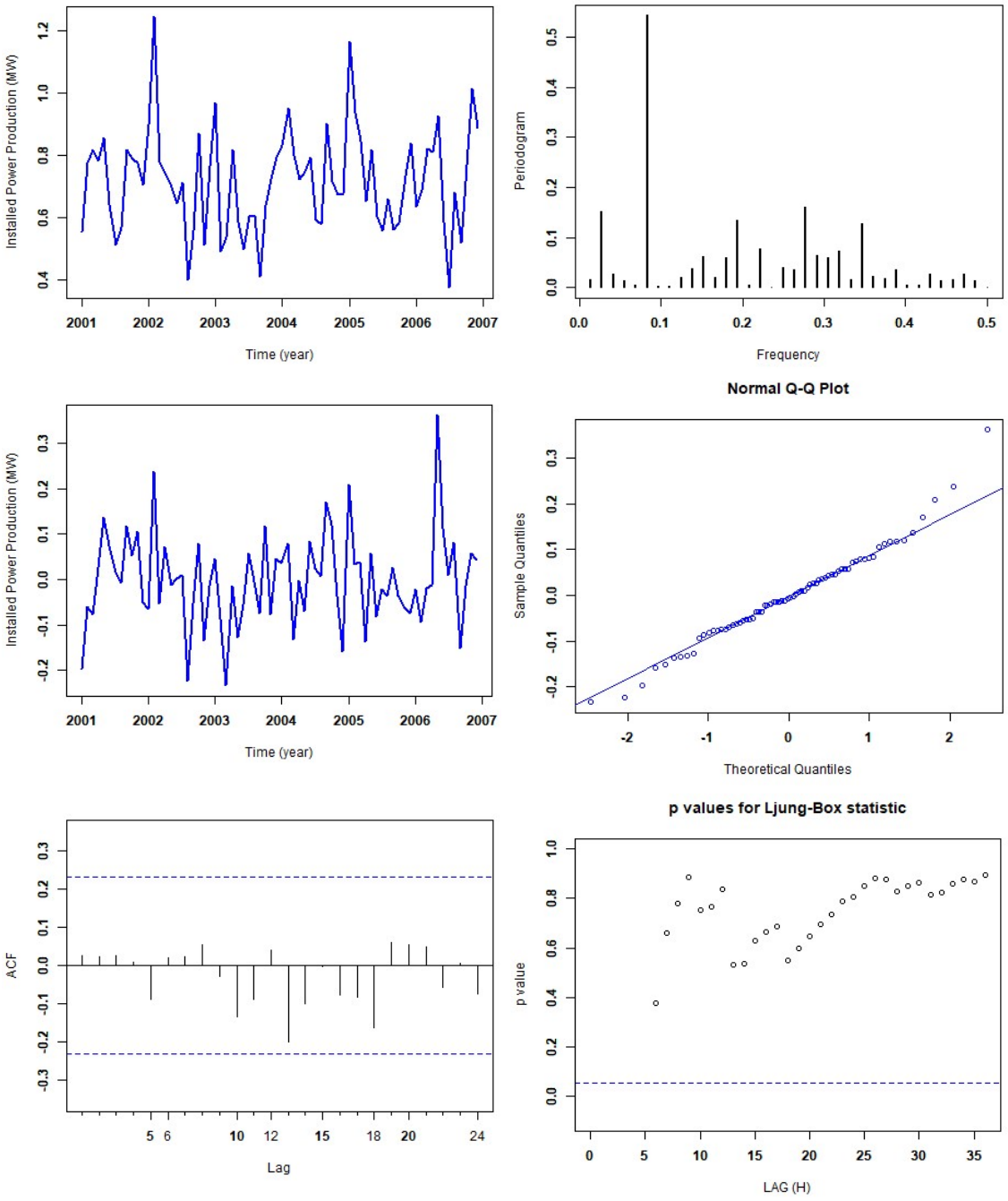


Figure B.16: Station 17: The fitted SARIMA model for installed power production is a SARIMA (1,0,3)(0,0,1)(12).

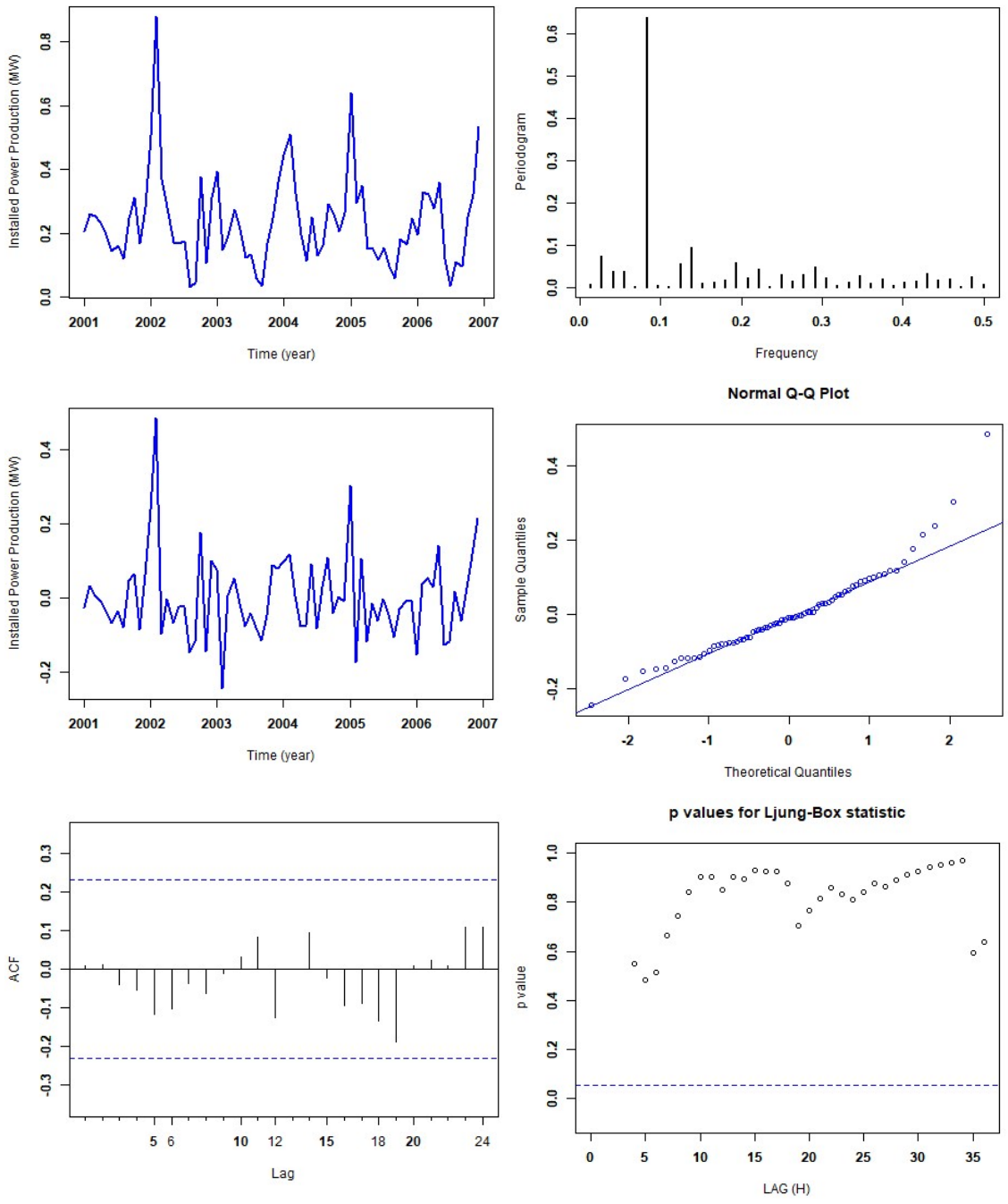


Figure B.17: Station 18: The fitted SARIMA model for installed power production is a SARIMA (1,0,0)(1,0,1)(12).

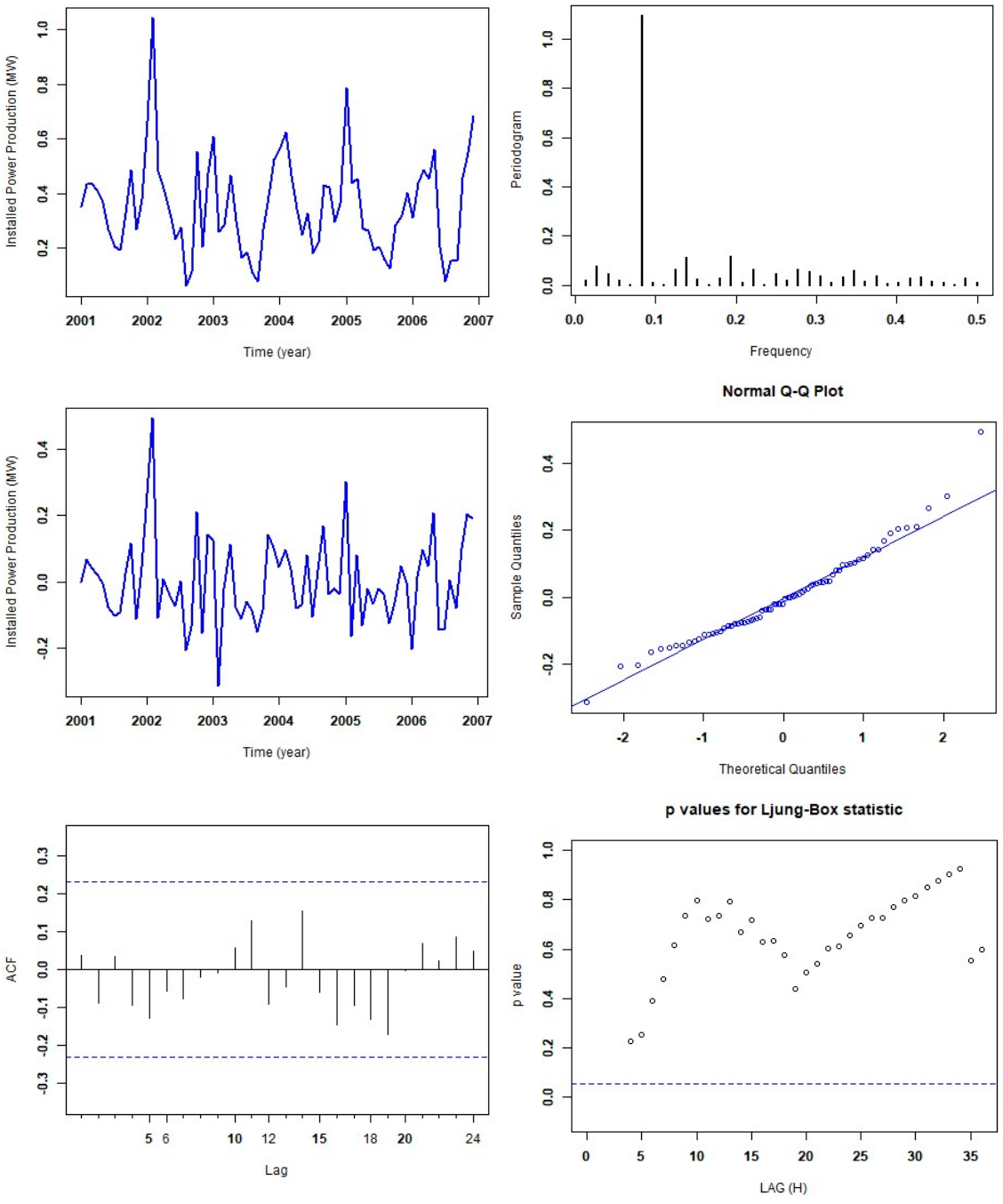


Figure B.18: Station 19: The fitted SARIMA model for installed power production is a SARIMA (1,0,0)(1,0,1)(12).

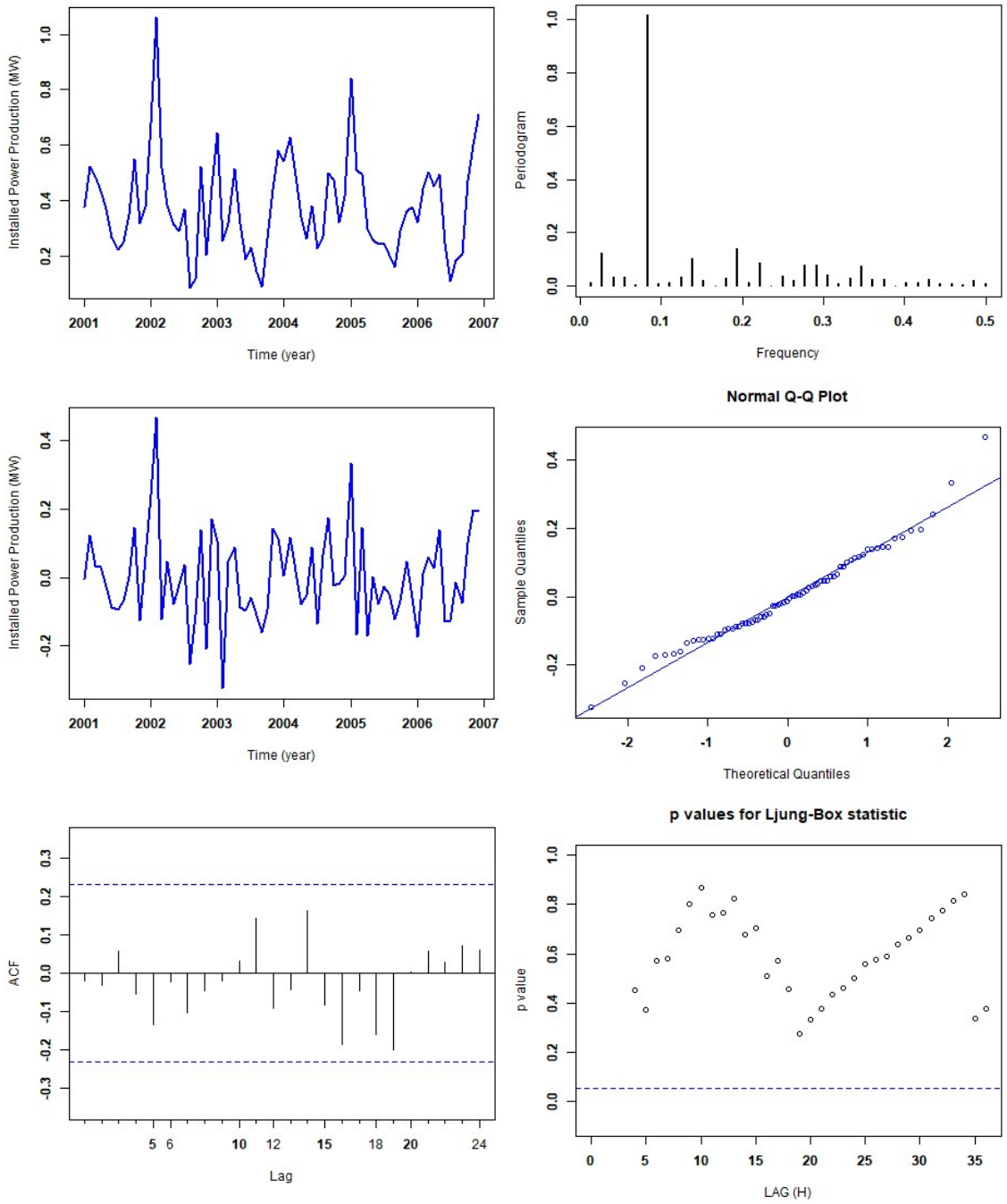


Figure B.19: Station 20: The fitted SARIMA model for installed power production is a SARIMA (0,0,1)(1,0,1)(12).

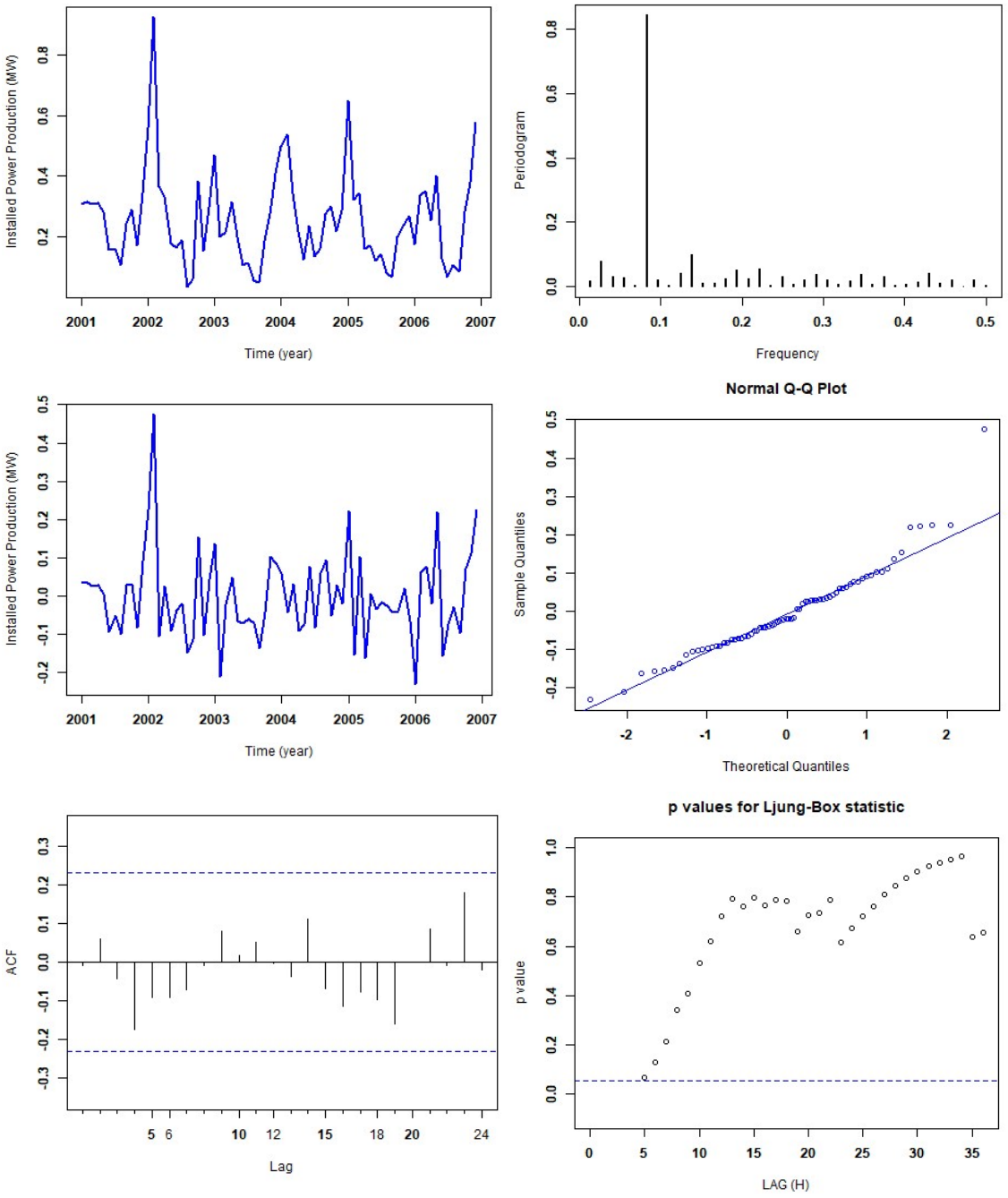


Figure B.20: Station 21: The fitted SARIMA model for installed power production is a SARIMA (1,0,0)(1,0,2)(12).

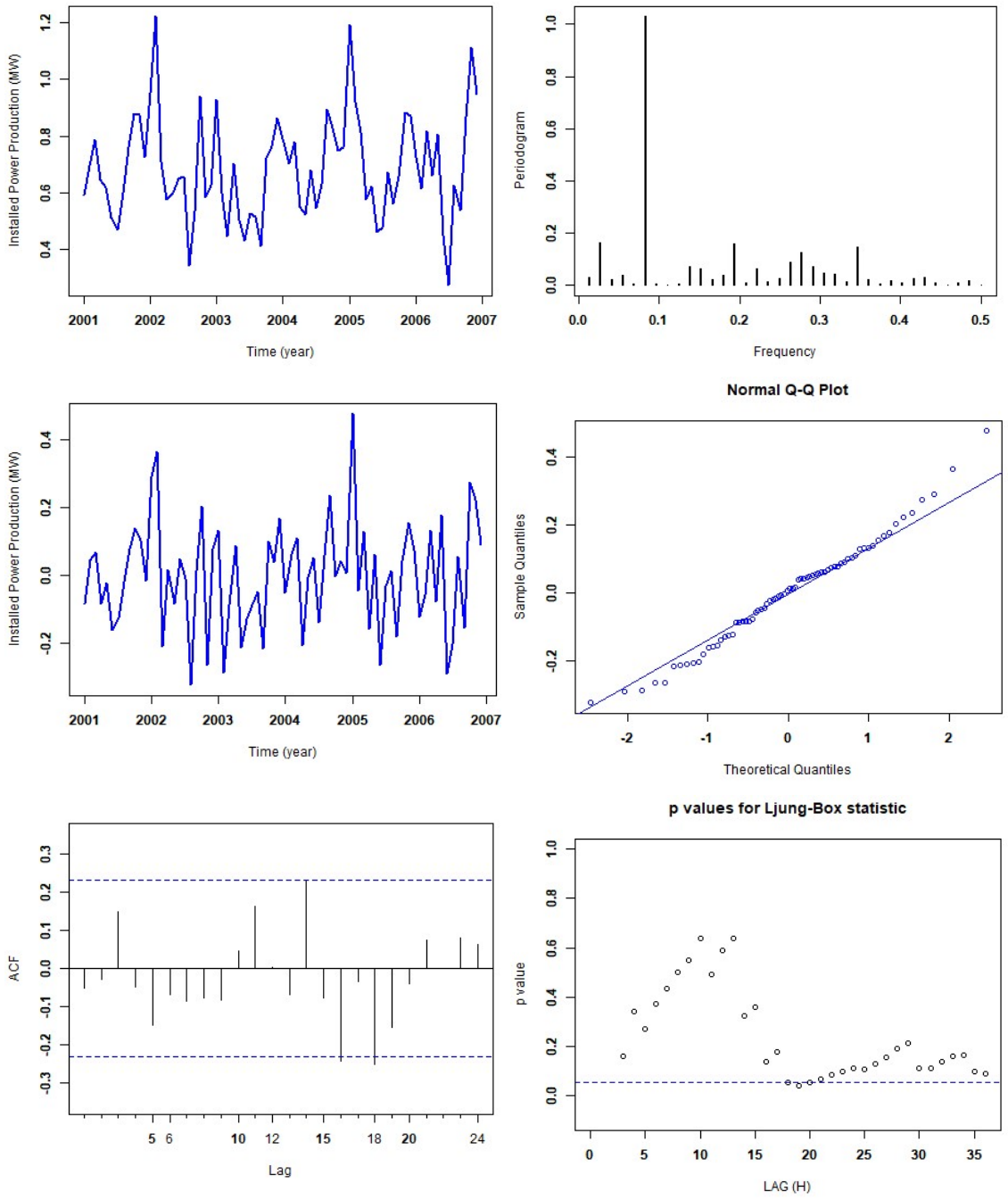


Figure B.21: Station 22: The fitted SARIMA model for installed power production is a SARIMA (0,0,1)(1,0,0)(12).

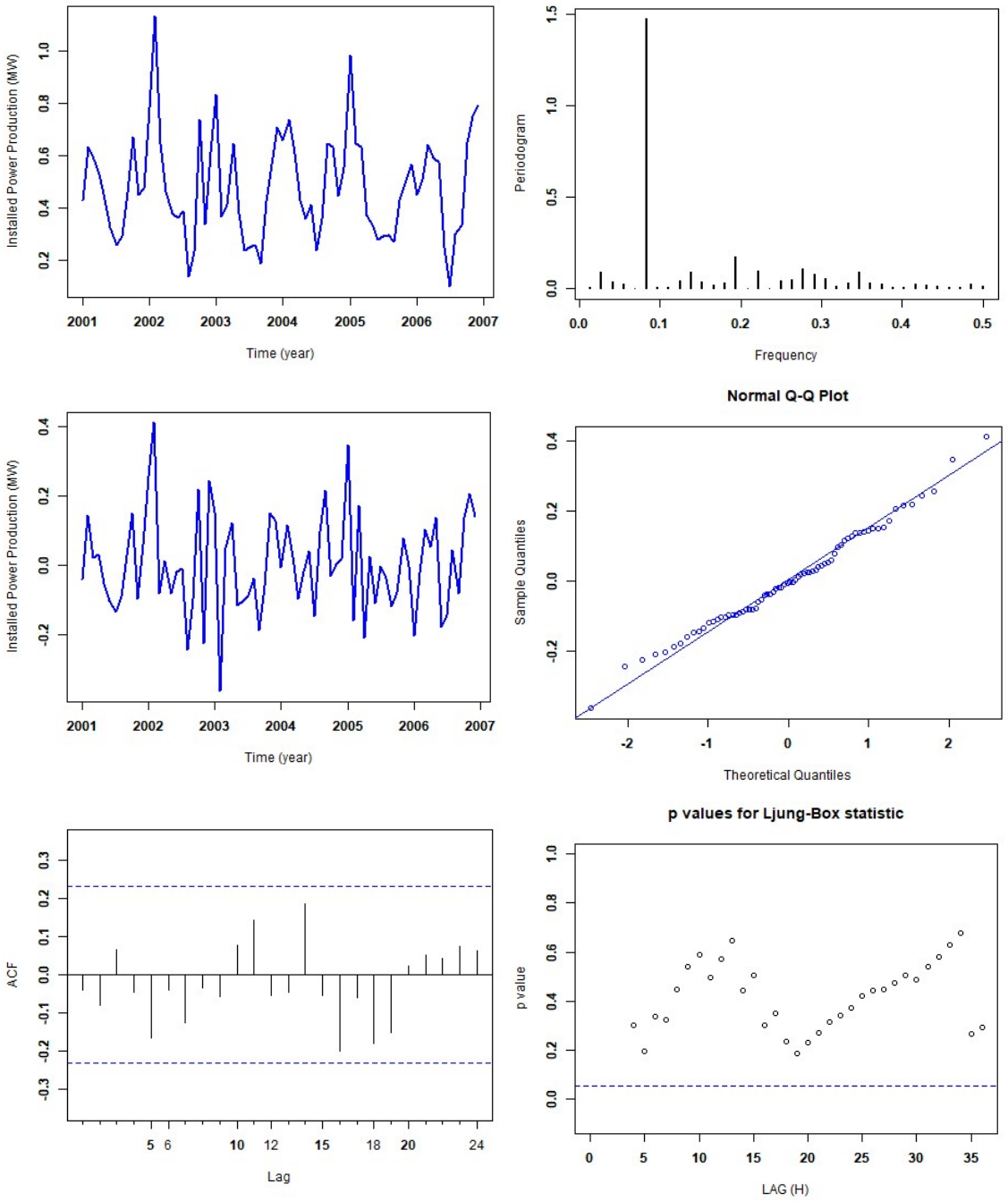


Figure B.22: Station 23: The fitted SARIMA model for installed power production is a SARIMA (0,0,1)(1,0,1)(12).

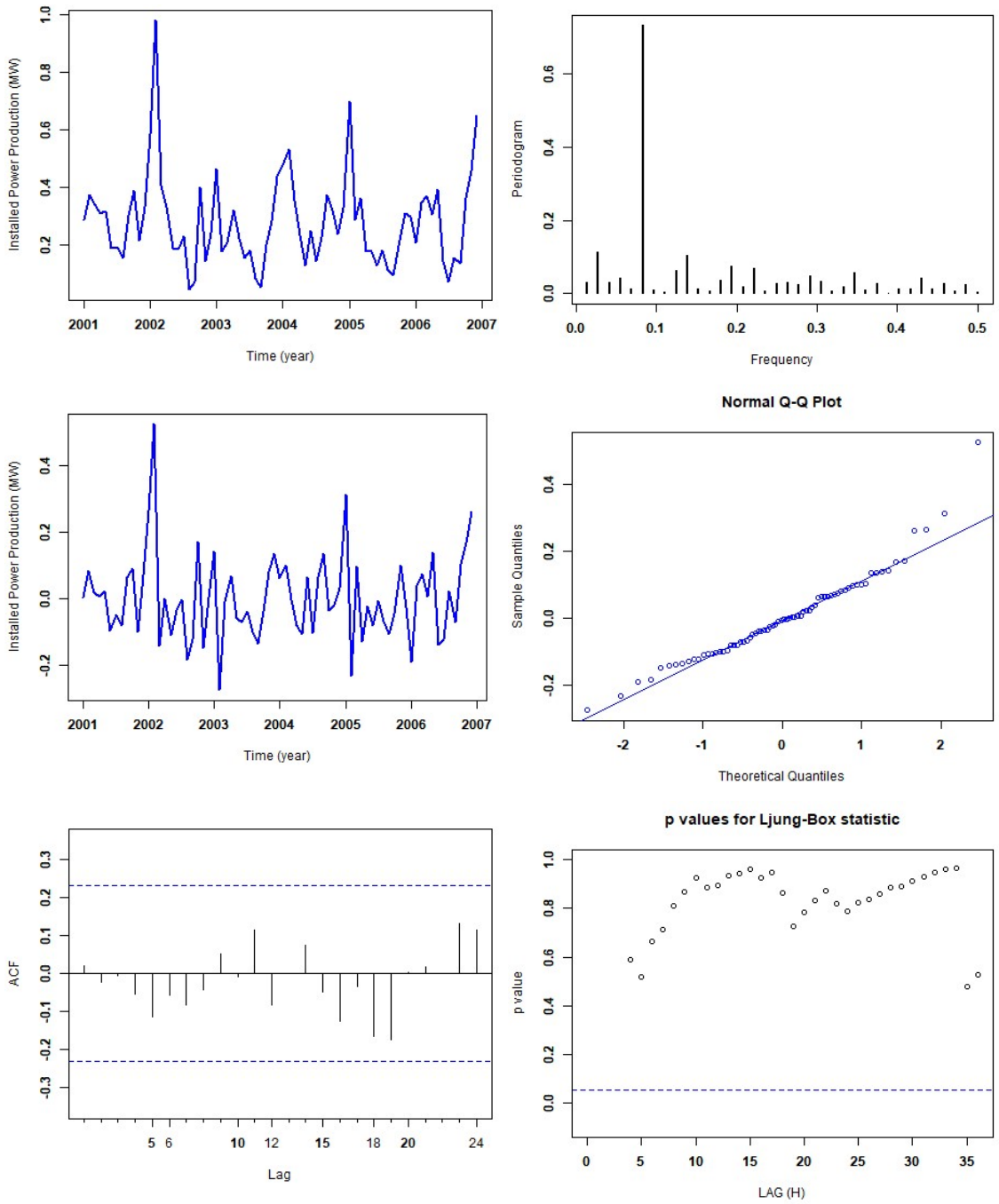


Figure B.23: Station 24: The fitted SARIMA model for installed power production is a SARIMA (1,0,0)(1,0,1)(12).

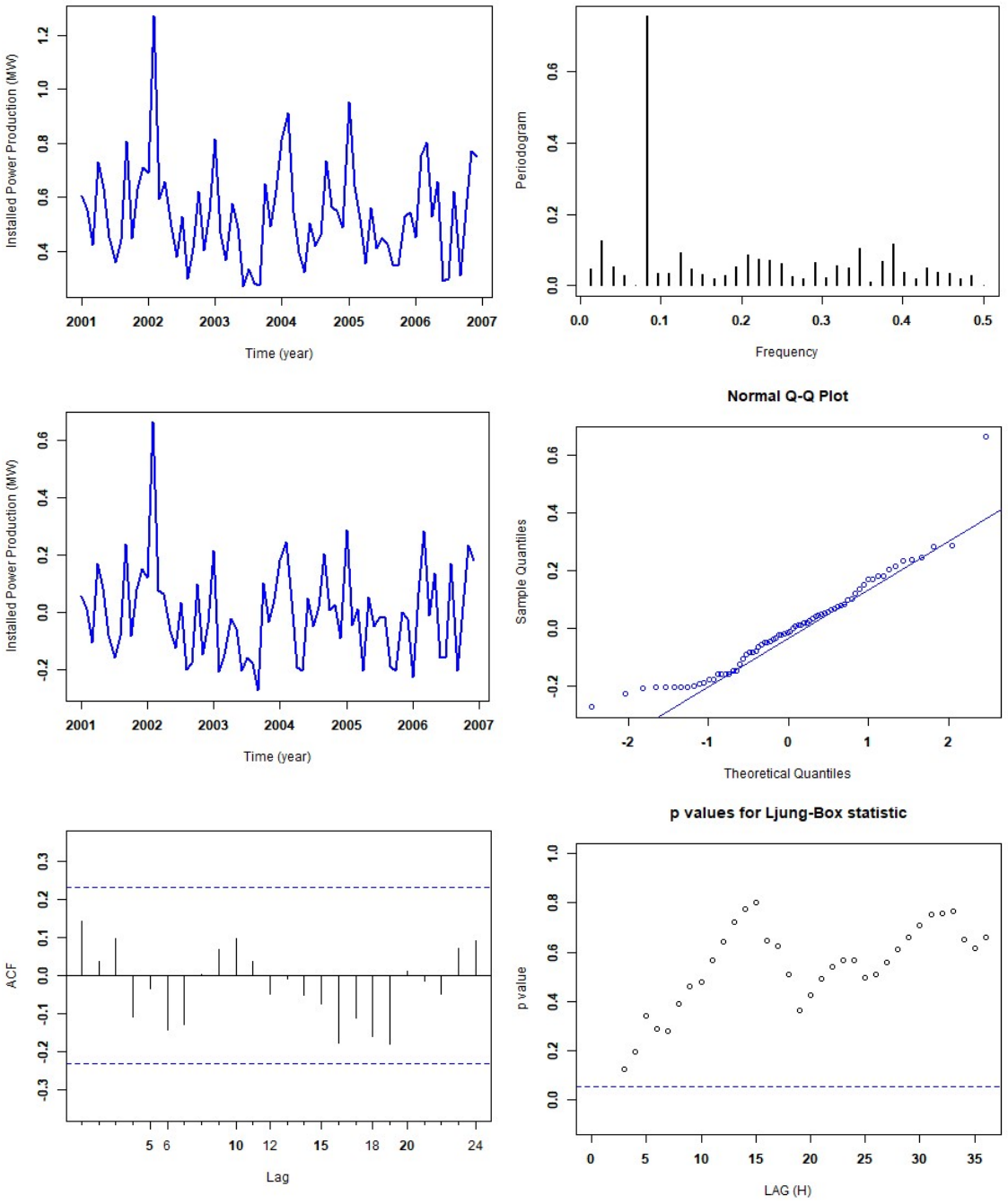


Figure B.24: Station 25: The fitted SARIMA model for installed power production is a SARIMA (0,0,0)(1,0,1)(12).

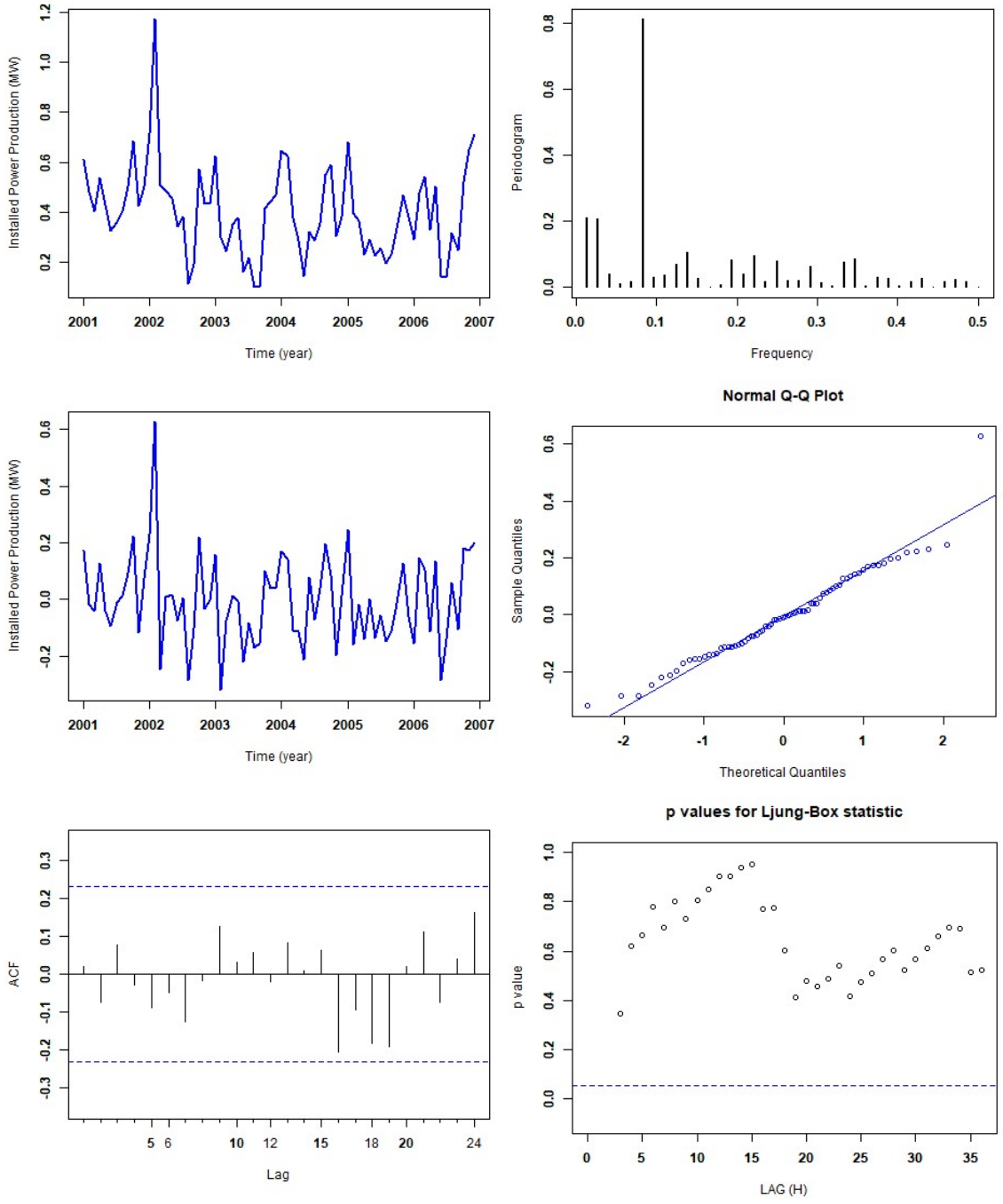


Figure B.25: Station 26: The fitted SARIMA model for installed power production is a SARIMA (1,0,0)(1,0,0)(12).

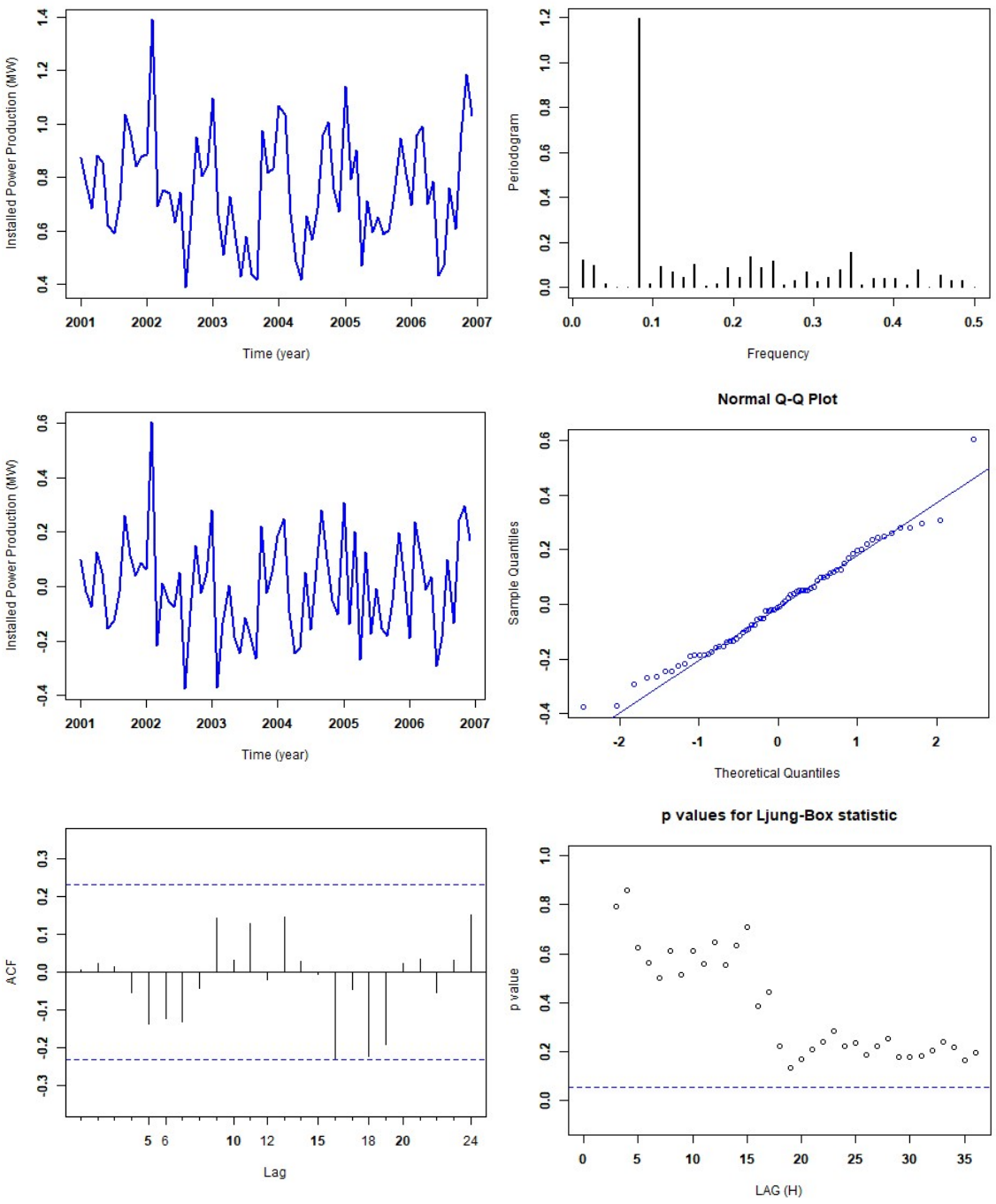


Figure B.26: Station 27: The fitted SARIMA model for installed power production is a SARIMA (0,0,1)(1,0,0)(12).

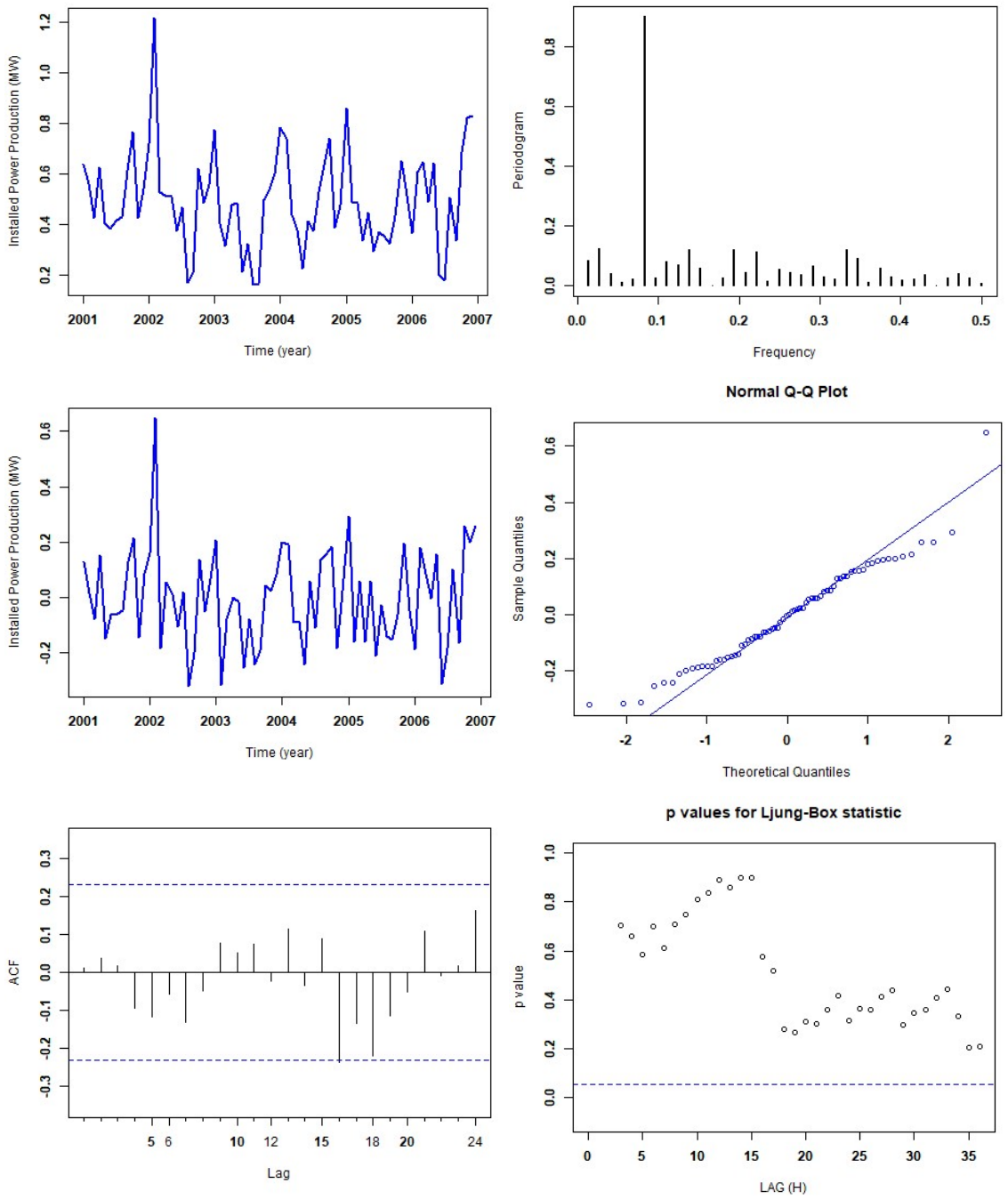


Figure B.27: Station 28: The fitted SARIMA model for installed power production is a SARIMA (0,0,1)(1,0,0)(12).

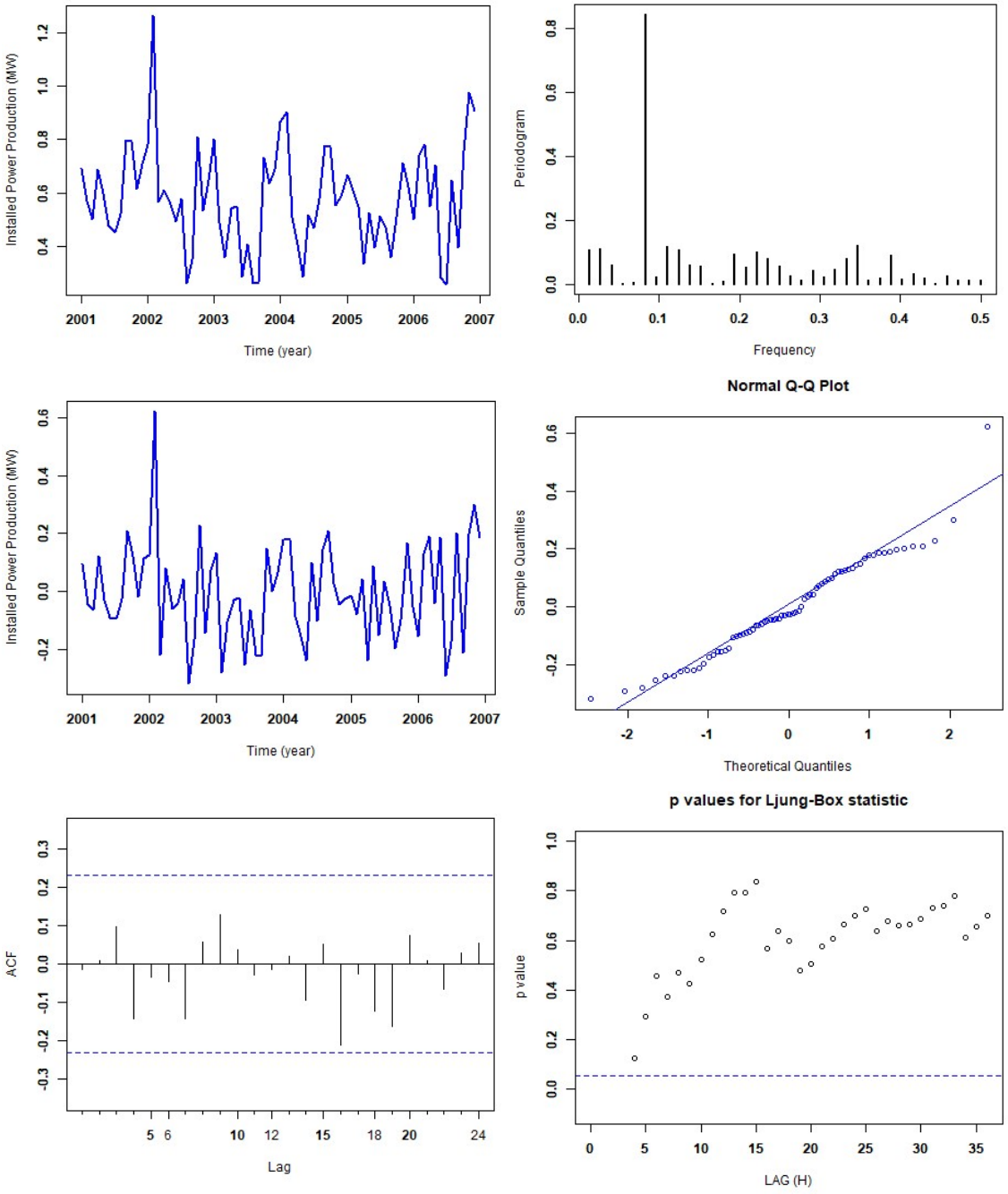


Figure B.28: Station 29: The fitted SARIMA model for installed power production is a SARIMA (0,0,1)(1,0,1)(12).

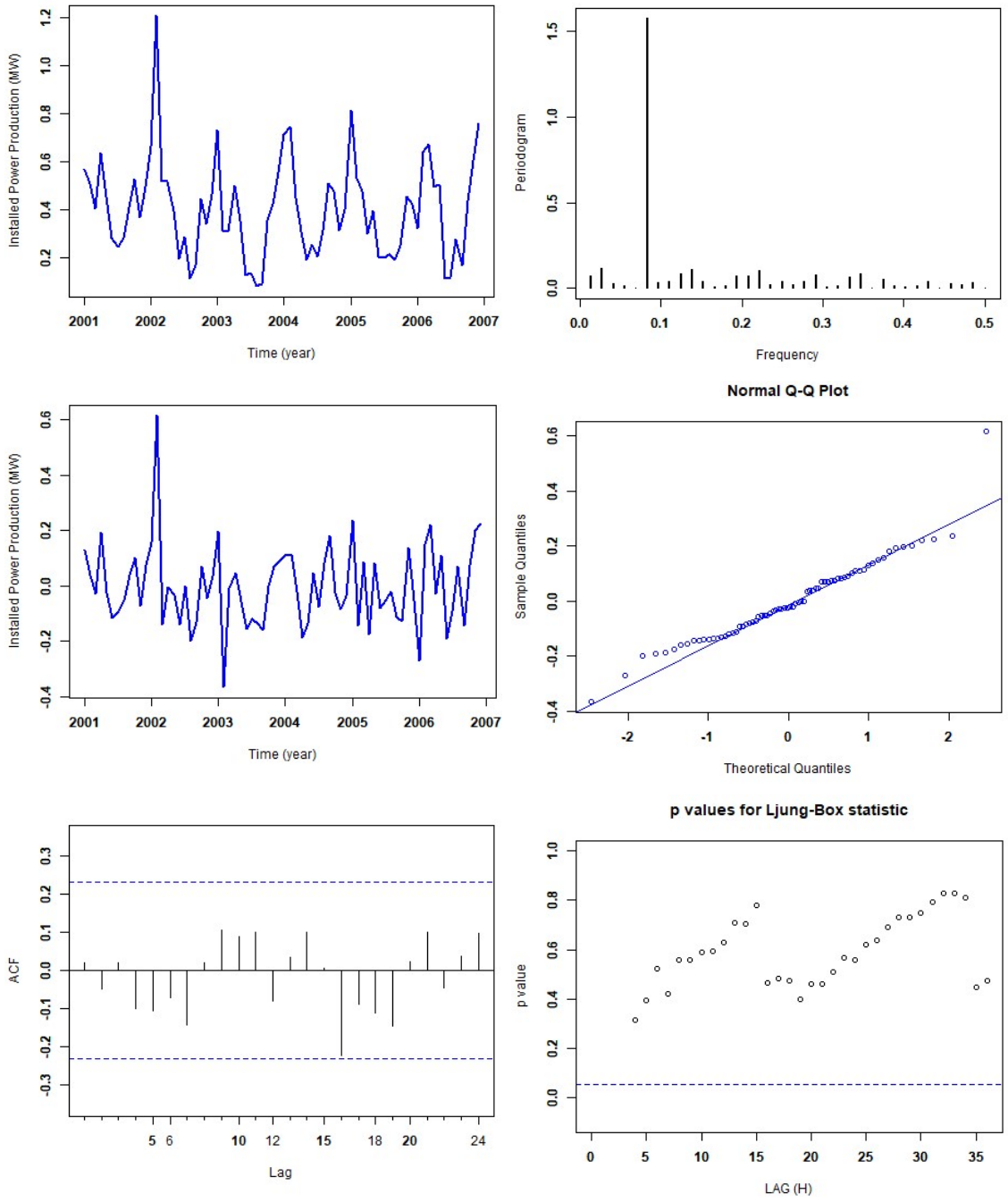


Figure B.29: Station 30: The fitted SARIMA model for installed power production is a SARIMA (1,0,0)(1,0,1)(12).

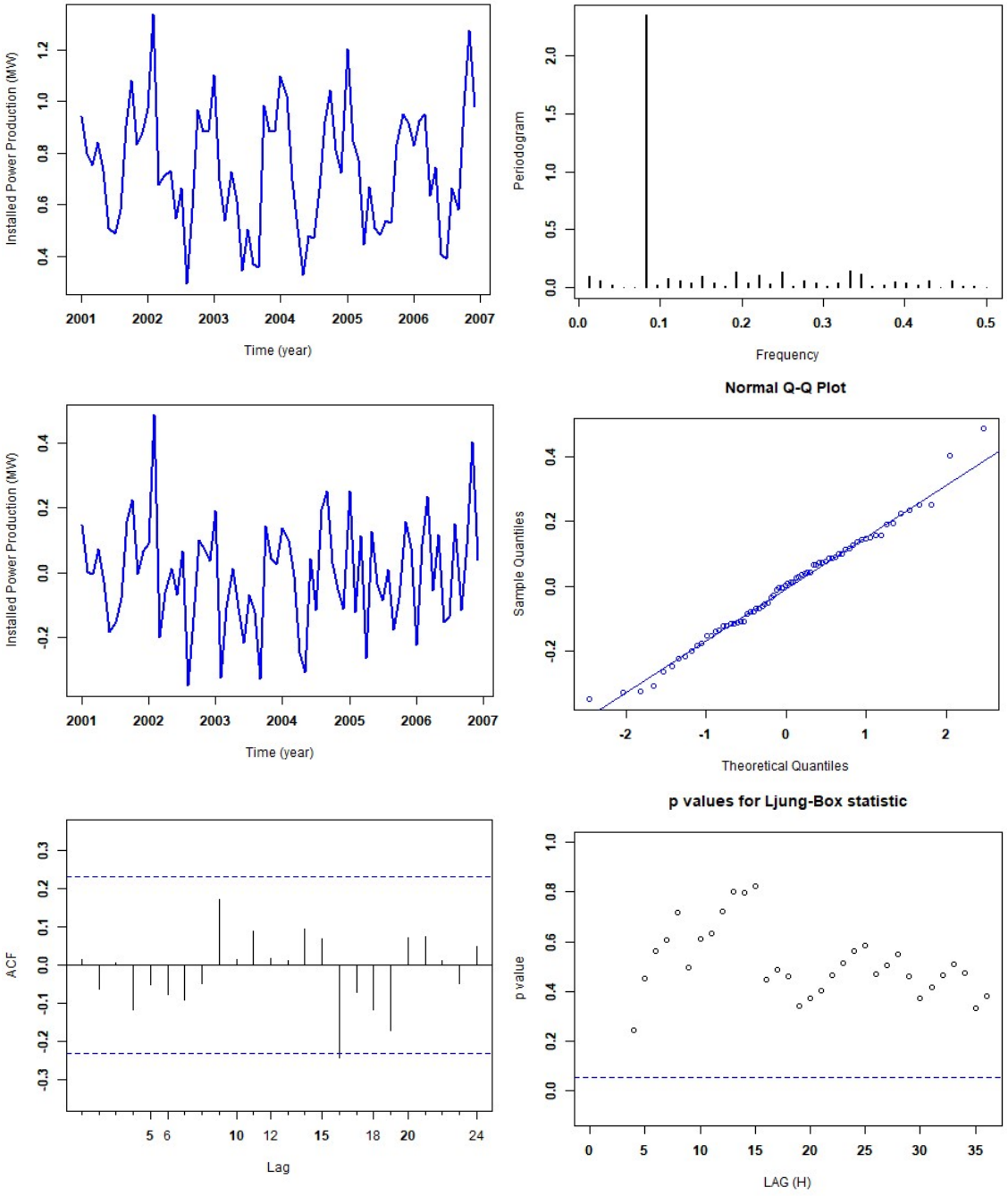


Figure B.30: Station 32: The fitted SARIMA model for installed power production is a SARIMA (1,0,0)(1,0,1)(12).

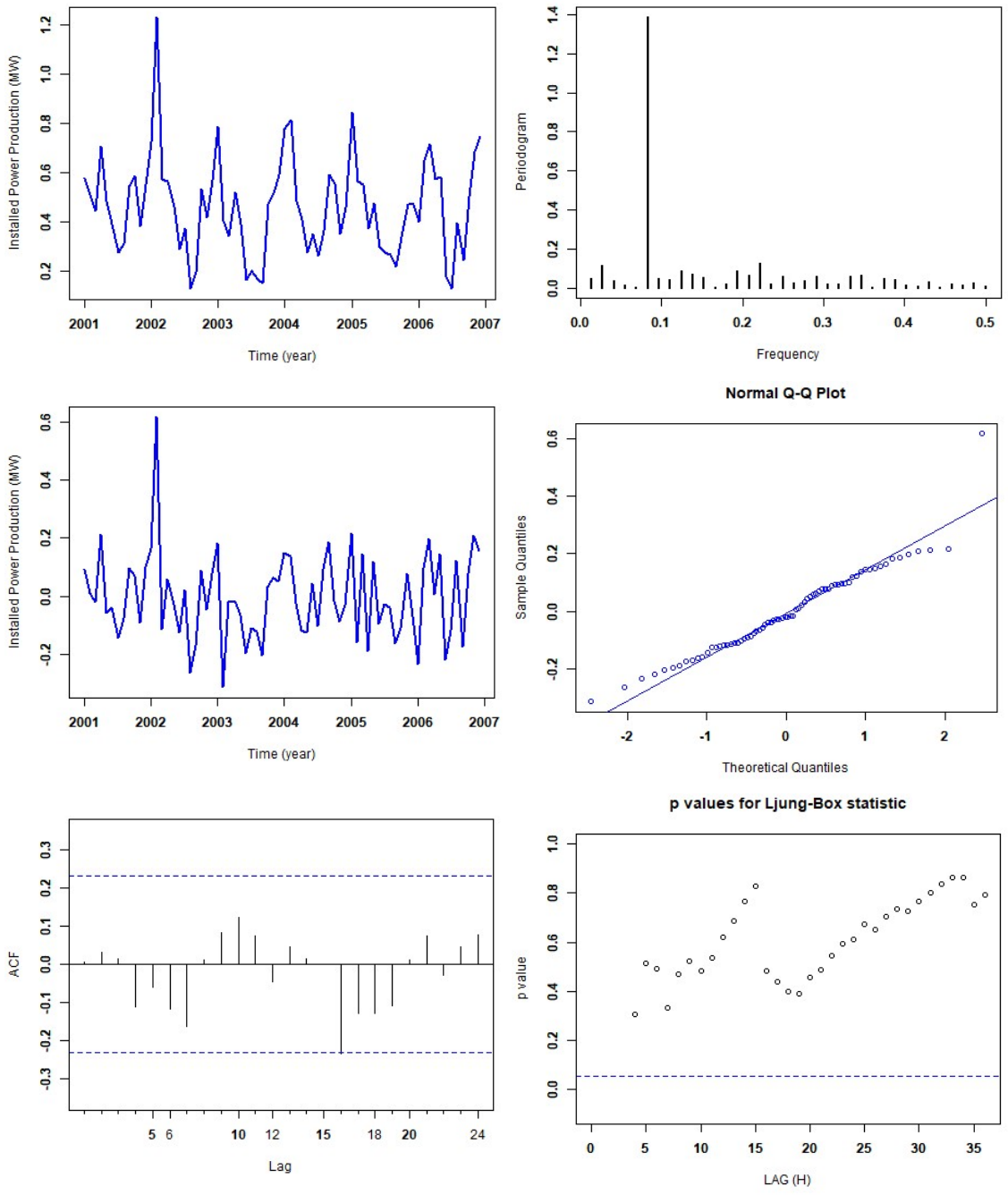


Figure B.31: Station 33: The fitted SARIMA model for installed power production is a SARIMA (0,0,1)(1,0,1)(12).

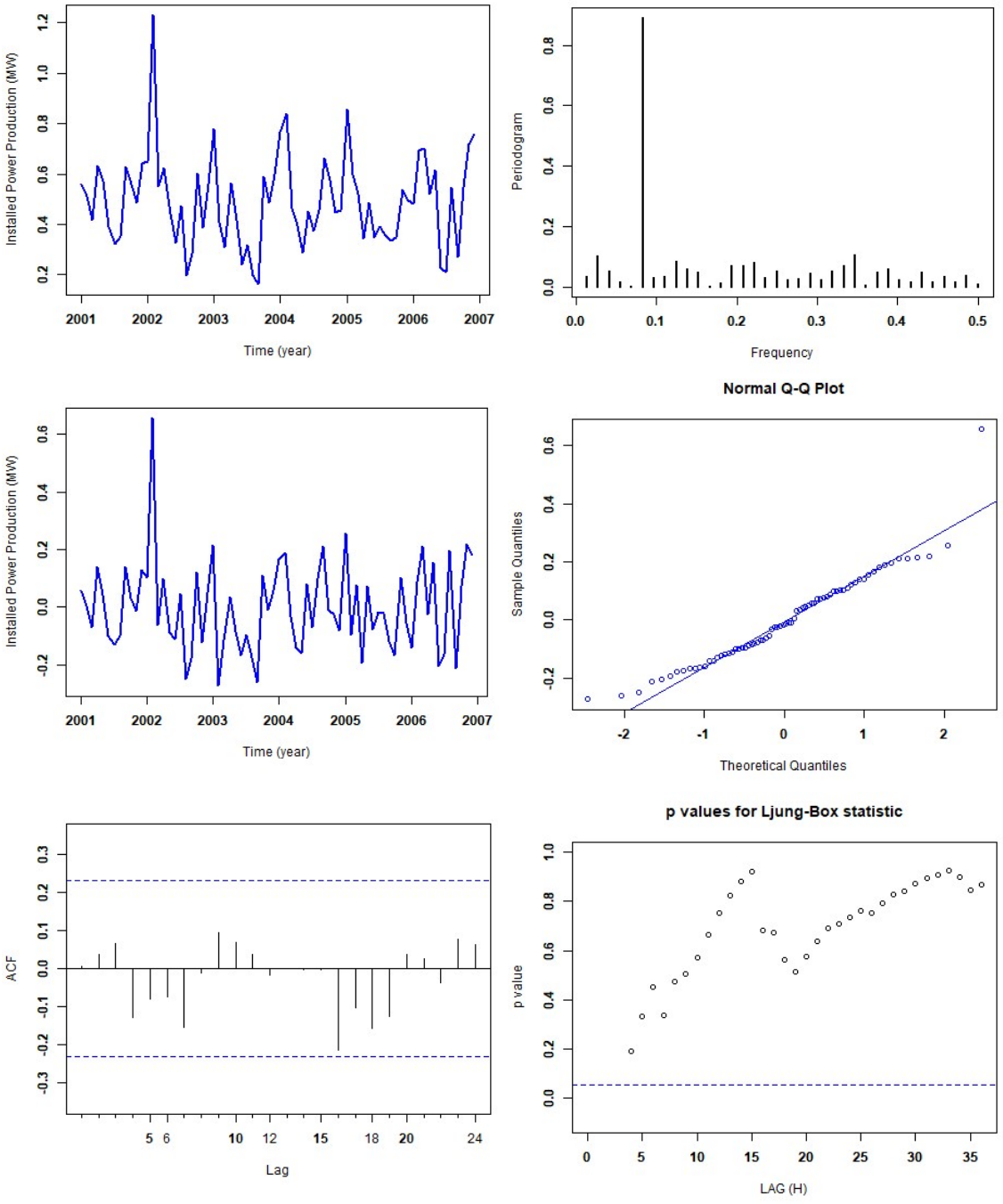


Figure B.32: Station 34: The fitted SARIMA model for installed power production is a SARIMA (0,0,1)(1,0,1)(12).

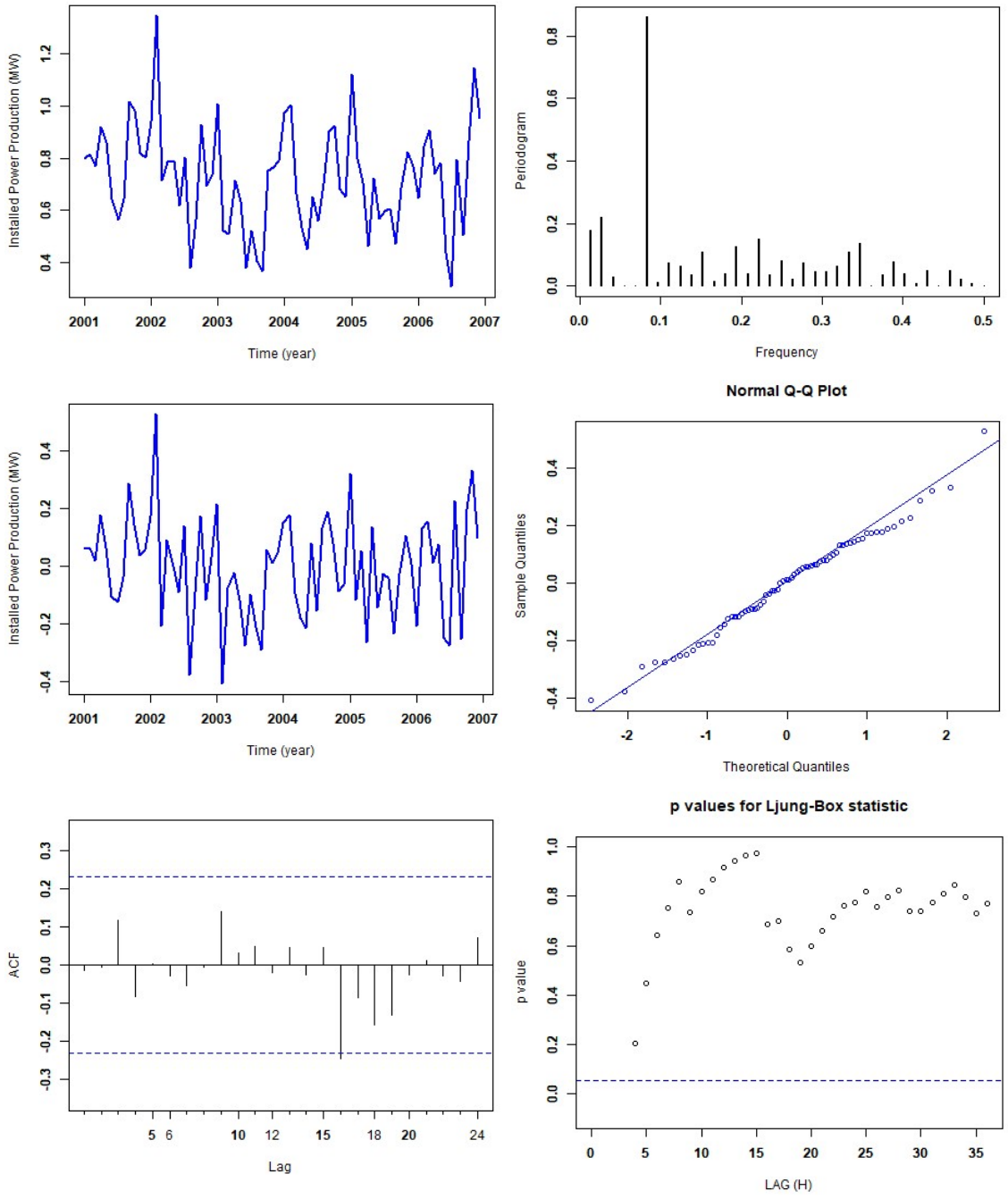


Figure B.33: Station 35: The fitted SARIMA model for installed power production is a SARIMA (0,0,1)(1,0,1)(12).

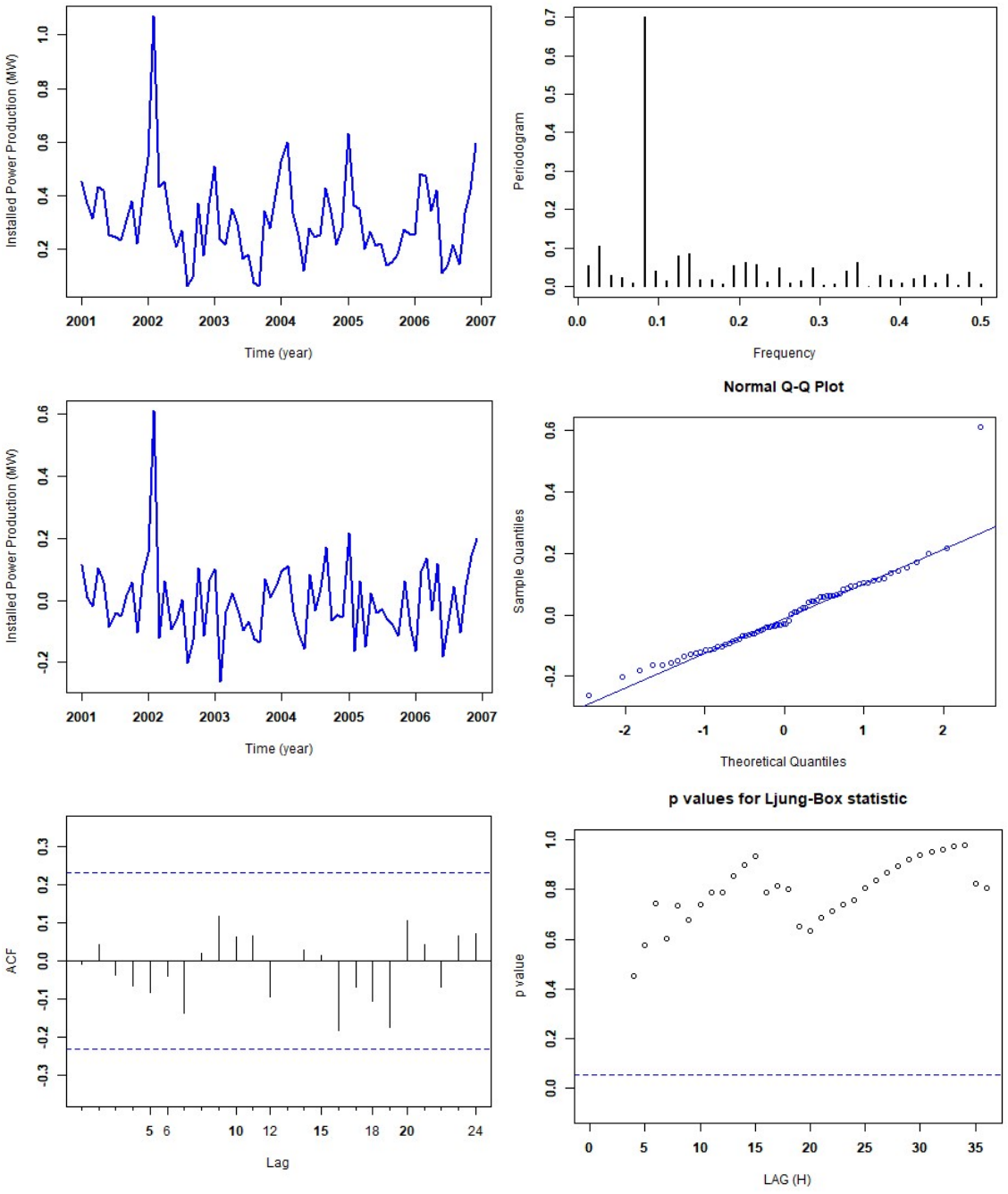


Figure B.34: Station 36: The fitted SARIMA model for installed power production is a SARIMA (1,0,0)(1,0,1)(12).

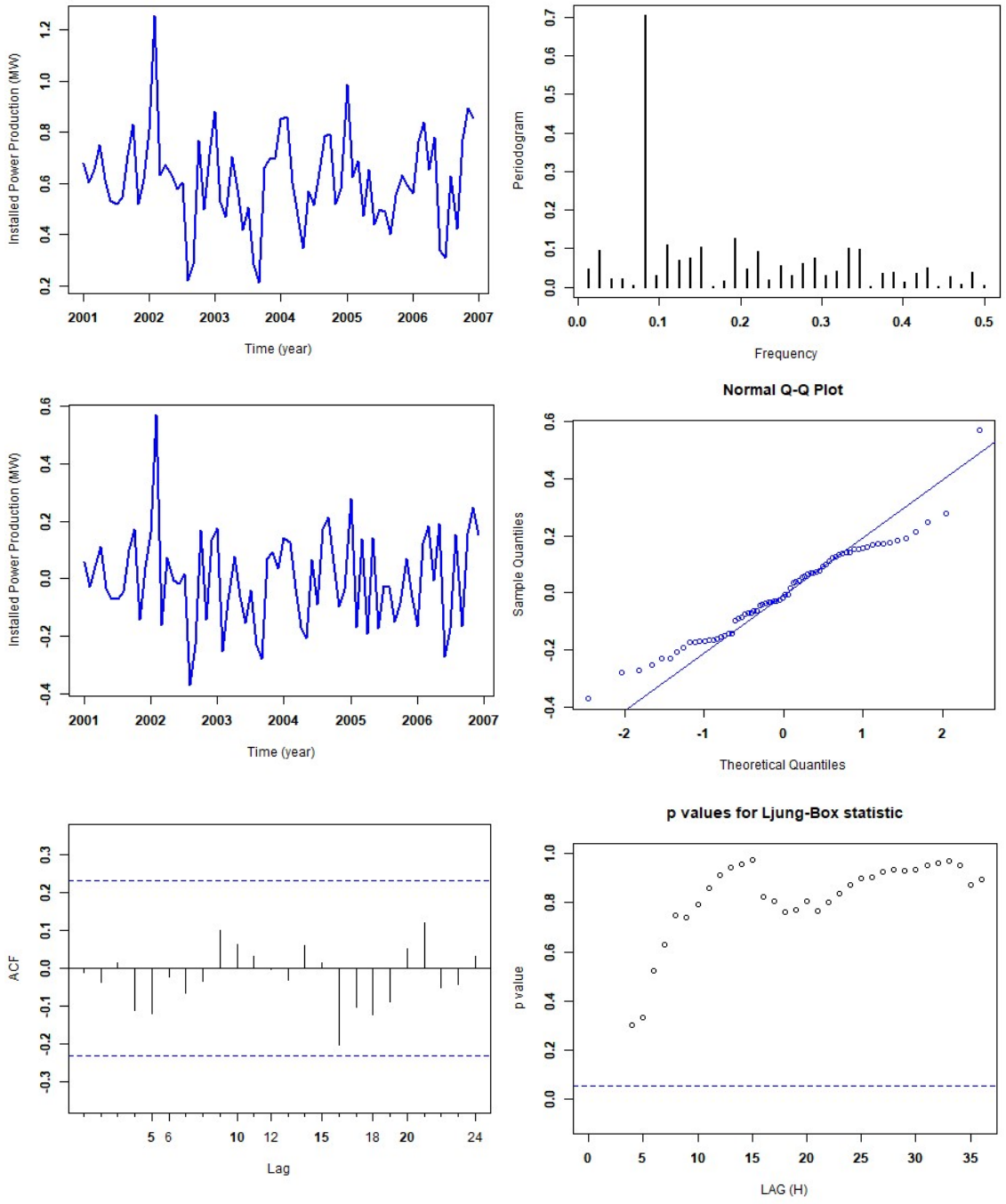


Figure B.35: Station 37: The fitted SARIMA model for installed power production is a SARIMA (0,0,1)(1,0,1)(12).

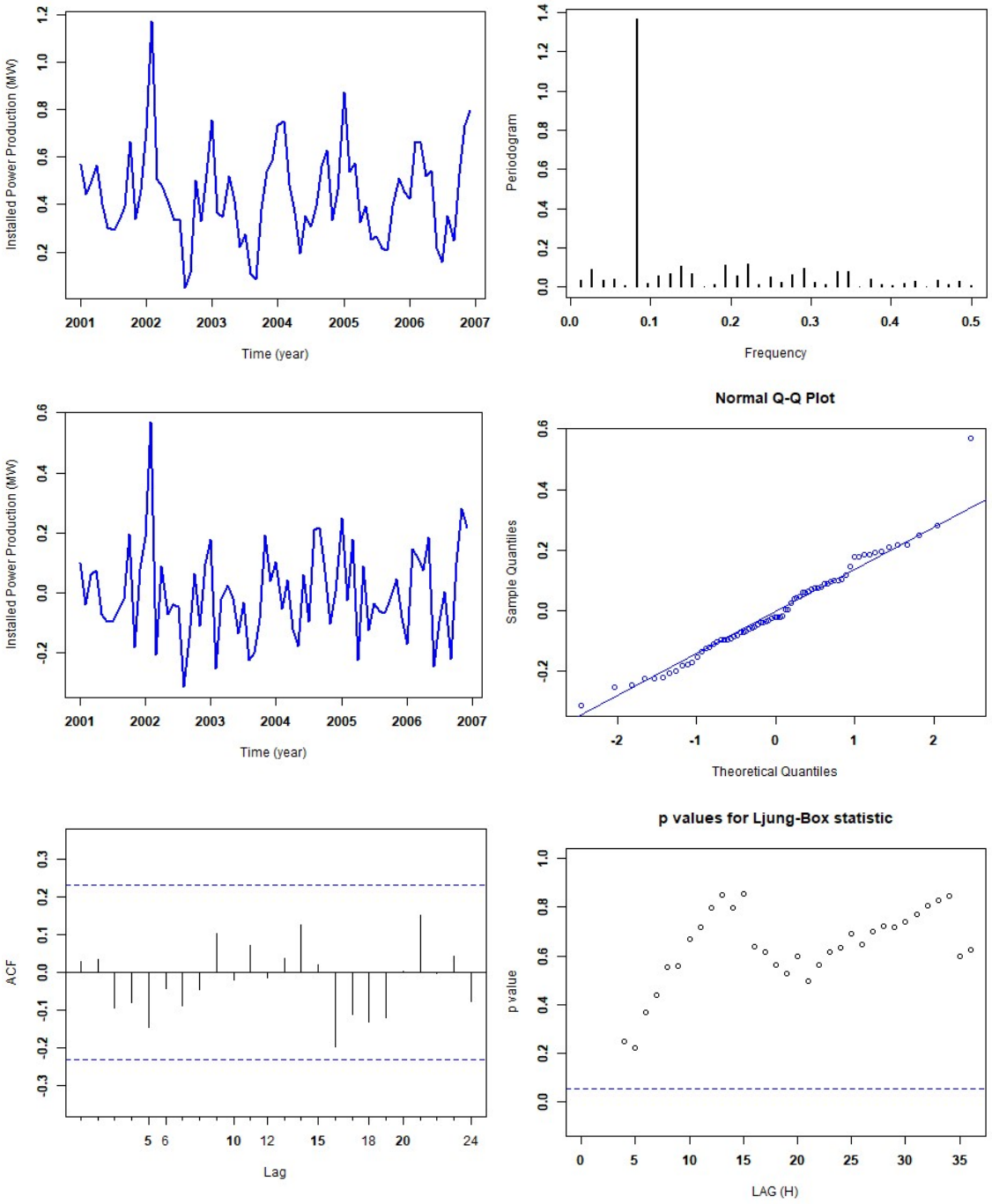


Figure B.36: Station 38: The fitted SARIMA model for installed power production is a SARIMA (0,0,1)(2,0,0)(12).

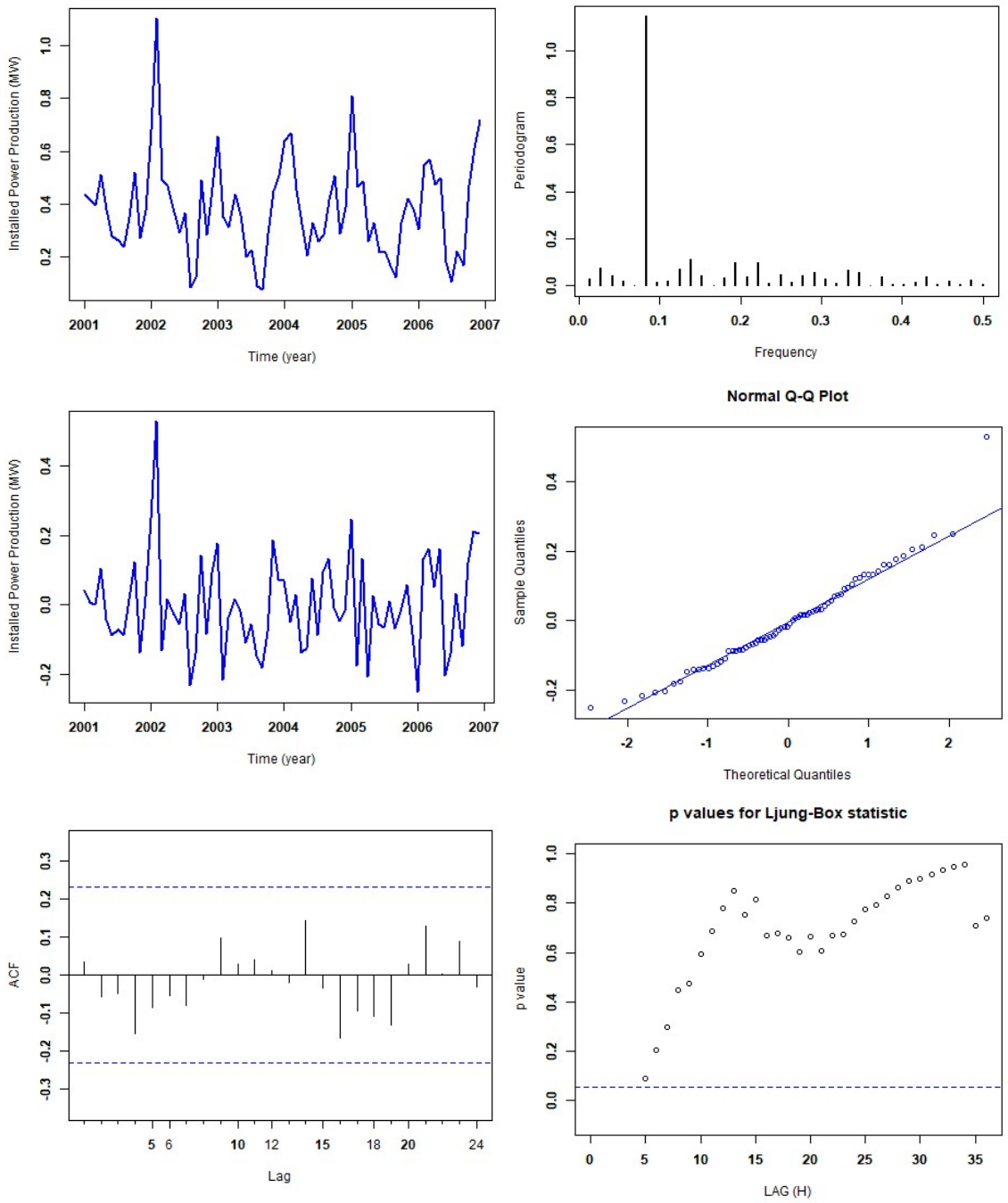


Figure B.37: Station 39: The fitted SARIMA model for installed power production is a SARIMA (1,0,0)(1,0,2)(12).

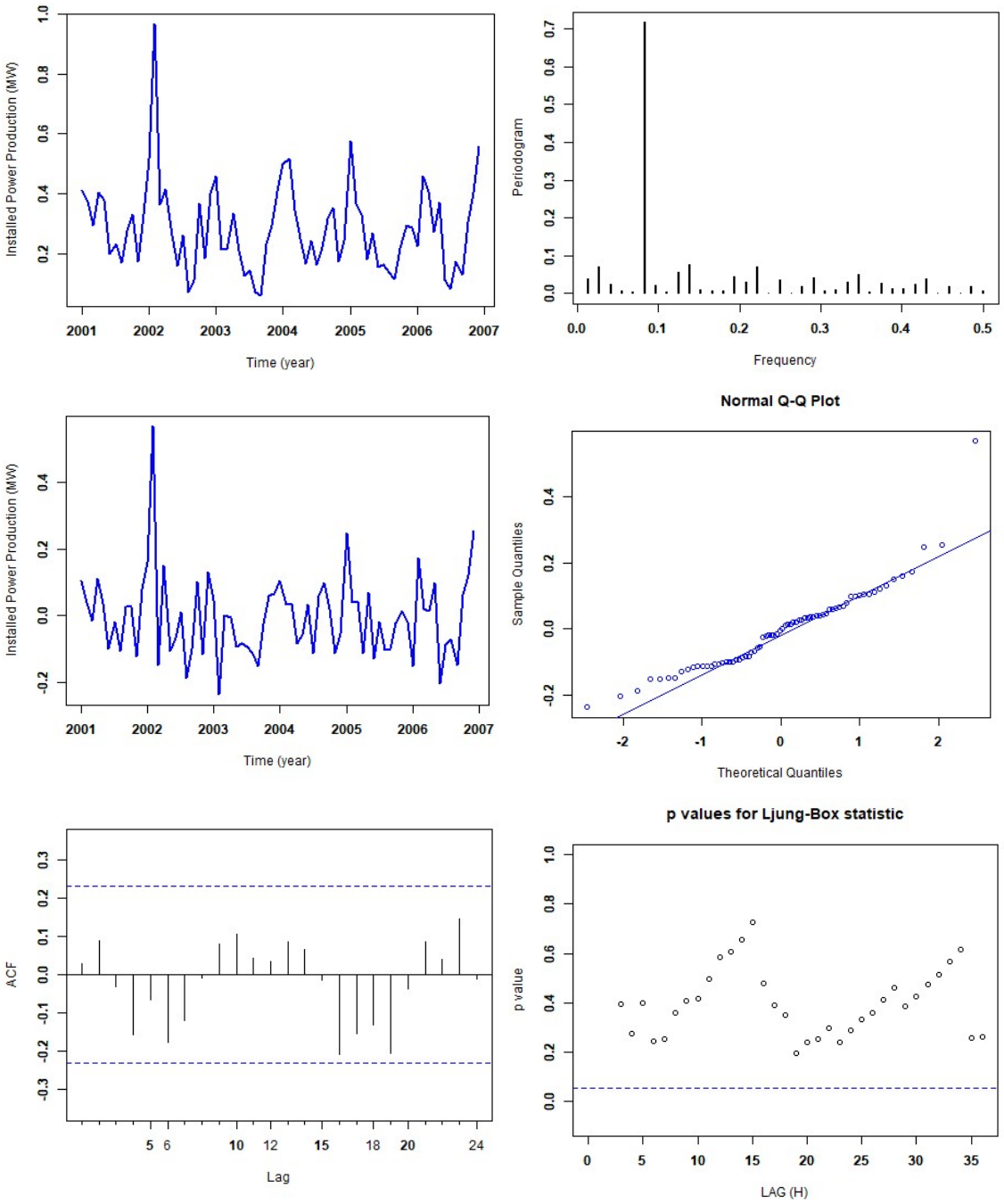


Figure B.38: Station 40: The fitted SARIMA model for installed power production is a SARIMA (0,0,1)(0,0,1)(12).

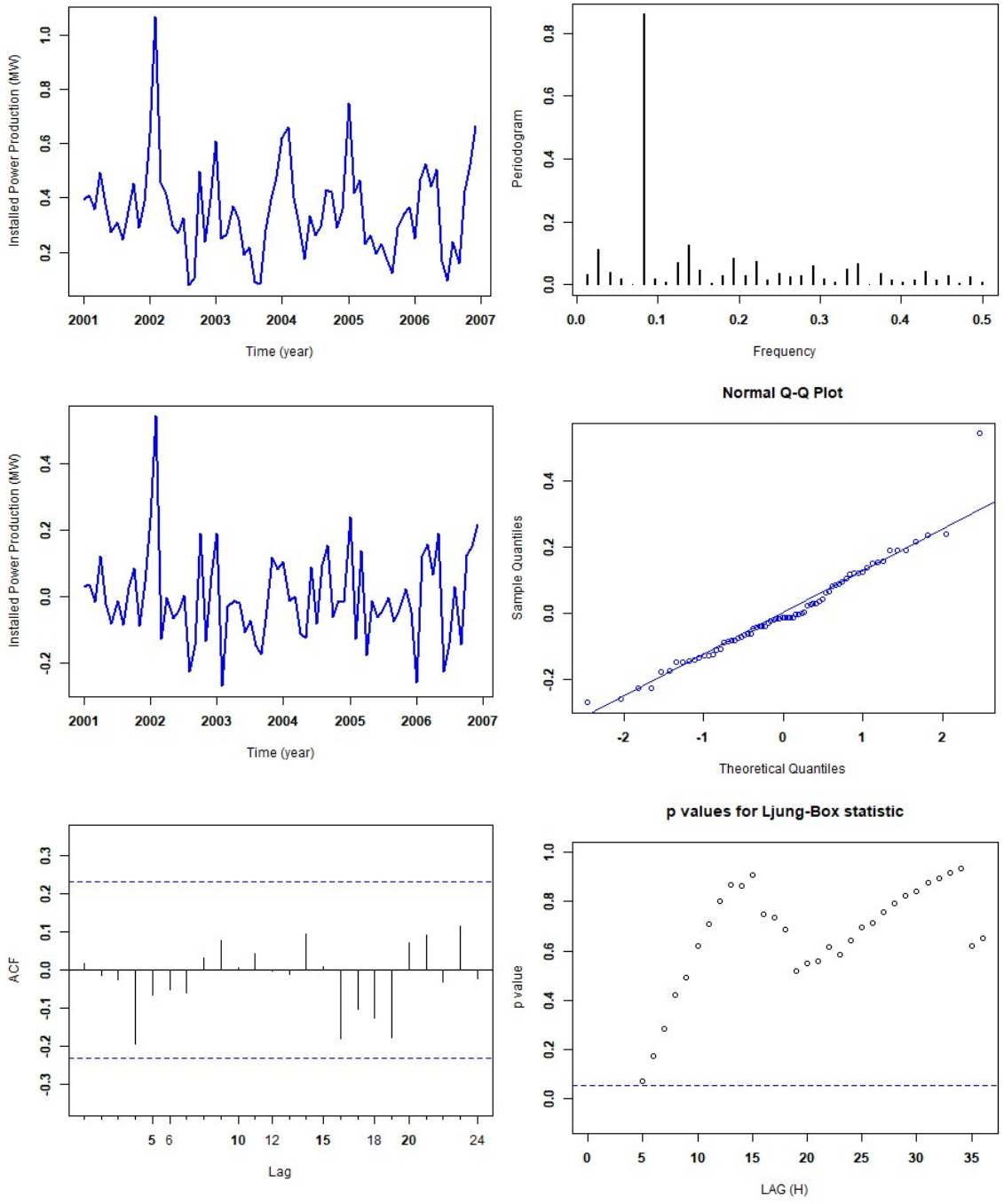


Figure B.39: Station 41: The fitted SARIMA model for installed power production is a SARIMA (1,0,0)(1,0,2)(12).

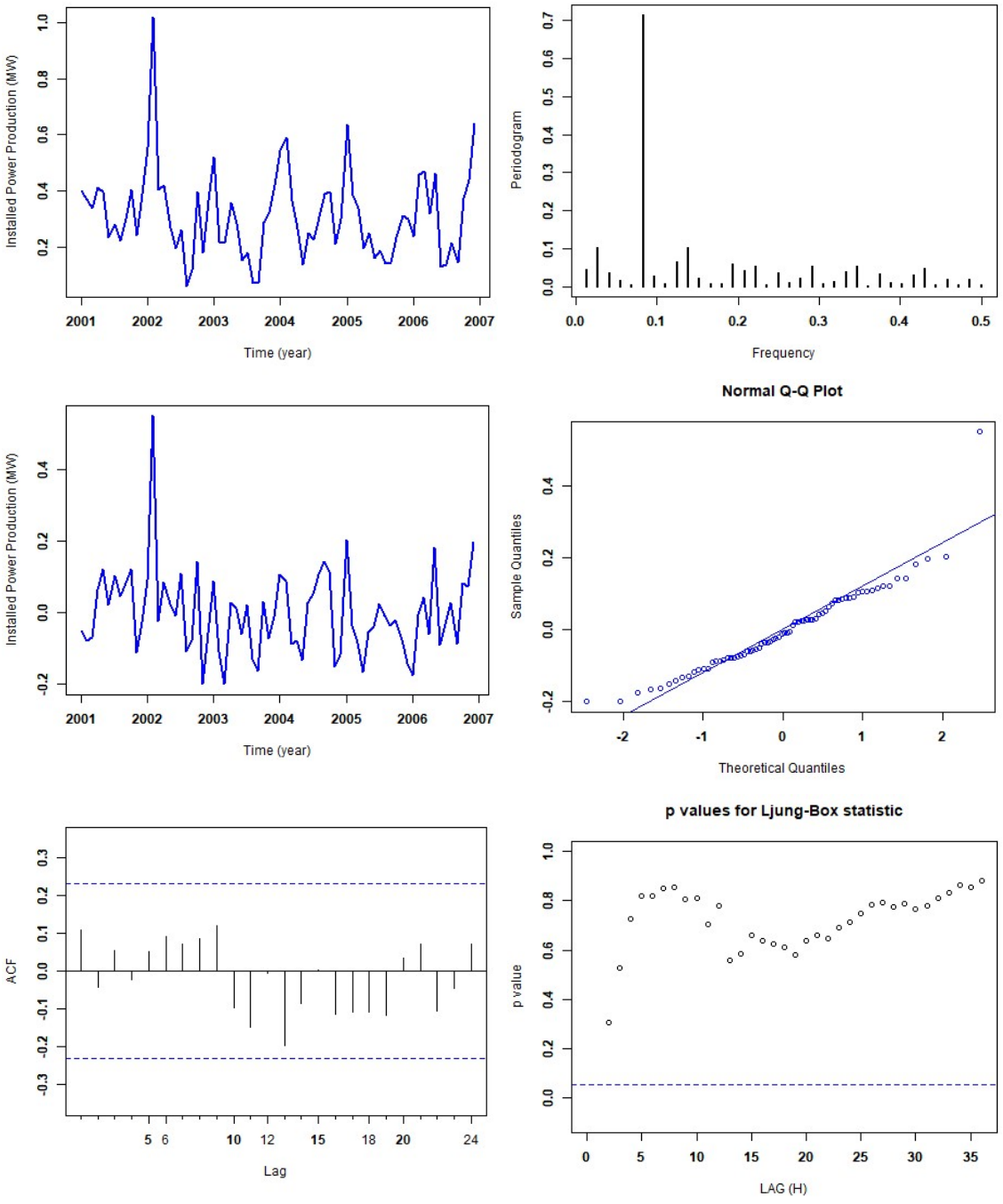


Figure B.40: Station 42: The fitted SARIMA model for installed power production is a SARIMA (0,0,0)(1,0,0)(12).

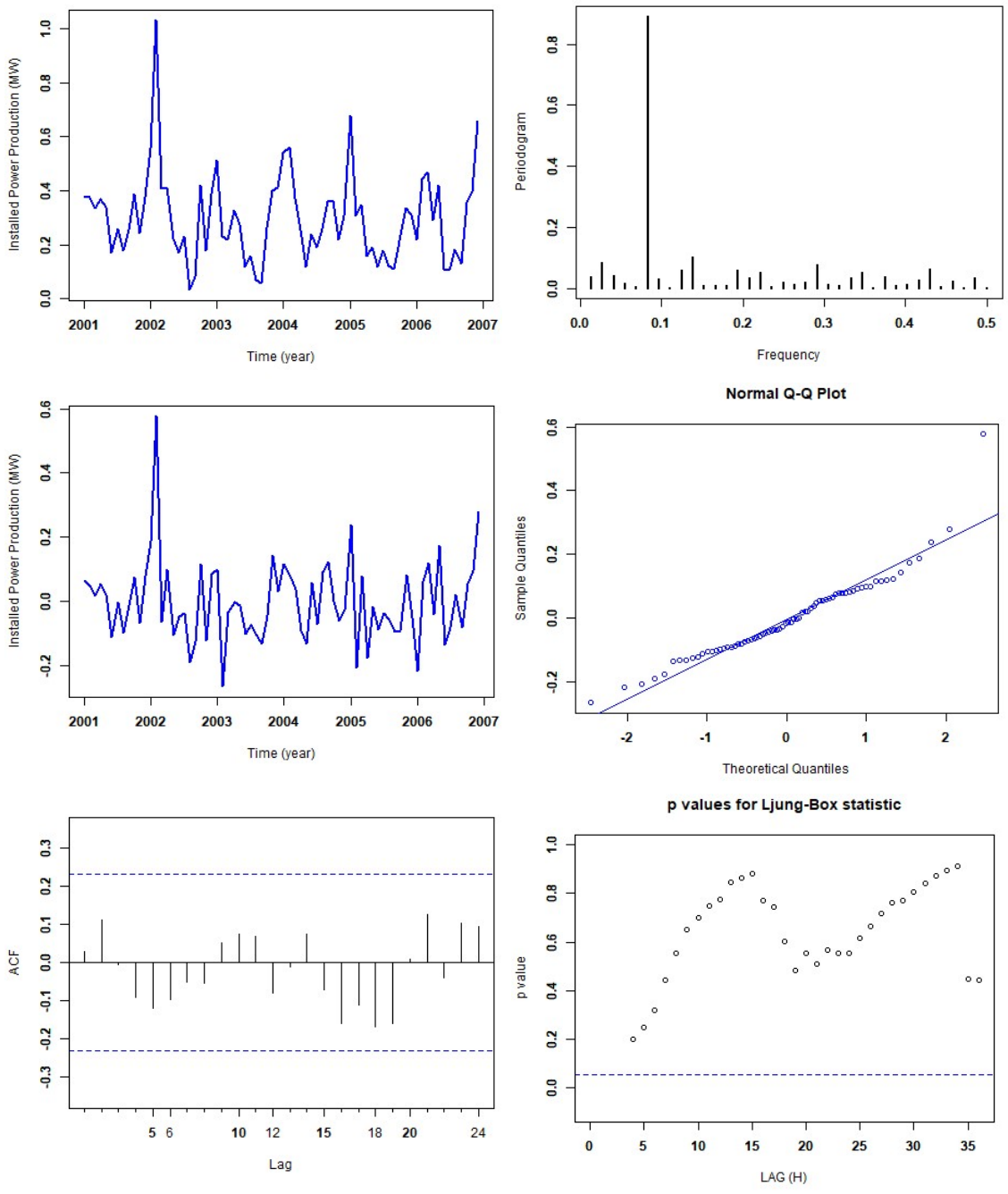


Figure B.41: Station 43: The fitted SARIMA model for installed power production is a SARIMA (0,0,1)(1,0,1)(12).

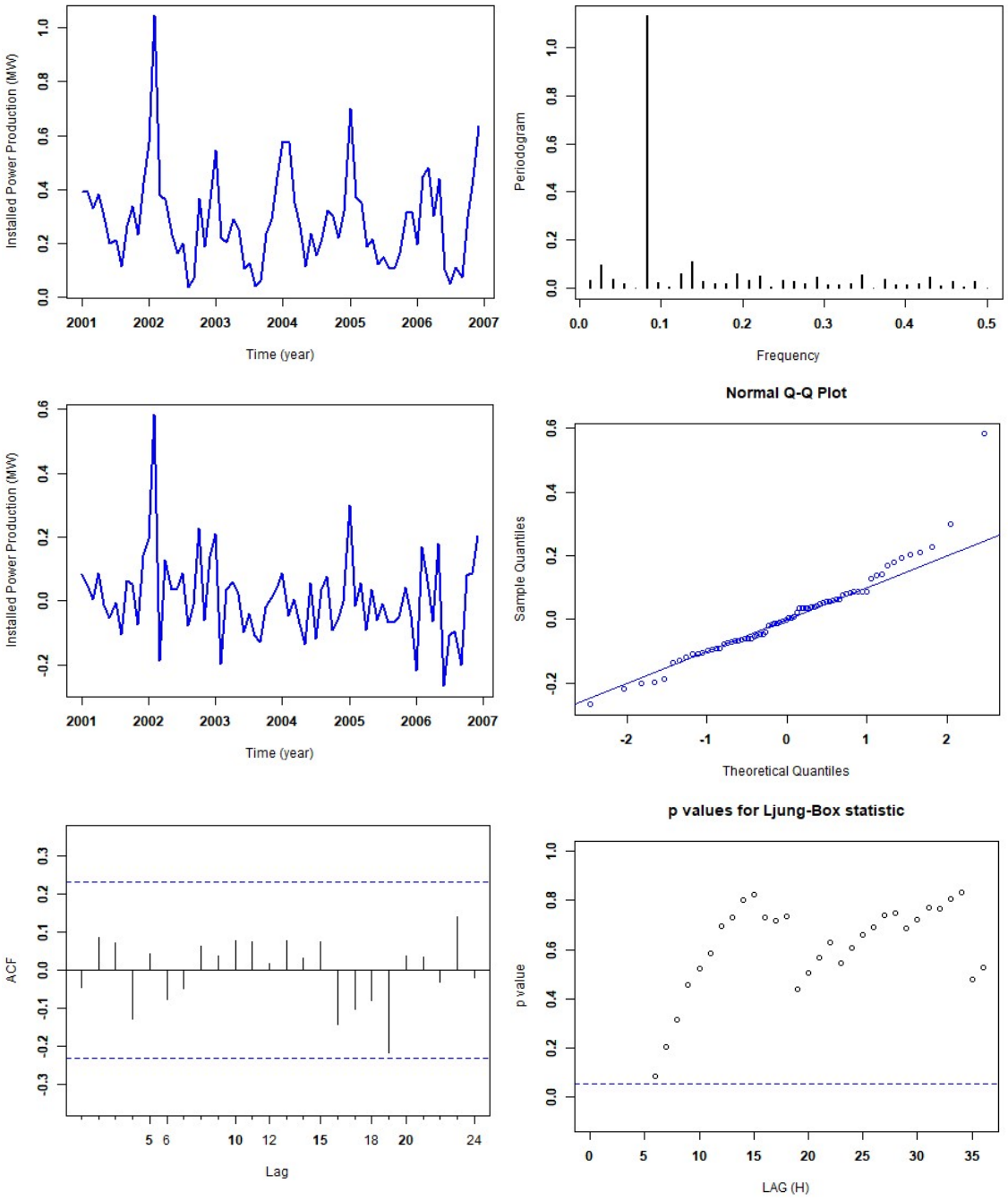


Figure B.42: Station 44: The fitted SARIMA model for installed power production is a SARIMA (2,0,1)(0,0,2)(12).

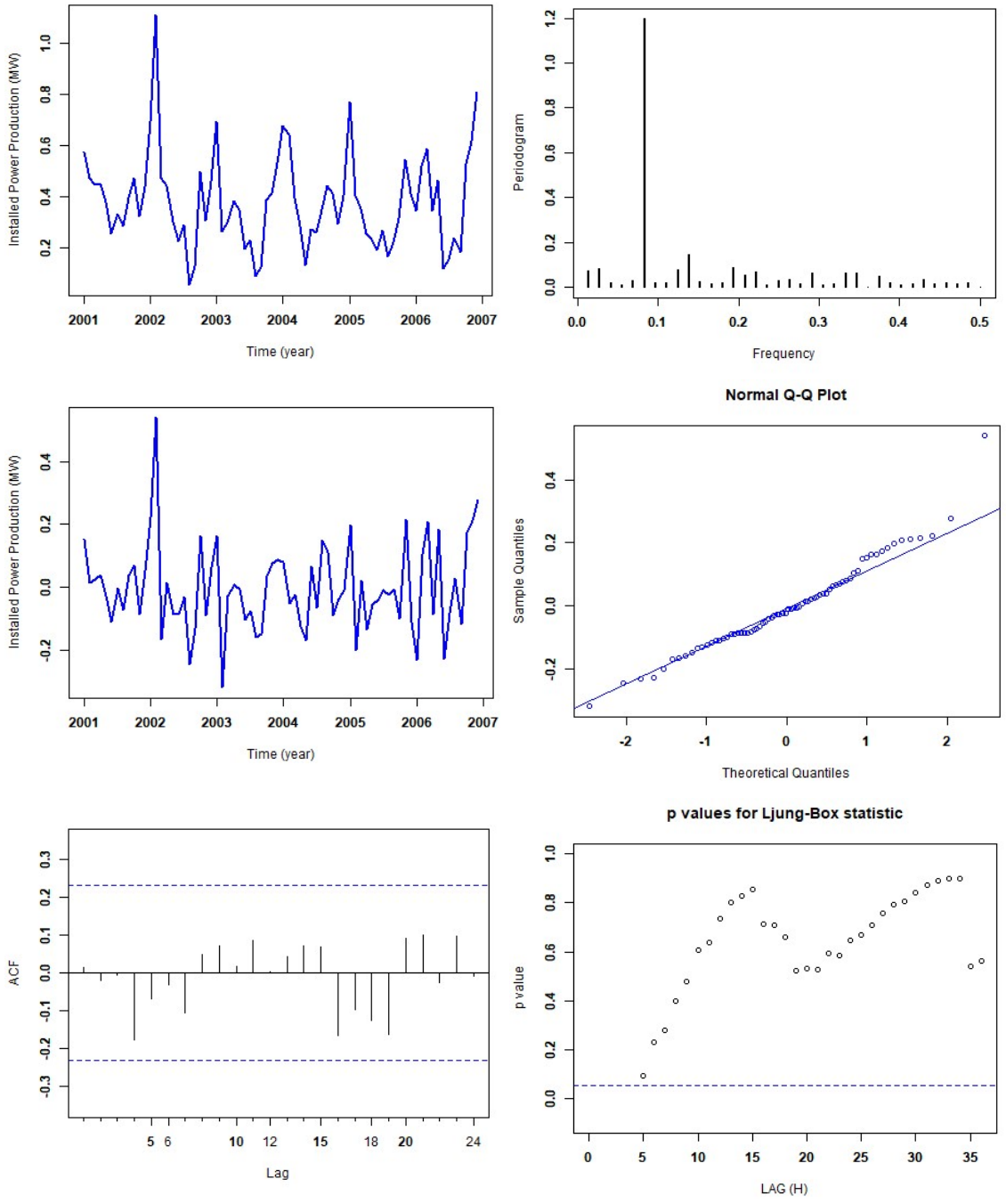


Figure B.43: Station 45: The fitted SARIMA model for installed power production is a SARIMA (1,0,0)(1,0,2)(12).

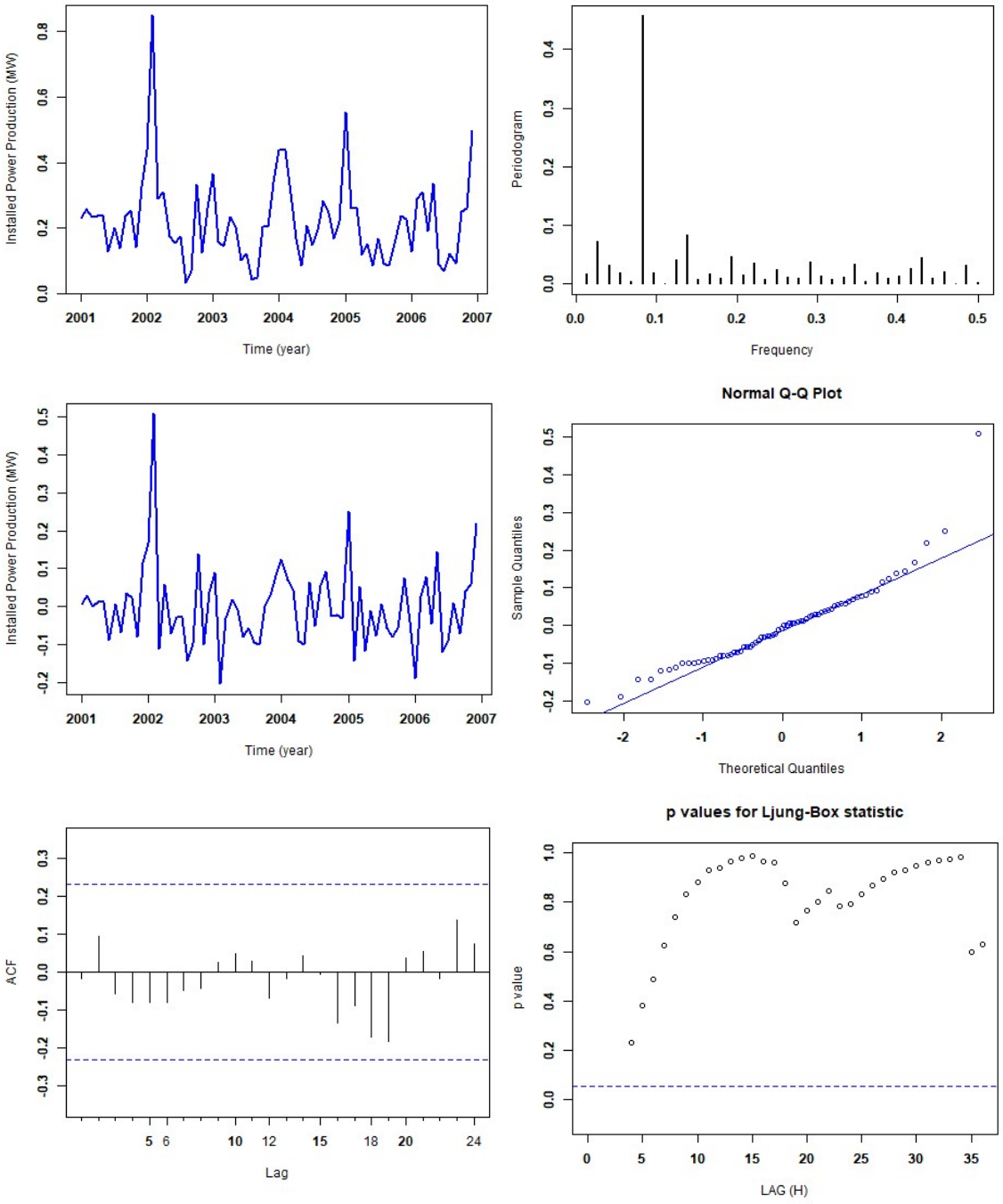


Figure B.44: Station 46: The fitted SARIMA model for installed power production is a SARIMA (1,0,0)(1,0,1)(12).

Appendix C

Figures for Times Series Forecasting

In Appendix C the predictions of monthly average wind power production for 2007 based on the data for the period 2001–2006. The blue line represents the original data, and the red line represents the SARIMA predictions. The green lines represent the interval of two standard deviations around the prediction (95.45% confidence interval), while the black lines represent the 68.27% confidence interval (based on one standard deviation and the normal probability assumption).

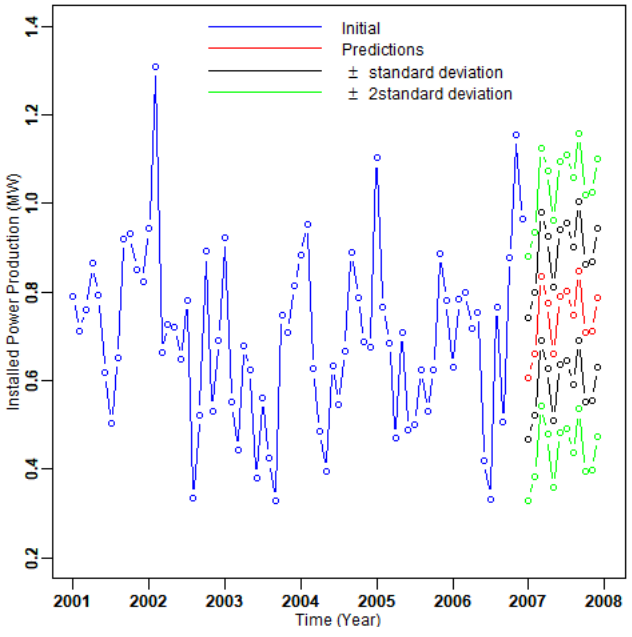


Figure C.1: Predictions for Station 2.

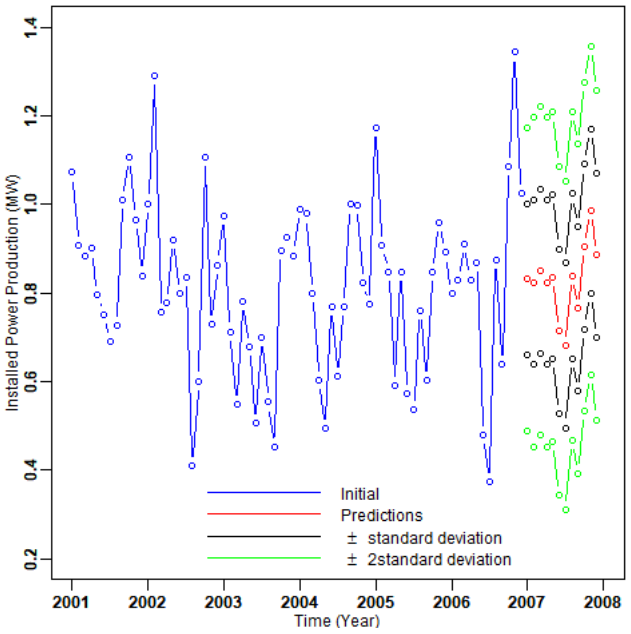


Figure C.2: Predictions for Station 3.

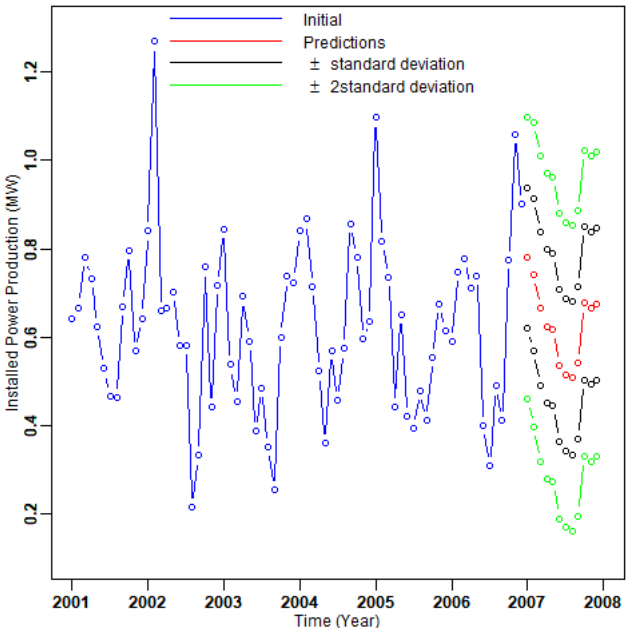


Figure C.3: Predictions for Station 4.

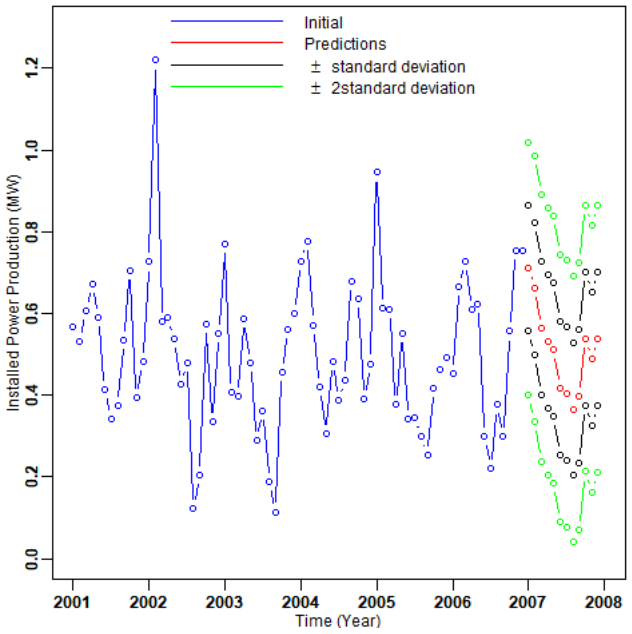


Figure C.4: Predictions for Station 5.

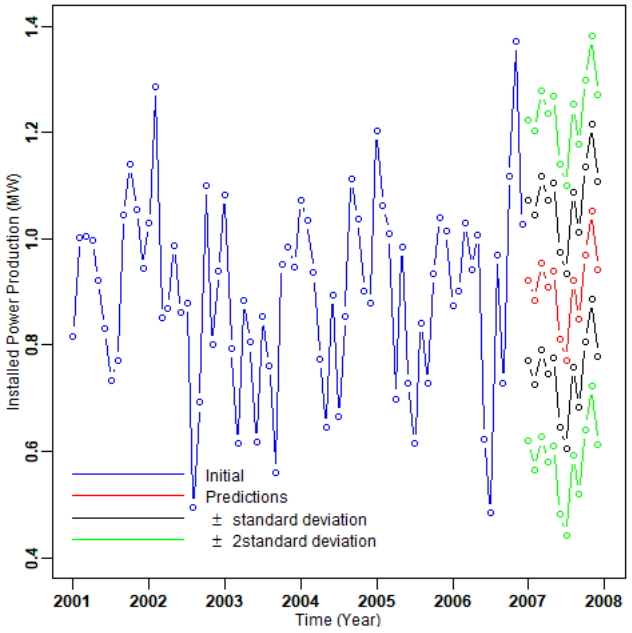


Figure C.5: Predictions for Station 6.

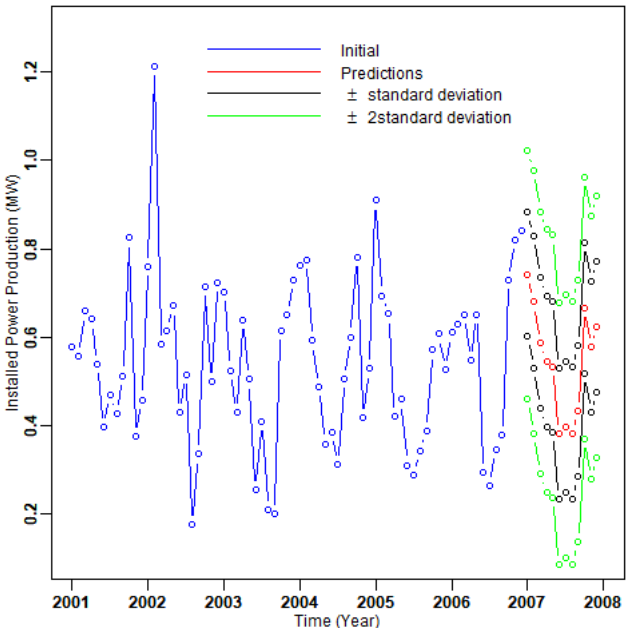


Figure C.6: Predictions for Station 7.

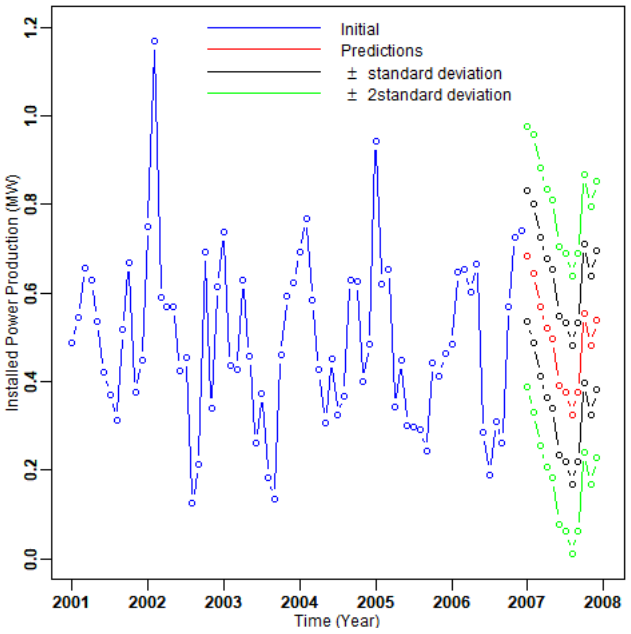


Figure C.7: Predictions for Station 8.

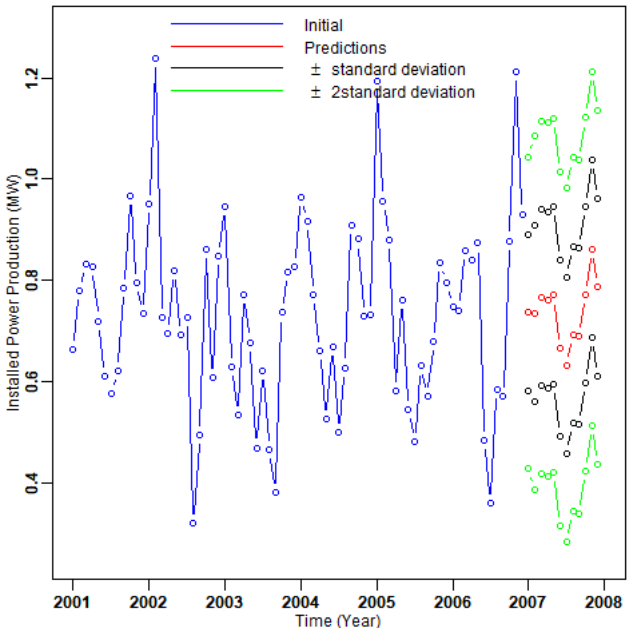


Figure C.8: Predictions for Station 9.

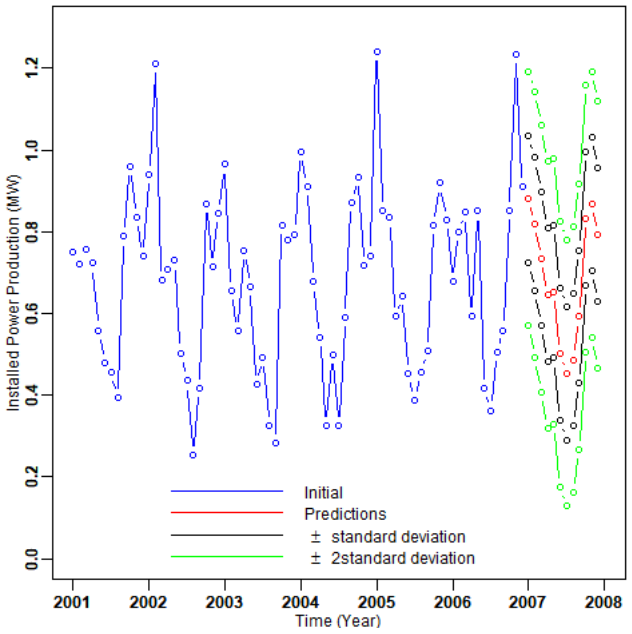


Figure C.9: Predictions for Station 10.

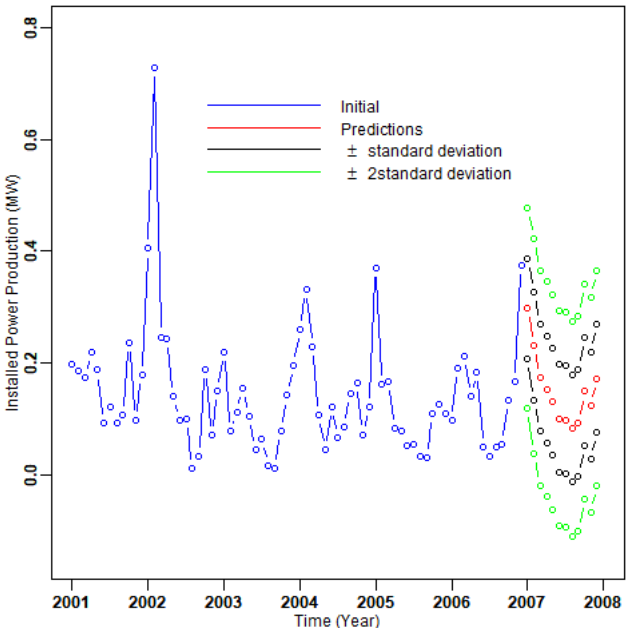


Figure C.10: Predictions for Station 11.

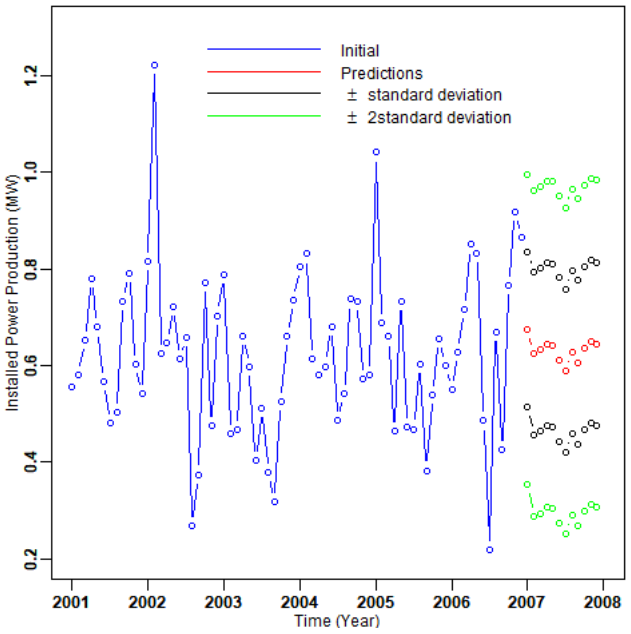


Figure C.11: Predictions for Station 12.

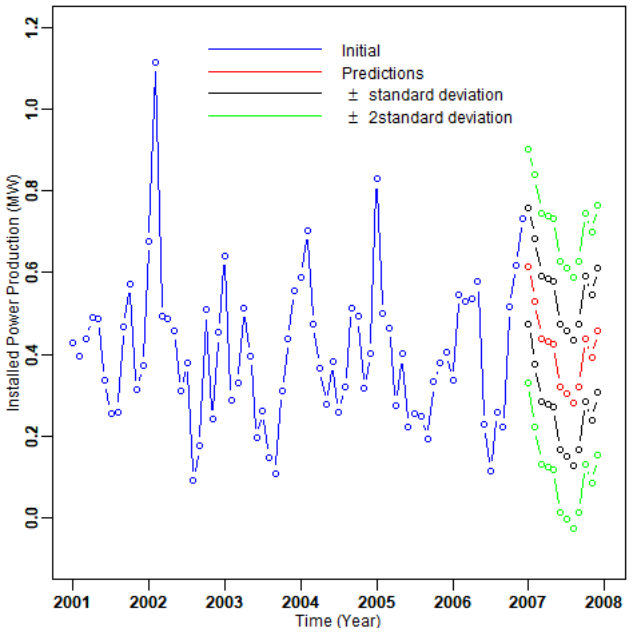


Figure C.12: Predictions for Station 13.

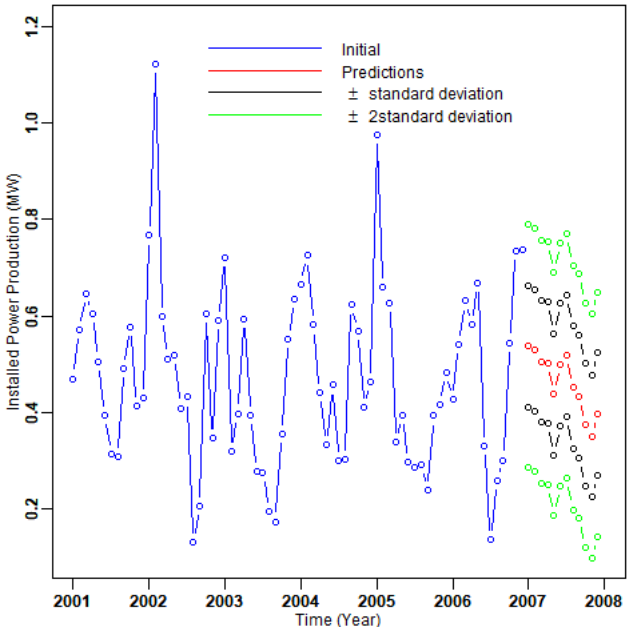


Figure C.13: Predictions for Station 14.

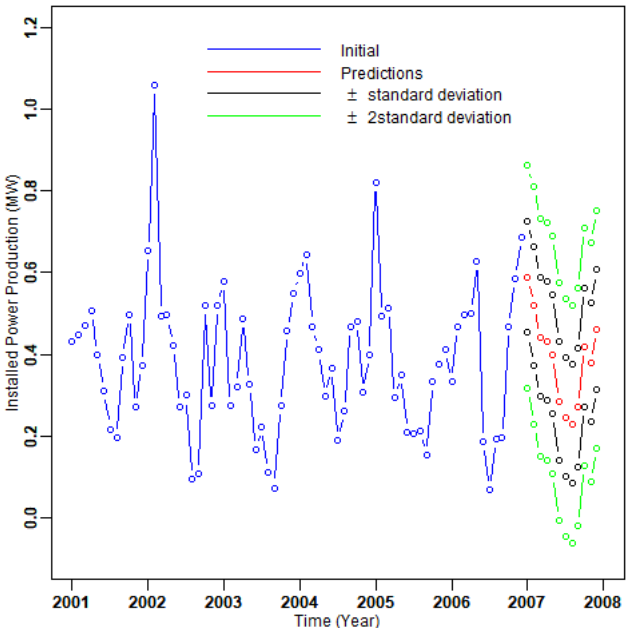


Figure C.14: Predictions for Station 15.

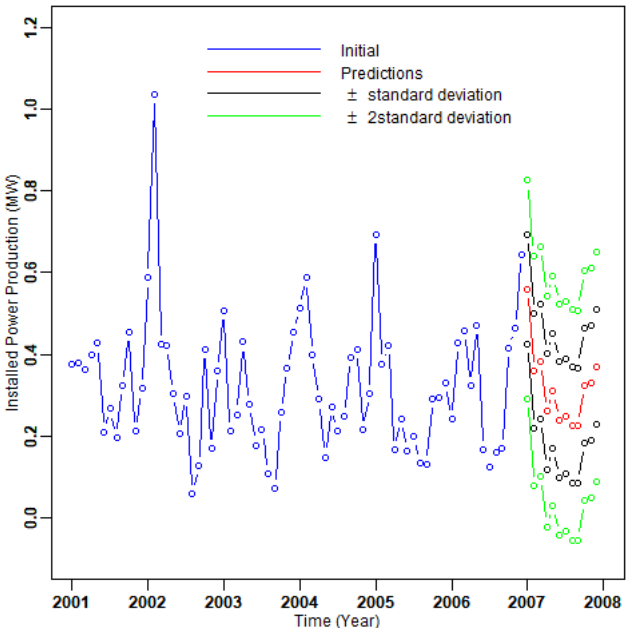


Figure C.15: Predictions for Station 16.

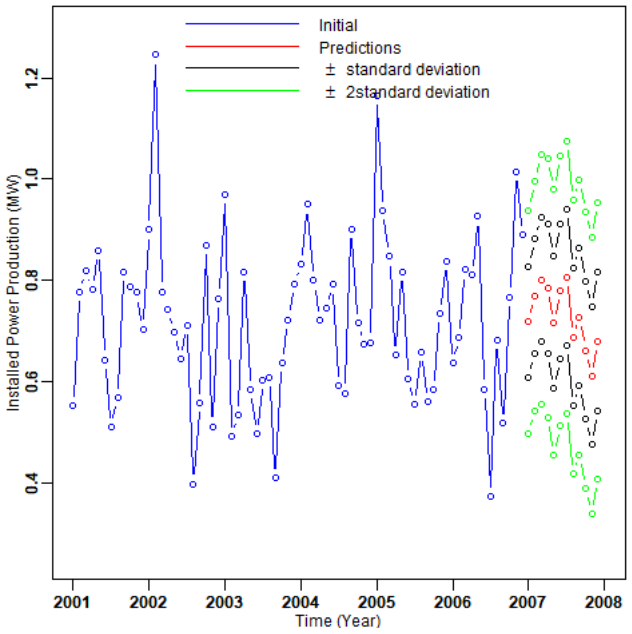


Figure C.16: Predictions for Station 17.

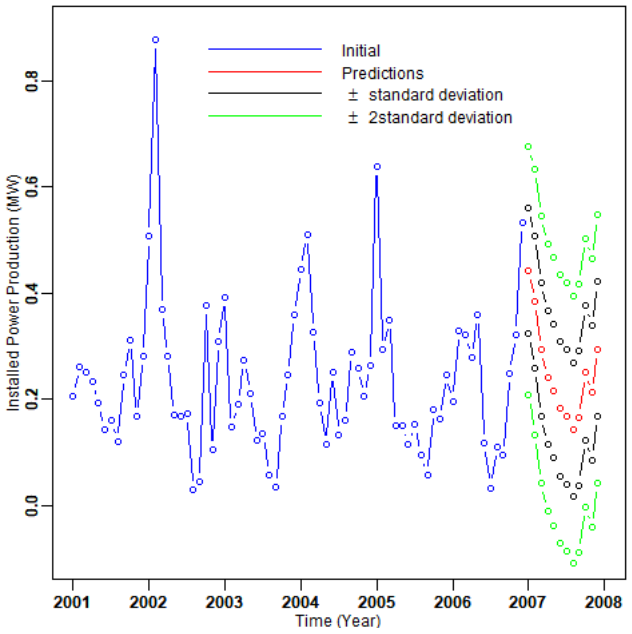


Figure C.17: Predictions for Station 18.

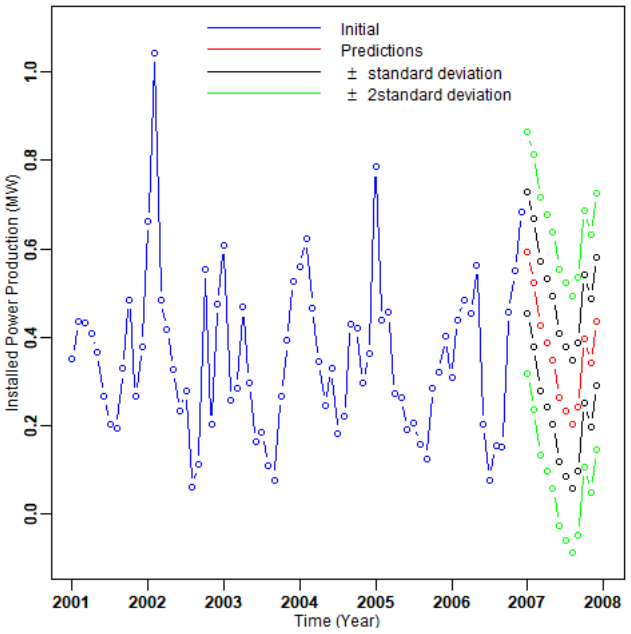


Figure C.18: Predictions for Station 19.

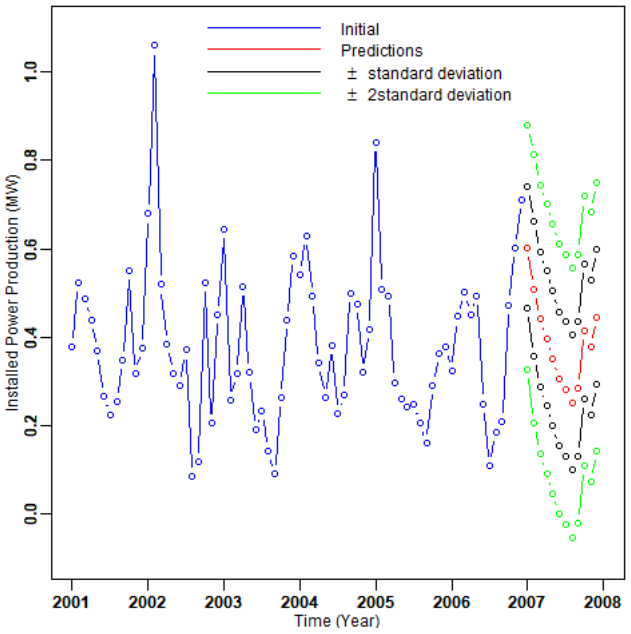


Figure C.19: Predictions for Station 20.

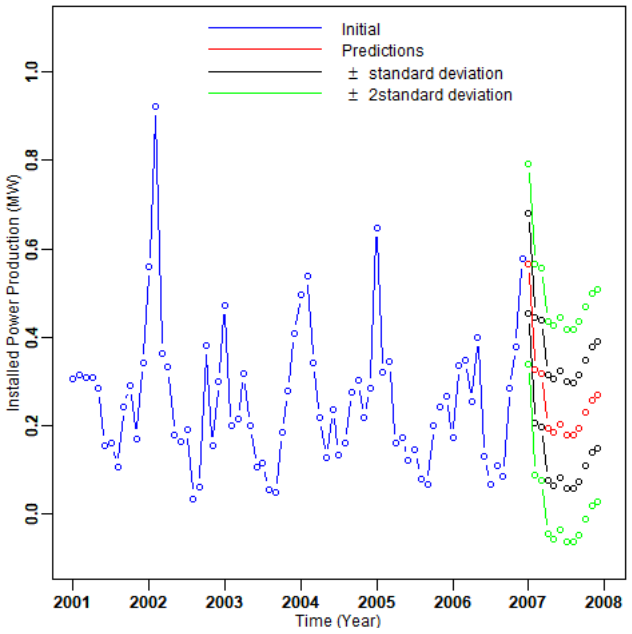


Figure C.20: Predictions for Station 21.

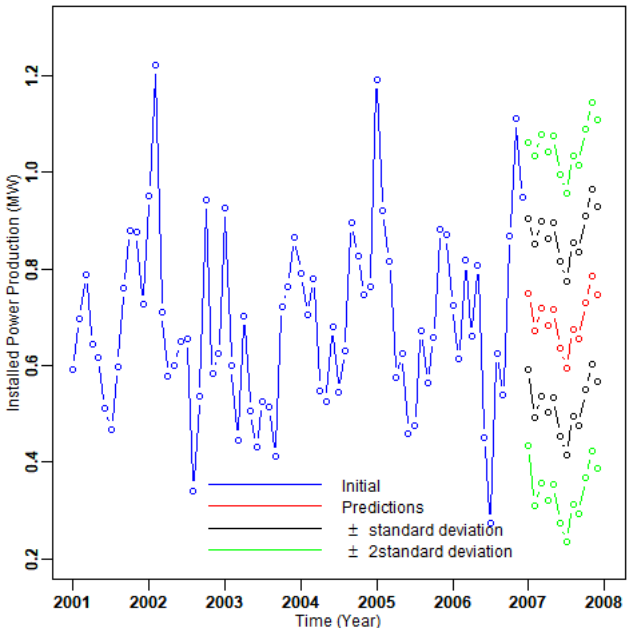


Figure C.21: Predictions for Station 22.

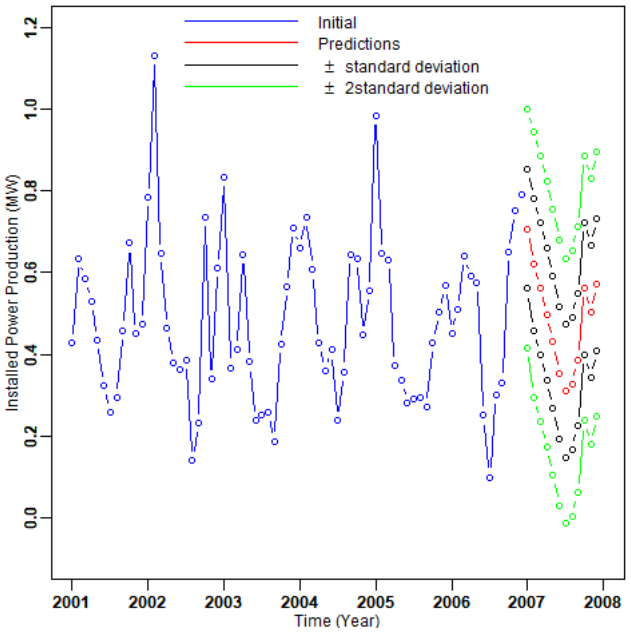


Figure C.22: Predictions for Station 23.

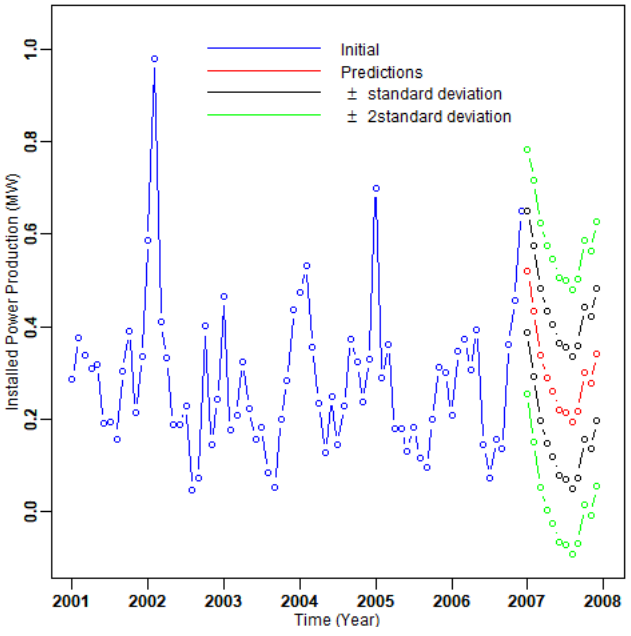


Figure C.23: Predictions for Station 24.

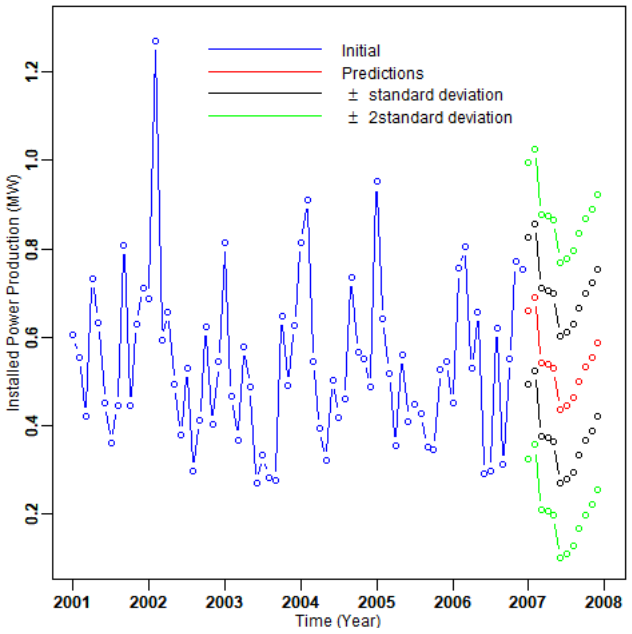


Figure C.24: Predictions for Station 25.

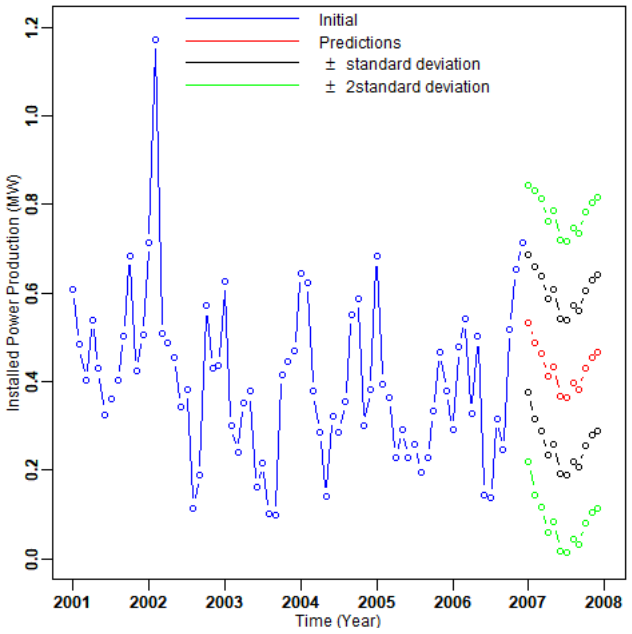


Figure C.25: Predictions for Station 26.

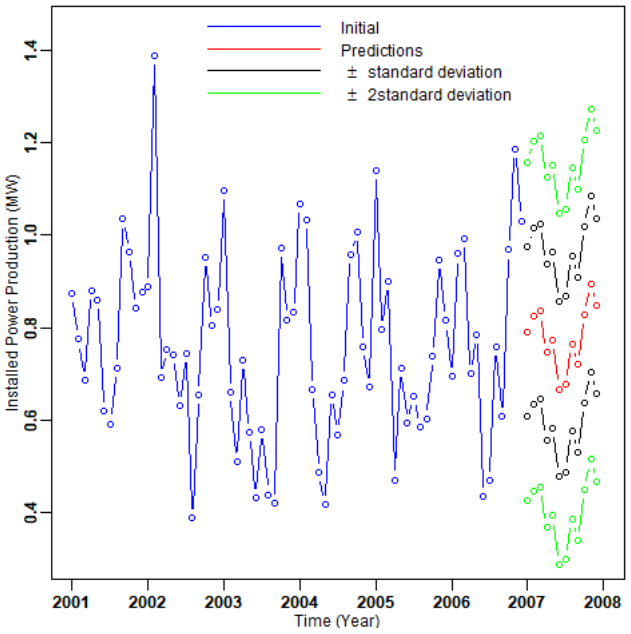


Figure C.26: Predictions for Station 27.

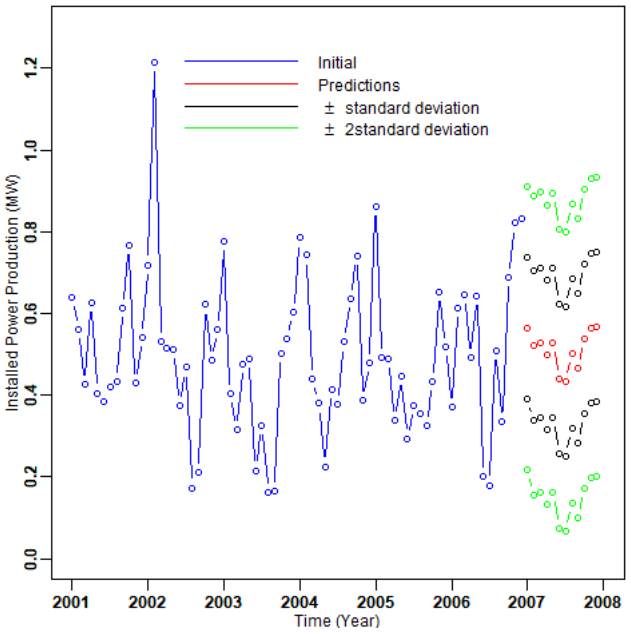


Figure C.27: Predictions for Station 28.

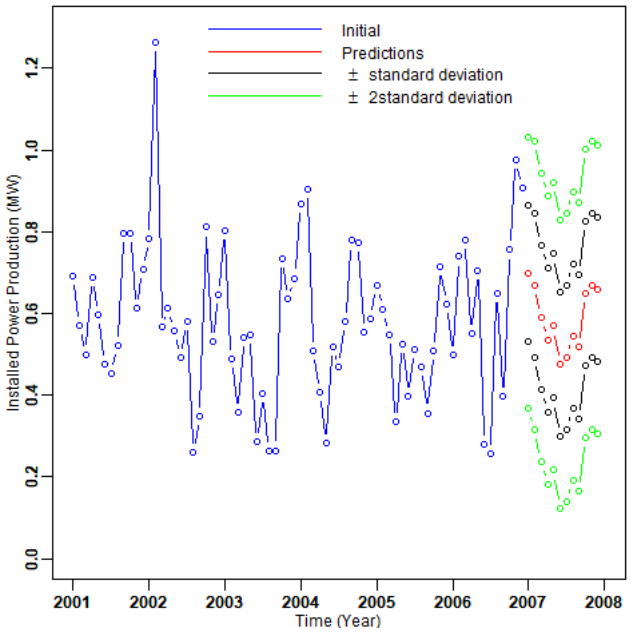


Figure C.28: Predictions for Station 29.

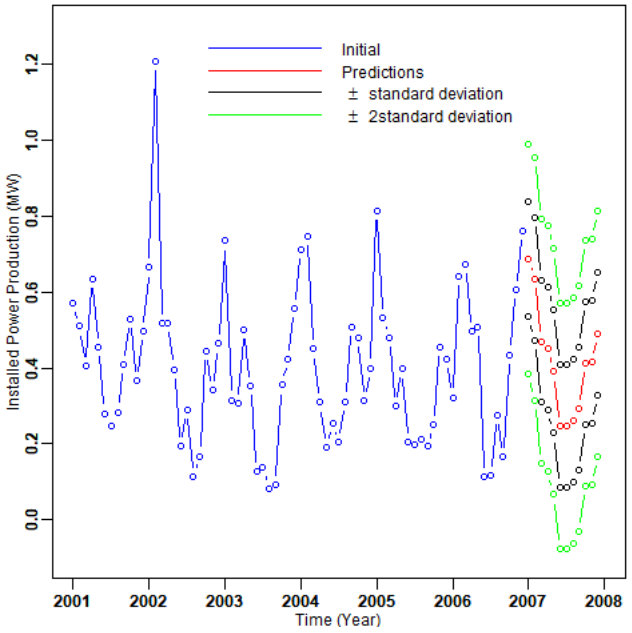


Figure C.29: Predictions for Station 30.

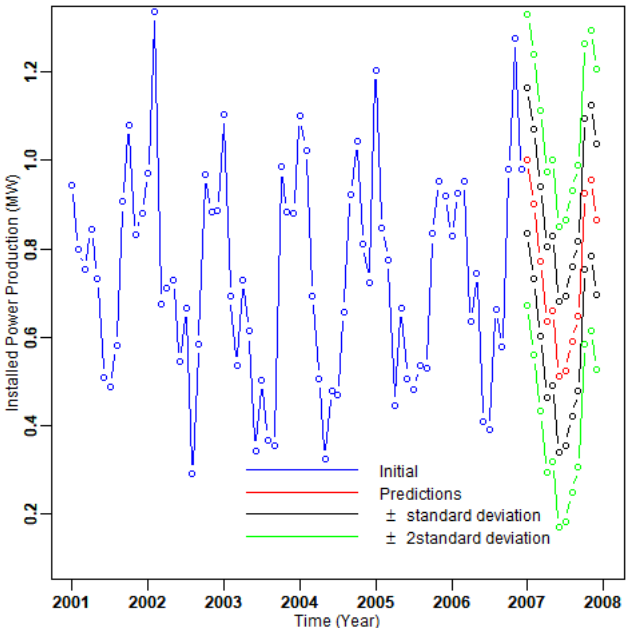


Figure C.30: Predictions for Station 32.

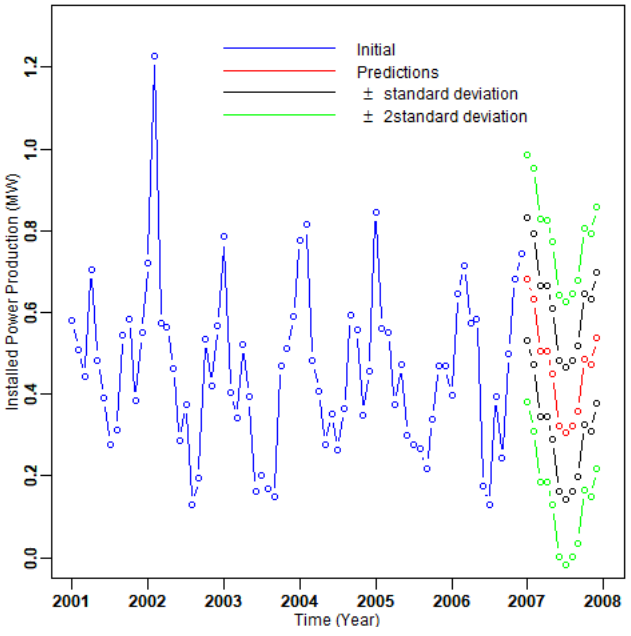


Figure C.31: Predictions for Station 33.

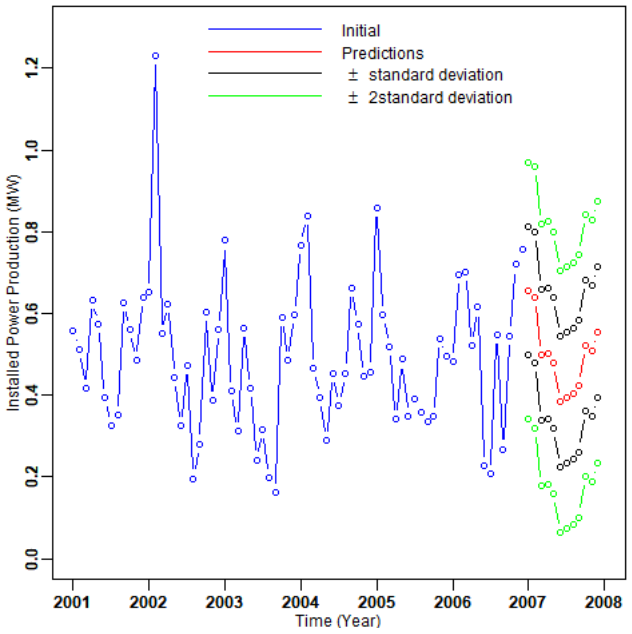


Figure C.32: Predictions for Station 34.

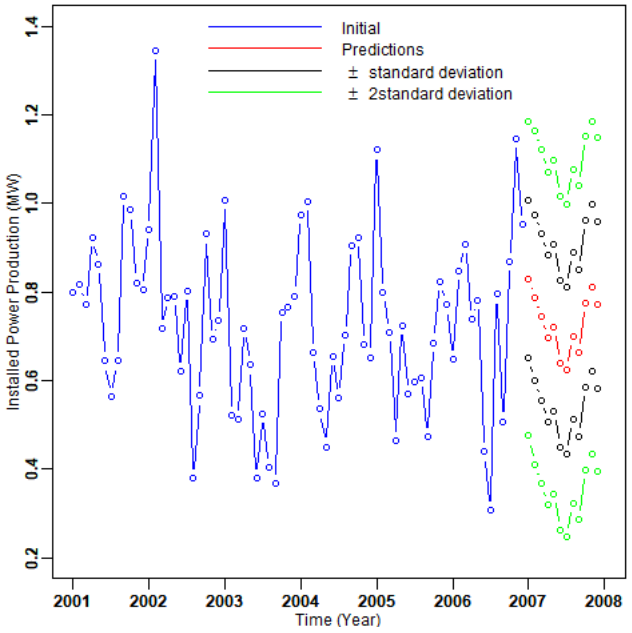


Figure C.33: Predictions for Station 35.

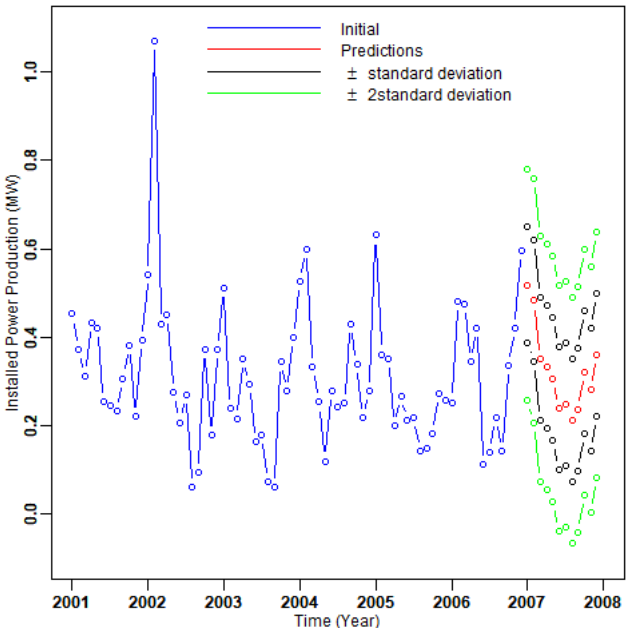


Figure C.34: Predictions for Station 36.

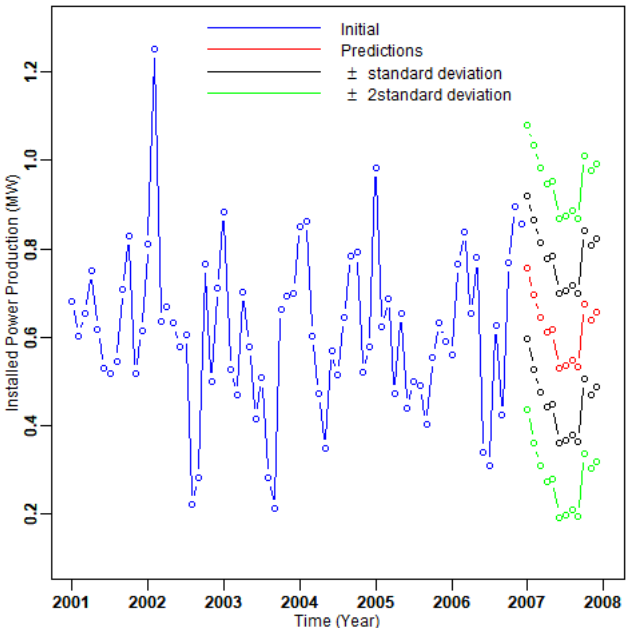


Figure C.35: Predictions for Station 37.

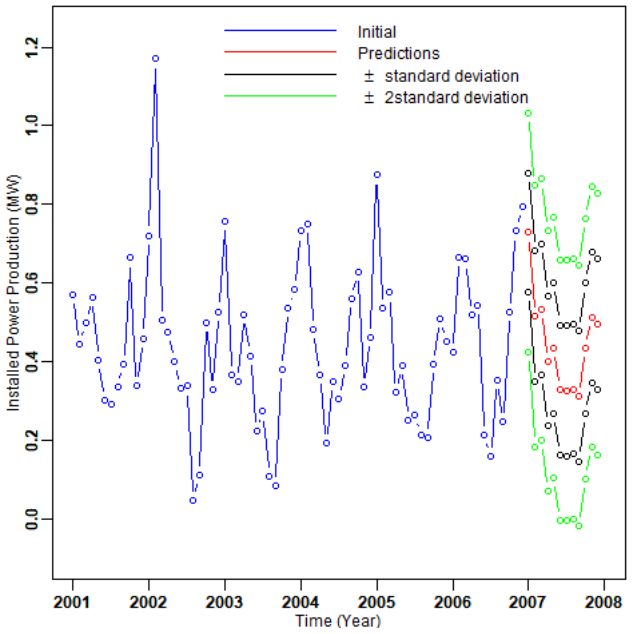


Figure C.36: Predictions for Station 38.

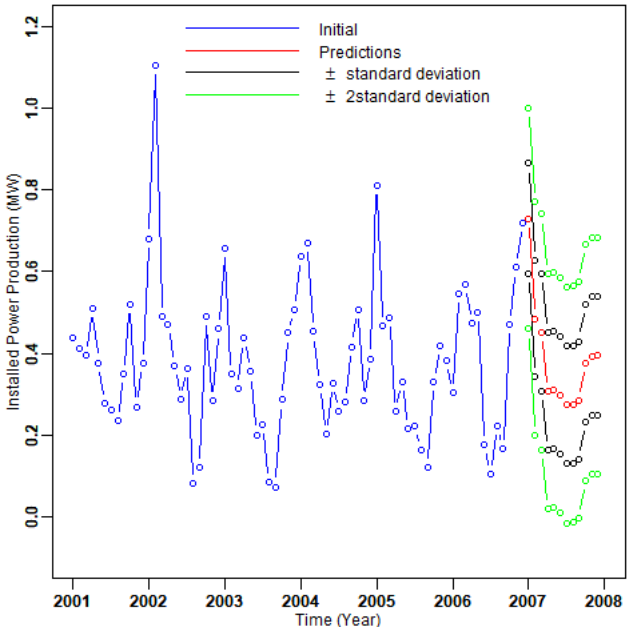


Figure C.37: Predictions for Station 39.

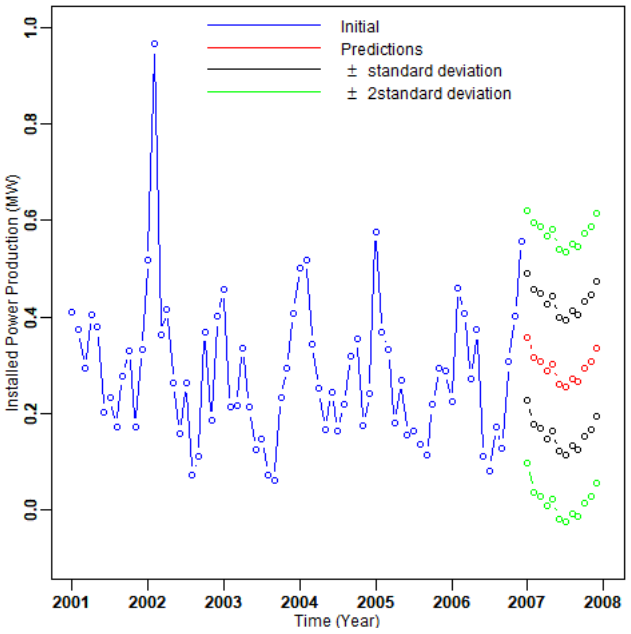


Figure C.38: Predictions for Station 40.

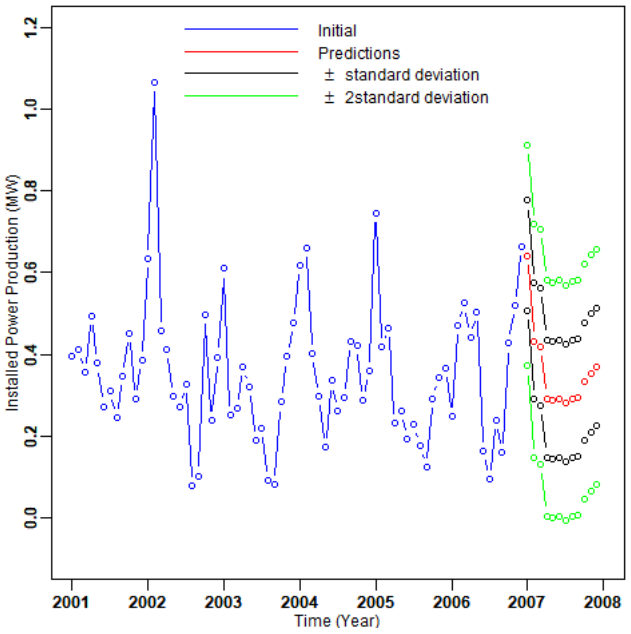


Figure C.39: Predictions for Station 41.

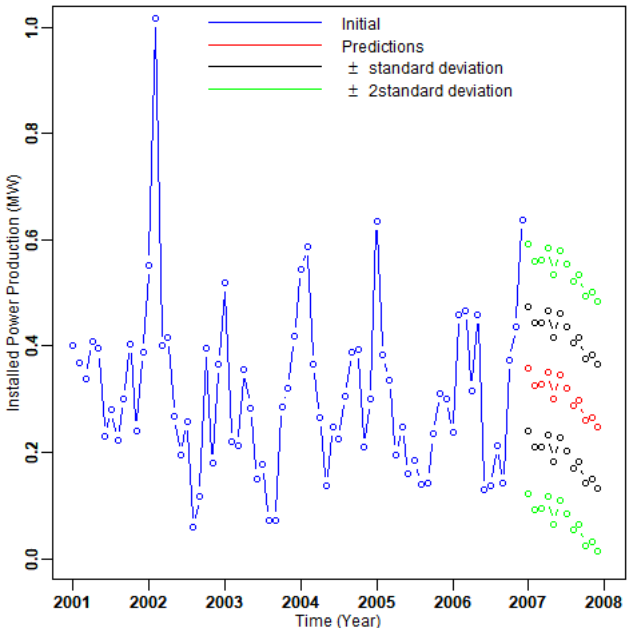


Figure C.40: Predictions for Station 42.

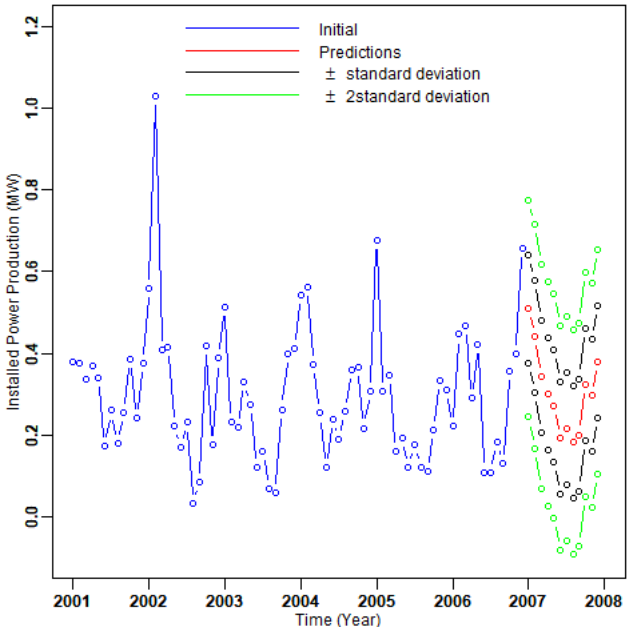


Figure C.41: Predictions for Station 43.

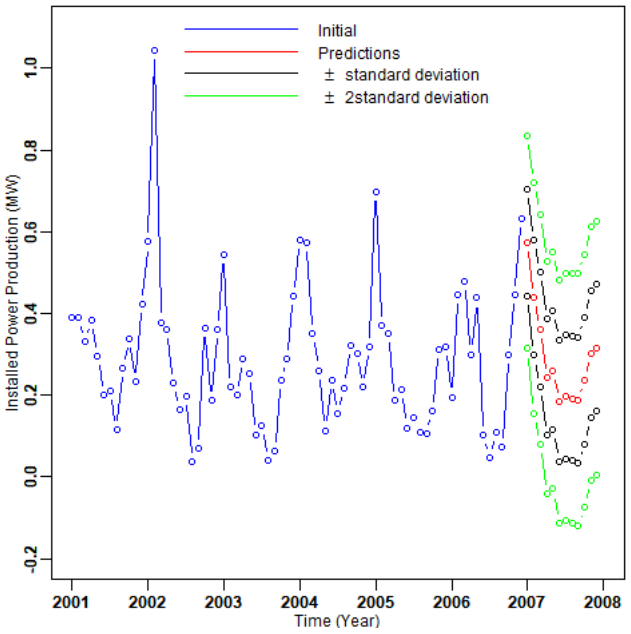


Figure C.42: Predictions for Station 44.

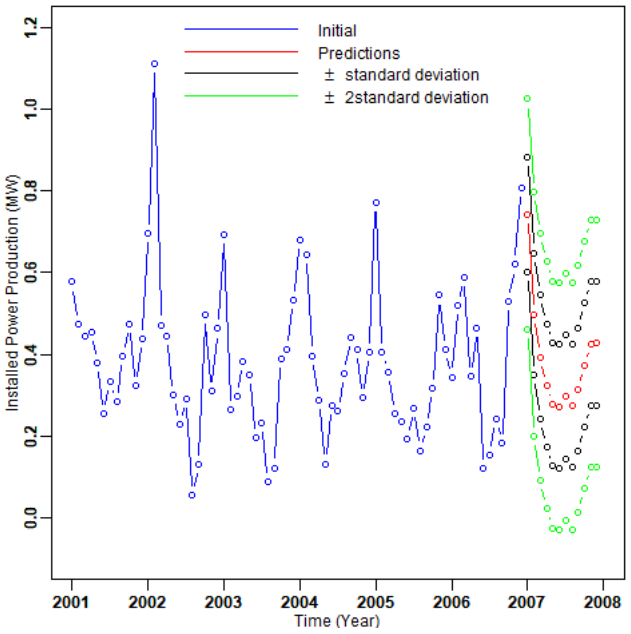


Figure C.43: Predictions for Station 45.

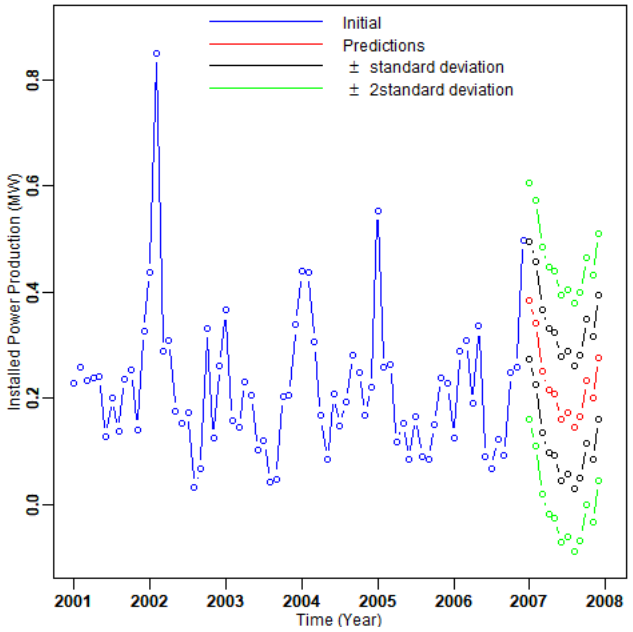


Figure C.44: Predictions for Station 46.

Appendix D

Tables with Cross-Validation Measures and Parameters of SARIMA models and Distribution fit

Table D.1: Table for the estimated parameters for Normal Distribution.

Station	Mean	Standard deviation
2	0.70	0.19
3	0.81	0.20
4	0.62	0.19
6	0.89	0.18
9	0.73	0.18
12	0.62	0.17
17	0.72	0.16
29	0.58	0.19
34	0.49	0.18
35	0.73	0.19
37	0.61	0.18

Table D.2: Table for the estimated parameters for Log-Normal Distribution

Station	Log(Mean)	Log(Standard deviation)
16	− 1.26	0.52
22	− 0.41	0.28
25	−0.67	0.33
46	−1.67	0.60

Table D.4: SARIMA model parameters for the residuals of installed power production. The SE is the standard error of the estimates and the p-value is used in the context of null hypothesis testing of zero correlation in order to quantify the idea of statistical significance of evidence.

Station	Model	Estimate	SE	p-values
2: (1,0,3)(1,0,1)(12)	AR1	0.86	0.09	0.00
	MA1	−0.86	0.12	0.00
	MA2	−0.33	0.14	0.0
	MA3	0.52	0.12	0.00
	SAR1	0.62	0.15	0.00
	SMA1	−1.00	0.25	0.00
3: (0,0,1)(1,0,0)(12)	MA1	0.42	0.14	0.004
	SAR1	0.31	0.12	0.01
4: (0,0,1)(1,0,1)(12)	MA1	0.41	0.12	0.00
	SAR1	0.98	0.16	0.00
	SAR2	−0.91	0.15	0.06
5: (1,0,0)(1,0,1)(12)	AR1	0.32	0.12	0.001
	SAR1	0.10	0.02	0.00
	SMA1	−0.96	0.18	0.00
6: (1,0,1)(1,0,0)(12)	AR1	−0.62	0.14	0.00
	MA1	0.95	0.08	0.00
	SAR1	0.32	0.13	0.02
7: (1,0,2)(1,0,1)(12)	AR1	0.87	0.07	0.00
	MA1	−0.61	0.12	0.00
	MA2	−0.39	0.12	0.00
	SAR1	$-0.98	0.09	0.00
	SMA1	−0.85	0.4	0.04
8 : (0,0,1)(1,0,1)(12)	MA1	0.36	0.12	0.004

Station	Model	Estimate	SE	p-values
	SAR1	1.00	0.02	0.00
	SMA1	−0.96	0.2	0.00
9: (0,0,1)(1,0,0)(12)	MA1	0.54	0.11	0.00
	SAR1	0.27	0.12	0.03
	MA1	0.32	0.14	0.02
10: (0,0,1)(1,0,1)(12)	SAR1	1.00	0.01	0.00
	SMA1	−0.78	0.35	0.03
	AR1	0.40	0.099	$2\ 10^{-4}$
11: (0,0,1)(1,0,1)(12)	SAR1	1.00	0.002	0.00
	SMA1	−0.96	0.21	0.00
12: (0,0,1)(1,0,0)(12)	MA1	0.34	0.12	0.007
	SAR1	0.09	0.13	0.5
	MA1	0.40	.011	$8\ 10^{-4}$
13:(0,0,1)(1,0,1)(12)	SAR1	1.00	0.01	0.00
	SMA1	−0.97	0.14	0.00
	AR1	0.93	0.06	0.00
14:(1,0,1)(1,0,0,)(12)	MA1	−1.00	0.04	0.00
	SAR1	−0.32	0.12	0.009
	MA1	0.37	0.11	0.02
15: (0,0,1)(1,0,1)(12)	SAR1	1.00	0.01	0.00
	SMA1	−0.97	0.15	0.00
	MA1	0.50	0.12	0.0001
16: (0,0,1)(1,0,1)(12)	SAR1	0.98	0.09	0.00
	SMA1	−0.87	0.4	0.04
	AR1	0.78	0.10	0.00
	MA1	− 1.08	0.12	0.00
17: (1,0,3)(0,0,1)(12)	MA2	−0.23	0.15	0.10
	MA3	0.69	0.11	0.00
	SAR1	−0.59	0.19	0.003
	AR1	0.37	0.12	0.03
18:(1,0,0)(1,0,1)(12)	SAR1	0.99	0.03	0.00
	SMA1	−0.96	0.21	0.00
	AR1	0.12	0.34	0.01
19: (1,0,0)(1,0,1)(12)	SAR1	0.01	0.99	0.00

Station	Model	Estimate	SE	p-values
	SMA1	0.17	−0.96	0.00
	MA1	0.47	0.11	10^{-4}
20: (0,0,1)(10,1)(12)	SAR1	0.1	0.02	0.00
	SMA1	−0.96	0.19	0.00
21: (1,0,0)(1,0,2)(12)	AR1	0.33	0.13	0.01
	SAR1	−0.74	0.23	0.002
	SMA1	−0.73	0.29	0.01
	SMA2	0.44	0.22	0.05
22: (0,0,1)(1,0,0)(12)	MA1	0.57	0.11	0.00
	SAR1	0.23	0.12	0.06
	MA1	0.50	0.11	0.0001
23: (0,0,1)(1,0,1)(12)	SAR1	0.99	0.05	0.00
	SAR2	−0.93	0.04	0.01
	AR1	0.39	0.12	0.002
24: (1,0,0)(1,0,1)(12)	SAR1	0.99	0.08	0.000
	SMA1	−0.94	0.3	0.003
25: (0,0,0)(1,0,1)(12)	SAR1	1.00	0.05	0.00
	SMA1	−0.95	0.03	0.03
26: (1,0,0)(1,0,0)(12)	AR1	0.45	0.11	0.0001
	SAR1	0.17	0.12	0.1
27: (0,0,1)(1,0,0)(12)	AR1	0.27	0.12	0.02
	SAR1	0.30	0.12	0.01
28: (0,0,1)(1,0,0)(12)	MA1	0.35	0.11	0.002
	SAR1	0.20	0.13	0.1
	MA1	0.36	0.12	0.005
29: (0,0,1)(1,0,1)(12)	SAR1	0.87	0.24	0.0004
	SMA1	−0.70	0.3	0.045
	AR1	0.35	0.12	0.005
30: (1,0,0)(1,0,1)(12)	SAR1	0.1	0.01	0.00
	SMA1	−0.94	0.14	0.005
	AR1	0.25	0.13	0.05
32: (1,0,0)(1,0,1)(12)	SAR1	0.95	0.08	0.00
	SMA1	−0.71	0.2	0.0009
	MA1	0.36	0.11	0.003
33: (0,0,1)(1,0,1)(12)				

Station	Model	Estimate	SE	p-values
	SAR1	0.99	0.09	0.00
	SMA1	−0.90	0.14	0.05
	MA1	0.21	0.12	0.1
34: (0,0,1)(1,0,1)(12)	SAR1	0.99	0.13	0.00
	SMA1	−0.91	0.48	0.00
	MA1	0.35	0.12	0.008
35: (0,0,1)(1,0,1)(12)	SAR1	0.80	0.3	0.01
	SMA1	−0.63	0.3	0.01
	AR1	0.34	0.12	0.005
36: (1,0,0)(1,0,1)(12)	SAR1	1.00	0.01	0.00
	SMA1	−0.96	0.14	0.00
	MA1	0.30	0.12	0.02
37: (0,0,1)(1,0,1)(12)	SAR1	0.93	0.25	0.001
	SMA1	−0.80	0.24	0.07
	MA1	0.43	1.00	0.001
38: (0,0,1)(2,0,0)(12)	SAR1	0.13	0.099	0.02
	SAR2	0.44	0.12	0.001
	AR1	0.36	0.12	0.004
	SAR1	0.77	0.18	0.000
39: (1,0,0)(1,0,2)(12)	SMA1	−0.76	0.24	0.002
	SMA2	0.40	0.21	0.06
	MA1	0.38	0.11	0.0007
40: (0,0,1)(0,0,2)(12)	SMA1	0.21	0.12	0.1
	SMA2	0.33	0.15	0.03
41: (1,0,0)(1,0,2)(12)	AR1	0.34	0.12	0.006
	SAR1	0.68	0.29	0.02
	SMA1	−0.65	0.03	0.06
	SMA2	0.36	0.19	0.06
42: (0,0,0)(1,0,0)(12)	SAR1	−0.21	0.12	0.08
43: (0,0,1)(1,0,1)(12)	MA1	0.26	0.11	0.02
	SAR1	1.00	0.014	0.00
	SMA1	−0.96	0.17	0.00
	AR1	1.40	0.10	0.00
	AR2	−0.52	0.10	0.00
44: (2,0,1)(0,0,2)(12)				

Station	Model	Estimate	SE	p-values
	MA1	−1.00	0.04	0.00
	SMA1	0.12	0.13	0.04
	SMA2	0.37	0.17	0.03
45: (1,0,0)(1,0,2)(12)	AR1	0.37	0.12	0.004
	SAR1	0.84	0.20	0.00
	SMA1	−0.81	0.27	0.004
	SMA2	0.29	0.20	0.015
46: (1,0,0)(1,0,1)(12)	AR1	0.33	0.12	0.007
	SAR1	0.99	0.05	0.000
	SMA1	−0.93	0.34	0.008

Table D.5: Cross validation performance measures calculated through the leave-one-out cross validation for the monthly average installed power production of the station 31. ME: mean error; MAE: mean absolute error; RMSE: root mean squared error.

Station	ME (MW)	MAE (MW)	RMSE (MW)
2	−0.03	0.14	0.17
3	0.004	0.16	0.19
4	0.02	0.14	0.26
5	0.009	0.12	0.14
6	0.016	0.15	0.19
7	−0,02	0.11	0.24
8	−0.002	0.12	0.14
9	0.01	0.13	0.17
10	0.02	0.14	0.18
11	−0.0097	0.06	0.07
12	0.01	0.14	0.17
13	0.007	0.11	0.14
14	−0.003	0.09	0.12
15	0.007	0.11	0.13
16	−0.002	0.10	0.12
17	0.01	0.08	0.12
18	0.004	0.09	0.11
19	0.006	0.1	0.12

Station	ME (MW)	MAE (MW)	RMSE (MW)
20	0.01	0.11	0.13
21	−0.004	0.096	0.12
22	0.02	0.13	0.17
23	0.02	0.12	0.15
24	0.004	0.10	0.12
25	0.006	0.13	0.16
26	−0.009	0.13	0.14
27	0.01	0.16	0.18
28	0.01	0.15	0.17
29	0.006	0.14	0.16
30	0.004	0.12	0.14
32	0.007	0.14	0.17
33	0.002	0.13	0.14
34	0.01	0.13	0.15
35	−0.002	0.14	0.16
36	−0.006	0.10	0.12
37	0.001	0.14	0.16
38	0.001	0.13	0.15
39	0.003	0.11	0.14
40	0.001	0.10	0.11
41	0.006	0.11	0.14
42	−0.006	0.09	0.11
43	−0.001	0.11	0.13
44	−0.007	0.11	0.13
45	0.0002	0.13	0.15
46	−0.006	0.08	0.0097

Table D.3: Table of the estimated parameter for the Weibull distribution in each station.

Station	Shape	Scale
5	2.82	0.56
7	3.09	0.60
8	2.79	0.55
10	3.38	0.76
11	1.44	0.15
13	2.43	0.46
14	2.71	0.53
15	2.27	0.43
18	1.74	0.27
19	1.74	0.27
20	2.29	0.43
21	1.77	0.29
23	2.06	0.54
24	1.89	0.32
26	2.34	0.46
27	3.97	0.84
28	2.74	0.56
30	3.44	0.75
31	2.09	0.45
32	3.46	0.82
33	2.47	0.52
36	2.08	0.35
38	2.34	0.50
39	2.19	0.43
40	2.05	0.32
42	2.10	0.35
43	1.91	0.33
44	1.74	0.32
45	−1.29	0.52

Table D.6: Seasonal model parameters of average monthly wind power. The Standard Error (SE) for a given variable is given by the Residual Standard Error divided by the square root of the sum of squares for the particular variable. The p-value is used to test the null hypothesis that the respective coefficient is zero.

Station	**coefficient**	**estimate**	**SE**	**p-value**
2	μ	0.70	0.02	2×10^{-16}
	sin	-0.04	0.03	0.1
	cos	0.15	0.03	7.55×10^{-7}
14	μ	0.47	0.02	2×10^{-16}
	sin	0.051	0.02	0.03
	cos	0.17	0.02	1.87×10^{-7}
17	μ	0.72	0.02	2×10^{-16}
	sin	0.04	0.02	0.0
	cos	0.12	0.02	5.53×10^{-6}
42	μ	0.31	0.01	2×10^{-16}
	sin	0.03	0.02	0.12
	cos	0.13	0.02	4.73×10^{-9}

References

[1] *Wind power program.* https://www.wind-power-program.com/turbine_characteristics.htm.

[2] Agou, V.D.: *Geostatistical Analysis of Precipitation on the island of Crete.* Master's thesis, Technical University of Crete, 2016.

[3] Akaike, A.: *A new look at the statistical model identification.* IEEE Transactions on Automatic Control, (6):716–723, 1974.

[4] Alencar, David Barbosa de, Carolina De Mattos Affonso, Roberto Célio Limão de Oliveira, Jorge Laureano Moya Rodríguez, Jandecy Cabral Leite, and José Carlos Reston Filho: *Different models for forecasting wind power generation: Case study.* Energies, 10(12), 2017, ISSN 1996-1073. https://www.mdpi.com/1996-1073/10/12/1976.

[5] Alexiadis, M. C., P.S. Dokopoulos, and H.S. Sahsamanoglou: *Wind speed and power forecasting based on spatial correlation models.* IEEE Transactions on Energy Conversion, 14(3):836–842, September 1999.

[6] Anonymous: *Profec.* https://profec-ventus.com/, visited on May 2020.

[7] Anonymous: *Wind sim.* http://windsim.com/services/bankable-aep-assessment.aspx, visited on May 2020.

[8] Armstrong, M: *Common problems seen in variograms.* Mathematical Geology, 16(3):305–313, 1984.

[9] Association, European Biomass Industry: *Renewable energy.* https://www.eubia.org/cms/wiki-biomass/renewable-energy/.

[10] Bank, Asian Development: *Guidelines for wind resource assessment best practices for countries initiating wind development.* Technical report, 2014, ISBN 978-92-9254-562-8 (PRINT), 978-92-9254-563-5 (PDF).

[11] Bochner, S., M. Tenenbaum, and H. Pollard: *Lectures on Fourier Integrals.* Annals of mathematics studies. Princeton University Press, 1959, ISBN 9780691079943.

[12] Carrillo, C., Josè Cidrès, Eloy Diaz-Dorado, and Andrès Obando Montaño: *An approach to determine the weibull parameters for wind energy analysis: The case of galicia (spain).* Energies, 7:2676–2700, April 2014.

[13] Cavanaugh, Joseph. E.: *Unifying the derivations for the Akaike and corrected Akaike information criteria.* Statistics & Probability Letters, (2):201–208, 1997.

[14] Chatfield: *The Analysis of Time Series: Theory and Practice.* Springer, 1975, ISBN 978-0-412-14180-5.

[15] Chilès, J. P. and P. Delfiner: *Geostatistics: Modeling Spatial Uncertainty.* Wiley series in probability and statistics. Wiley, New York, 2h edition, 2012, ISBN 9780470183151.

[16] Christakos, G.: *Random Field Models in Earth Sciences.* Academic Press, San Diego, 1992, ISBN 0-12-174230-X.

[17] Coles, S.: *An Introduction to Statistical Modeling of Extreme Values.* Springer, London, 2001, ISBN 1-85233-459-2.

[18] D'Amico, Guglielmo, Filippo Petroni, and Flavio Prattico: *Wind speed forecasting at different time scales: a non parametric approach.* Physica A: Statistical Mechanics and its Applications, 406:59–66, July 2014.

[19] Ellis, G.W. and A.S. Cakmark: *Time series modelling of strong ground motion from multiple event earthquakes.* Soil Dynamics and Earthquake Engineering, 10(1):42–54, January 1991.

[20] Friedland, Carol J., T. Andrew Joyner, Carol Massarra, Robert V. Rohli, Anna M. Treviño, Shubharoop Ghosh, Charles Huyck, and Mark Weatherhead: *Isotropic and anisotropic kriging approaches for interpolating surface-level wind speeds across large, geographically diverse regions.* Geomatics, Natural Hazards and Risk, 8(2):207–224, 2017.

[21] Fuller, W.A.: *Introduction to Statistical Time Series*. John Wiley and Sons, 1995, ISBN 0-471-28715-6.

[22] Giwhyun, Lee, Ding Yu, G. Marc, and Xie Le: *Power curve estimation with multivariate environmental factors for inland and offshore wind farms*. American Statistical Association, 110(509):56–67, 2015.

[23] Giwhyun, Lee, Ding Yu, and Genton Marc.G.: *A kernel plus method for quantifying wind turbine performance upgrades*. Wind Energy, 18(7):1207–1219, 2014.

[24] Goovaerts, P.: *Geostatistics for Natural Resources Evaluation*. Applied geostatistics series. Oxford University Press, New York, 1997, ISBN 9780195115383. http://books.google.gr/books?id=CW-7tHAaVR0C.

[25] Granger, C.W.J and A.O Hughes: *A new look at some old data: the Beveridge wheat prices series*. Royal Statistical Society, 134(3):413–428, 1971.

[26] Hansen, J., M. Sato, R. Ruedy, K. Lo, D.W. Lea, and M. Medina-Elizade: *Global temperature change*. Proceedings of the National Academy of Sciences, 103:14288–14293, 2006.

[27] Hristopoulos, D. T.: *Random Fields for Spatial Data Modeling: A Primer for Scientists and Engineers*. Springer, Dordrecht, the Netherlands, 2020.

[28] Hristopoulos, Dionissios T.: *Stochastic local interaction (sli) model: Bridging machine learnig and geostatics*. Computers & Geosciences, 85:26–37, May 2015.

[29] Hu, Hongda, Zhiyong Hu, Kaiwen Zhong, Jianhui Xu, Pinghao Wu, Yi Zhao, and Feifei Zhang: *Long-term offshore wind power prediction using spatiotemporal kriging: A case study in china's guangdong province*. Energy Exploration & Exploitation, 38(3):703–722, 2020. https://doi.org/10.1177/0144598719889368.

[30] Huang, Z.and Chalabi, Z.S: *Use of time-series analysis to model and forecast wind speed*. Wind Engineering and Industrial Aerodynamics, 56(2-3):311–322, May 1995.

[31] Hur, J. and R. Baldick: *Spatial prediction of wind farm outputs using the augmented kriging-based model*. In *2012 IEEE Power and Energy Society General Meeting*, pages 1–7, 2012.

[32] Isaaks, E.H. and R.M. Srivastava: *Applied Geostatistics*. Oxford University Press, 1989, ISBN 9780195050134. https://books.google.gr/books?id=vC2dcXFLI3YC.

[33] Ivan, Komusanac, Fraile Daniel, and Brindley Guy: *Wind energy in europe in 2018 energy*. https://windeurope.org/wp-content/uploads/files/about-wind/statistics/WindEurope-Annual-Statistics-2018.pdf.

[34] Jiang, R. and D.N.P. Murthy: *A study of Weibull shape parameter: Properties and significance*. Reliability Engineering & System Safety, 96(12):1619–1626, 2011.

[35] Jinfu, Liu, Ren Guorui, Yufeng Guo Jie, Wan, and Yu Daren: *Variogram time–series analysis of wind speed*. Renewable Energy, 99:483–491, July 2016.

[36] Jing, Shi, Qu Xiuli, and Zeng Songtao: *Short-term wind power generation forecasting: Direct versus indirect arima-based approaches*. International Journal of Green Energy, 8(1):100–112, 2011. https://doi.org/10.1080/15435075.2011.546755.

[37] Jooyoung, Jeon and W.Taylor James: *Using conditional kernel density estimation for wind power density forecasting*. American Statistical Association, 107(497):66–79, March 2012.

[38] Journel, A. G. and C. J. Huijbregts: *Mining Geostatistics*. New York : The Blackburn Press, 2003, ISBN 1930665911 (PBK). Mining. Statistical Mathematics (BNB/PRECIS).

[39] Kariniotakis, G.N, G.S. Stavrakis, and E.F. Nogaret: *Wind power forecasting using advanced neural networks models*. IEEE Transactions on Energy Conversion, 11(4):762–767, December 1996.

[40] Krige, D. G.: *A statistical approach to some basic mine valuation problems on the witwatersrand*. Journal of the Chemical, Metallurgical and Mining Society of South Africa, 52(6):119–139, 1951.

[41] Kwiatkowski, D.; Phillips, P. C. B.; Schmidt P.; Shin Y.: *Testing the null hypothesis of stationarity against the alternative of a unit root*. Econometrics, 54:15–178, 1992.

[42] Lenzi, A., P. Pinson, L.H. Clemmensen, and *et. al.*: *Spatial models for probabilistic prediction of wind power with application to annual-average and high temporal resolution data.* Stochastic Environmental Research and Risk Assessment, 31:1615–1631, September 2017.

[43] Lydia, M., Suresh Kumar, A. immanuel Selvakumar, and G. Edwin Prem Kumar: *A comprehensive review on wind turbine power curve modeling techniques.* Renewable and Sustainable Energy Reviews, 30:452–460, 2014.

[44] Medeiros, Marcelo and Lacir Soares: *Modeling and forecasting short-term electricity load: A comparison of methods with an application to brazilian data.* International Journal of Forecasting, 24:630–644, October 2008.

[45] Montgomery, D.C, C.L Jennings, and K. Murat: *Introduction to Time Series Analysis and Forecasting.* Willey.

[46] Olea, R.: *A six-step practical approach to semivariogram modeling.* Stochastic Environmental Research and Risk Assessment, 20(5):307–318, 2006, ISSN 1436-3240. http://dx.doi.org/10.1007/s00477-005-0026-1.

[47] Olea, R.A.: *Geostatistics for Engineers and Earth Scientists.* Springer US, 1999, ISBN 9780792385233. http://books.google.gr/books?id=bKoD2mM0RHUC.

[48] Ommar, Ellabban, Abu Rub Haitham, and Blaabjerg Frede: *Renewable energy resources: Current status, future prospects and their enabling technology.* Reneable and Sustainable Energy Reviews, 39:748–764, November 2014.

[49] Papoulis, A. and S.U. Pillai: *Probability, Random Variables and Stochastic Process.* McGraw-Hill Inc., New York, 4h edition, 2002.

[50] Pavlides, A. G.: *Development of New Geostastical Methods for Spatial Analysis and Applications in Reserves Estimation and Quality Characteristics of Coal Deposits.* PhD thesis, Technical University of Crete, 2016.

[51] Peter, J.Brockwell and A.Davis Richard: *Introduction to Time Series and Forecasting.* Springer.

[52] project.org, R core R-core@R: *Simulate from an arima model.* https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/arima.sim, visited on April 2020.

[53] REN21: *Renewables 2016 global status report*. Technical report, Paris REN21 Secretariat, 2016, ISBN 9783981810707.

[54] REN21: *Renewables global futures report:great debates towards 100%renewable energy*. Technical report, Paris REN21 Secretariat, 2016, ISBN 9783981810745.

[55] REUK.co.uk: *Wind turbine tip speed ratio*. http://www.reuk.co.uk/wordpress/wind/wind-turbine-tip-speed-ratio/, visited on April 2020.

[56] Schwarz, Gideon: *Estimating the dimension of a model*. The Annals of Statistics, (2):461–464, 1978.

[57] Seguro, J.V and T.W. Lambert: *Modern estimation of the parameters of the Weibull wind speed distribution for wind energy analysis*. Wind Engineering and Industrial Aerodynamics, 85(1):75–84, March 2000.

[58] Shahriar, Shafiee and Topal Erkan: *When will fossil fuel reserves be diminished?* Energy Policy, 37:181–189, 2009.

[59] Shumway, Robert. H and S. Stoffer, Dacid: *Time Series Analysis and Its Applications With R Examples*. Springer, 4th edition, 2016.

[60] Soman, Saurabh, Hamidreza Zareipour, O.P. Malik, and Paras Mandal: *A review of wind power and wind speed forecasting methods with different time horizons*. In *North American Power Symposium 2010, NAPS 2010*, pages 1–8, October 2010.

[61] Thomaidis, Nikolaos: *Designing strategies for optimal spatial distribution of wind power*. SSRN Electronic Journal, October 2012.

[62] Vanmarcke, E.: *Random Fields: Analysis and Synthesis*. World Scientific, Princeton University, USA, 2010.

[63] Wikipedia: *Wind energy*. https://en.wikipedia.org/wiki/Wind_power#cite_note-GWEC_Market-1, visited on July 2019.

[64] Wikipedia: *Wind power by country*. https://en.wikipedia.org/wiki/Wind_power_by_country, visited on May 2019.