Karlsruhe Institute of Technology

Evaluating the benefit of grid-based weather information in energy forecasting

Bachelors's Thesis of

Marcel Herm

at the Department of Informatics
Institute for Automation and Applied Informatics (IAI)

Reviewer:	Prof. Dr. Veit Hagenmeyer
Second reviewer:	Prof. Dr. Achim Streit
Advisor:	Nicole Ludwig, M.Sc
Second advisor:	Marian Turowski, M.Sc

Summer Term – 2019

Karlsruher Institut für Technologie
Fakultät für Informatik
Postfach 6980
76128 Karlsruhe

I declare that I have developed and written the enclosed thesis completely by myself, and have not used sources or means without declaration in the text.

PLACE, DATE

. . . .

(Marcel Herm)

Abstract

As the share of electricity from regenerative sources is growing constantly, the weather becomes an increasingly important factor in the analysis of electricity markets. Hence, this thesis uses local weather data to predict electricity spot prices. More precisely, we include wind speed and temperature from individual German weather stations into time series and statistical learning models. However, as the available weather information is vast and renewable power is not generated everywhere, we use random forests and Bayesian structural time series to perform a feature selection. Overall, we manage to improve our forecasting accuracy of the EPEX electricity prices by up to 7.69 % in terms of root mean squared error and up to 8.19 % in terms of mean absolute error.

Contents

List of Figures

List of Tables

1. Introduction

According to Li et al. (2009), especially temperature and perceived temperature have a great impact on energy demand. Consequently, several authors combine forecasting energy time series using weather data. They mostly either focus on forecasting photovoltaic (PV) electricity generation as in Bofinger and Heilscher (2014) and Sperati et al. (2016) or on electricity generation from wind as in Davò et al. (2016) and Alessandrini et al. (2015).

It is notable, that works using station-based data often try to do some sort of geographic interpolation to be able to obtain values for every possible position. Considering this thesis, there is no such problem, as the used grid-based data already provides such distributed values. Thus, when using grid-based data, the step of interpolation can be omitted, and therefore, less effort is required. Some works also regard only forecasting for specific locations or accumulated values for bigger areas. With grid-based data, more general predictions can be made regarding the target location, as there is no binding to a certain locality.

Another interesting point is, that the works that forecast weather related time series all used grid-based data, though some of them also used station-based data to refine their forecasts. Among the papers that aimed for forecasting power or similar often only station-based data is used which leads to the assumption that less effort has been made in these fields, as it is still more complex to acquire grid-based data. However, there remains the possibility that station-based data is more suitable, even though this means a trade-off in terms of flexibility. Of course it is also possible that this has to do with the fact, that there is no grid-based power data available as this may harm privacy issues.

2. Related Work

This chapter gives an overview of related work in the field of energy forecasting considering grid-based data. After describing the general approach of research, there will be some sources ordered by degree of relation to this thesis.

In order to find relevant literature, *arXiv*,*Google Scholar* and *BASE* have been used in order to find suitable reading.
In the process of research, the following criteria have been used to identify relevant literature:

- The title suggests working with geographic or grid-based data
- The title implies being applied in the field of energy networks
- The title suggests to aim at forecasting values
- If not mentioned in the title, does the abstract or introduction suggest working with geographic or grid-based data
- If not mentioned or further explained in the title, does the abstract or introduction suggest to aim at forecasting or rather how is forecasting done

When it comes to weather based prediction of power, a lot of papers have been published. Some of them use grid-based data, but in the process of research, there didn't come up any that had a focus on the aspect of how to most suitably include grid-based data to forecast energy time-series. However, there has been related work, even though not having the exact same subject, which is evaluating how to properly include grid-based data for energy time series forecasting. Since in research it is important to acquire reference knowledge, these papers are employed as a such. Subsequently, they will be outlined and explained regarding their type of data, the forecast time series, applied forecasting methods and the forecast horizon.

As the related works are presented ordered by relevance, the first papers are quite close to this thesis, forecasting some sort of electricity generation with grid-based data.
The first paper presented is Davò et al. (2016) who utilize grid-based wind speed data generated by applying the Regional Atmospheric Modelling System (RAMS) with boundary conditions from ECMWF. Furthermore, they acquire grid-based data of solar radiation energy per square meter coming from National Oceanic and Atmospheric Administration - Earth System Research Laboratory (NOAA/ESRL) which was provided for an online competition hosted by *Kaggle* as one of the two forecast time series is the solar irradiance. Reference power data is obtained from *Terna* as the other predicted time series is the wind power produced over Sicily. Here, a Principal Component Analysis (PCA) is employed which is an important point as grid-based data is even more prone to the curse of dimensionality due to the two additional dimensions besides time. In terms of forecasting they apply Neural Networks (NN) and an Analog Ensemble (AnEn). The forecast horizon has a range of 0 to 72 hours and the output is a prediction of both, wind power and solar radiation.
Similar to this thesis, De Felice et al. (2015) use grid-based data from ECMWF to forecast the electricity demand, though for Italy. Therefore Linear Regression (LR) and a Support Vector Machines (SVM) are applied. Given that power prediction is a rather complex problem, the non-linear SVM performs better than a simple LR.
Another application of grid-based data from ECMWF is proposed in Sperati et al. (2016) for solar power prediction in Italy. They implement a Probability Density Function (PDF) combined with NN, Variance Deficit (VD), Ensemble Model Output Statistics (EMOS) and Persistence Ensemble (PE). The time series forecast includes a range of 0-72 hours.
The second group of related works covers those that also forecast electricity generation, but in contrast to the first group, make use of station-based data.
Alessandrini et al. (2015) utilize non-gridded wind and power data from a wind farm in northern Sicily in Italy with which they forecast generated wind power. Here, a novel approach, an AnEn, which originally is used for meteorological ensemble forecasts, is applied to the data to retain a probabilistic prediction for the next 0-132 hours.
A work where station-based weather data is applied to forecast low voltage load in the United Kingdom, has been published by Haben et al. (2018). They implement Kernel Density Estimation (KDE), Simple Seasonal Linear Regression (SSLR), Autoregressive model using an average weekly profile (ARWD), Autoregressive model using an average weekly profile including annual seasonality (ARWDY) and Holt-Winters-Taylor Exponential Smoothing Method (HWT-ESM) and compare them for forecast horizons of up to 4 days.

Bofinger and Heilscher (2014) also acquire data from local weather stations to forecast solar power generation. The data is then refined with grid-based data from ECMWF by applying Model Output Statistics (MOS) and Inverse Distance Weighting (IDW), spatially interpolated and then simulated for Germany in order to predict a temporal range of 24-120 hours.
Then, there is e. g. Aguiar et al. (2016), using both, grid-based and station-based weather data to improve Global Horizontal Solar Irradiance (GHI) forecasts on Gran Canaria Island. In order to achieve this, NN are applied. As the authors consider intra-day forecasting, the forecast horizon here is limited to a range of 1 to 6 hours.
The last group of related works deals with those that either utilize grid-based, station-based or both types of data to forecast various time series.
In terms of this thesis, a very interesting paper is Diagne et al. (2013), where grid-based weather data is used for solar radiation forecasting, which is similar to GHI which is forecast in the previous paper. Also different data sources are compared, specifically ECMWF, Fifth-generation Mesoscale Model (MM5) and Weather Research and Forecasting Model (WRF). The paper focuses on Autoregressive (AR) methods including ARMA, Autoregressive Integrated Moving Average (ARIMA) and Coupled Autoregressive and Dynamical System (CARDS), NN and Wavelet Neural Networks (WNN) considering short time ranges from 5 min up to 6h.
Similarly, Salcedo-Sanz et al. (2018) also utilize grid-based weather data to forecast solar radiation in Australia. The evaluated methods are combinations of Coral Reefs Optimization (CRO), Extreme Learning Machine (ELM), Grouping Genetic Algorithm (GGA), Multivariate Adaptive Regression Splines (MARS), Support Vector Regression (SVR) for a forecast horizon of 24 hours.
A different paper from Ludwig et al. (2015), investigates the usage of station-based weather data from Deutscher Wetterdienst (DWD) for electricity price forecasting in Germany. The price history is obtained from European Power Exchange (EPEX SPOT). The work does not consider NN, but rather compares Least Absolute Shrinkage Selection Operation (LASSO) and Random Forests (RF) in addition to ARMA and ARMAX models. A desirable side effect from RF is the output of the variable importance which is useful in order to filter variables by order of their relevance. As a this work has a focus on short-term forecasts, the forecast time series here is the electricity price for the next day, thus having a forecast horizon of 24 hours.

Of course, there where other papers considered such as Kamińska-Chuchmala (2014), in particular due to its subject of forecasting internet traffic load and the high correlation between internet traffic load and electricity load, as can be read in Morley et al. (2018). They apply Ordinary Kriging (OK) to spatially interpolate station-based data. As interpolation itself is not topic of this thesis, this subject has not been considered for further research. However, due to the great similarity between these fields, it is an has great for related research and also influenced this work. Furthermore it underlines the benefit of grid-based data by illustrating the effort processing station-based data which is not neccessary for grid-based data. Another great example utilizing station-based data is Fairley et al. (2017) investigating marine energy generation evaluating implications for electricity supply are being discussed. This combines localization issues and the electricity network. Unfortunately, the aim here is not to forecast but only to examine the problem.

Table 2.1 provides an overview about the mentioned related works regarding the type of the used weather data, used methods, the place of origin of the data, the forecast horizon and the forecast time series.

paper	type of weather data	methods	location	forecast horizon	forecast time series
Davò et al. (2016)	grid-based	PCA,AnEn,NN	Sicily	0-72h	wind power,solar radiation
De Felice et al. (2015)	grid-based	LR,SVM	Italy	1-2 months	electricity demand
Sperati et al. (2016)	grid-based	PDF,NN,VD,EMOS,PE	Italy	0-72h	solar power
Alessandrini et al. (2015)	station-based	AnEn	Sicily	0-132h	wind power
Haben et al. (2018)	station-based	KDE,SSLR,ARWD, ARWDY,HWT-ESM	United Kingdom	up to 4 days	low voltage electricity load
Bofinger and Heilscher (2014)	mixed	MOS,IDW	Germany	24-120h	solar power
Aguiar et al. (2016)	mixed	NN	Gran Canaria Island	1-6h	solar radiation
Diagne et al. (2013)	grid-based	ARMA,ARIMA,CARDS, NN,WNN	-	5 min-6h	solar radiation
Salcedo-Sanz et al. (2018)	grid-based	ELM,CRO,MARS, MLR,SVR,GGA	Australia	24h	solar radiation
Ludwig et al. (2015)	station-based	ARMA,ARMAX,LASSO,RF	Germany	24h	energy prices
This thesis	grid-based	LR,ARMA,ARMAX,PCA	Germany	1-24h	electricity load

Table 2.1. List of related works regarding the type of the used weather data, used methods, place of origin of the data, forecast horizon and forecast time series.

3. Methodology

In this chapter, the used forecasting methods are explained more in detail and why they were used, but also other methodological aspects of this thesis will be outlined.

3.1. Data acquisition

In order to acquire the needed weather data, ECMWF's Python-API has been used for automated data acquisition.

3.2. Forecasting methods

For time series forecasting, often used methods are e.g. ARMA models as mentioned in Hyndman and Athanasopoulos (2018), but also NN, where it is common to reduce the number input variables in order to speed up computation, which is desirable for the huge amount of grid-based data that grows quadratically with size. There are also some papers that use regression models other than ARMA such as LR,MLR or SVM.
E.g. Aguiar et al. (2016) uses NN to do intra-day forecasting of solar radiation within 1-6h on Gran Canaria and, as in this thesis, grid-based data from ECMWF is used.

3.2.1. Linear Regression

3.2.2. ARIMA

One of the most frequently used methods for forecasting in time series forecasting is ARMA, which is a combination of Autoregressive and Moving Average terms. ARIMA is quite similar to this, but also involves a differentiation term to consider the trend of past data.

3.2.3. ARIMAX

3.2.3.1. only calendar variables as exogenous inputs

3.2.3.2. additional weather variables as exogenous inputs

3.3. Feature Selection

3.3.1. Principal Component Analysis

4. Evaluation

4.1. Data

Of course choosing data sources as well as sorting and cleaning the data also requires a certain amount of time and effort. Thus it will be explained hereinafter how this has been done for the data used in this thesis.

4.1.1. ECMWF

The data used in this thesis originates from ECMWF, which is a research institute that produces global numerical weather predictions and other data.
It is time series based and for each timestamp there is a 2-dimensional array referred to by longitude and latitude respectively.

It must be mentioned that, as the data used has been reanalysed, so the expected error is likely to be smaller than if working with real-time data.

As data parameters there are also longitude and latitude, where the longitude is chosen to be from 5.5 to 15.5 and the latitude from 47 to 55.5. As the resolution of the used grid is at 0.25°, this results in a total of 1435 grid points per timestamp. As the range of the data from ECMWF extends from 2015/1/1 to 2019/3/31(TODO update), there is a total of 1551 days with each 12 timestamps due to the 2 hours frequency and thus 18612 timestamps. Considering that there is a value for each point in the grid and every timestamp, there are 26708220 values for each variable.

Plotting the data on a map results in a result as can be seen in **??** and Figure 4.1. In this case the temperature measured at 2 meters is visualized.

In order to reduce complexity, a shapefile of the NUTS dataset was used. The shapefile contains all countries in the EU. The shape of Germany was filtered from this data and

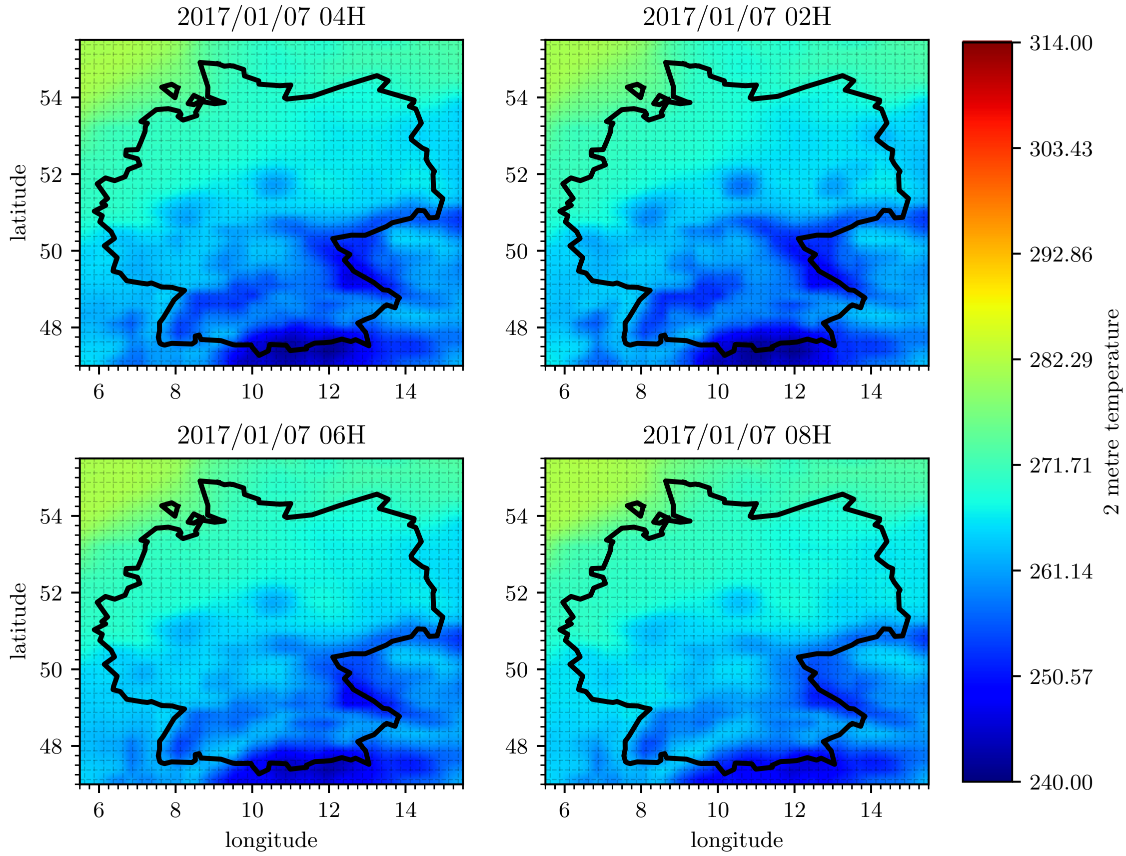


Figure 4.1. Map showing 4 times with highest temperature variance in Germany, where top left is highest, top right second highest, bottom left third highest and bottom right fourth highest variance (TODO put this in text).

each point in the dataset is checked whether it is within Germany or not. The result can be seen in Section 4.1.1. The filtered map first is saved in a numpy.ndarray and then applied on the data to mask unwanted data visualized in Section 4.1.1.

The initial dataset contains a set of variables listed in Table 4.1 where also the units, min and max are shown for each variable respectively.

4.1.2. Load data

Besides weather data used to refine the forecasting results, also historic load data was needed. Therefor data has been retained from *Open Power System Data*. Figure A.1 shows the distribution of loads over time with one point per day at 12 am UTC time. The color

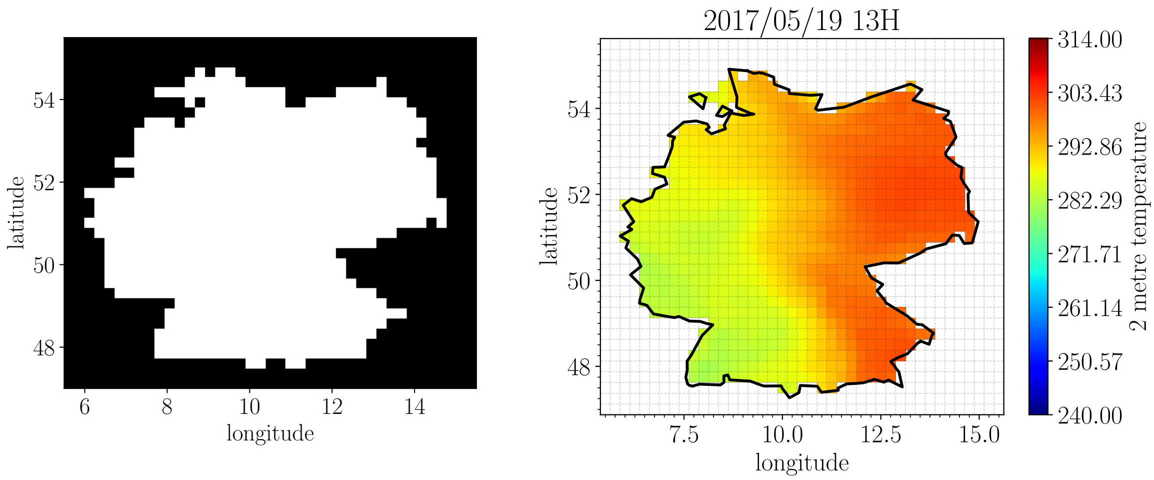


Figure 4.2. 2D boolean numpy.ndarray (left) used to filter grid squares that are within germany. It was created by using a shapefile of germany from *Eurostat* and checking for each point of the grid whether it is within the shape. When applied to the weather data, only relevant data within Germany is obtained (right).

variable name	**units**	**min**	**max**
10 metre U wind component	$m\ s^{-1}$	-18.56	21.92
10 metre V wind component	$m\ s^{-1}$	-21.51	20.00
2 metre temperature	K	240.97	313.26
Leaf area index, high vegetation	$m^2\ m^{-2}$	0.00	4.90
Leaf area index, low vegetation	$m^2\ m^{-2}$	0.00	3.84
Low cloud cover	$(0\ -\ 1)$	0.00	1.00
Soil temperature level 1	K	257.91	313.64
Surface latent heat flux	$J\ m^{-2}$	-2203977.00	359411.00
Surface net thermal radiation	$J\ m^{-2}$	-663417.00	142945.02
Surface sensible heat flux	$J\ m^{-2}$	-1703159.00	801354.00
Total cloud cover	$(0\ -\ 1)$	0.00	1.00
Total column rain water	$kg\ m^{-2}$	0.00	2.73
Total sky direct solar radiation at surface	$J\ m^{-2}$	-0.12	3088320.00

Table 4.1. List of exogenous weather variables used to forecast the load including min, max values from *ECMWF*.

shows the mean temperature measured at 2 meters.

4.1.3. Population

For further improvement and in order to figure out important point in the grid, population data for germany has been selected from *Eurostat*.

4.2. Programming part

4.2.1. Programming Language

For the programming part, Python 3.6+ has been chosen, as there is a variety of libraries to process all used file formats and because it tends to be a time saving language, also for visualization.

4.2.2. Documentation

In regard to coding styles, especially when it comes to docstrings, the numpy conventions were used. The three major points for this were first, that it is a popular and often used style, then it is also a visually oriented style which means, that it is easy to read and last it is supported by several (TODO check which, sphinx?!) autodoc tools that create a HTML based documentation from existing source code with docstrings.

4.3. Preparation

4.4. Results

4.4.1. ARIMA

Describe the results you have obtained using your methods described above. Again use proper visualization methods.

4.4.2. Experiment 1

…

4.4.3. Experiment 2

…

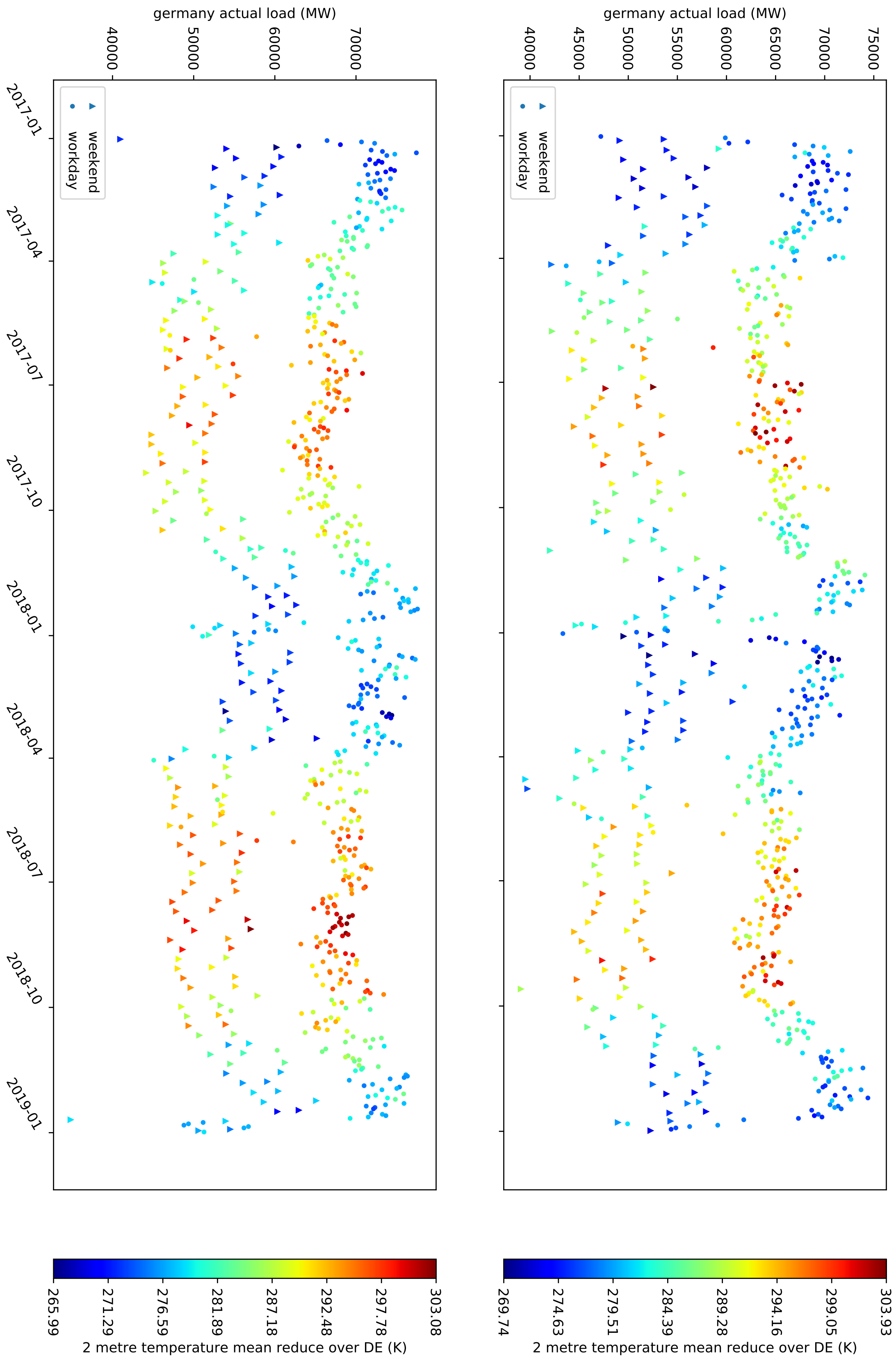


Figure 4.3. Load curve with mean of 2 meter measured temperature in Germany as color from 2015/1/1 to 2018/12/31 with one single point per day at 12am UTC time respectively.

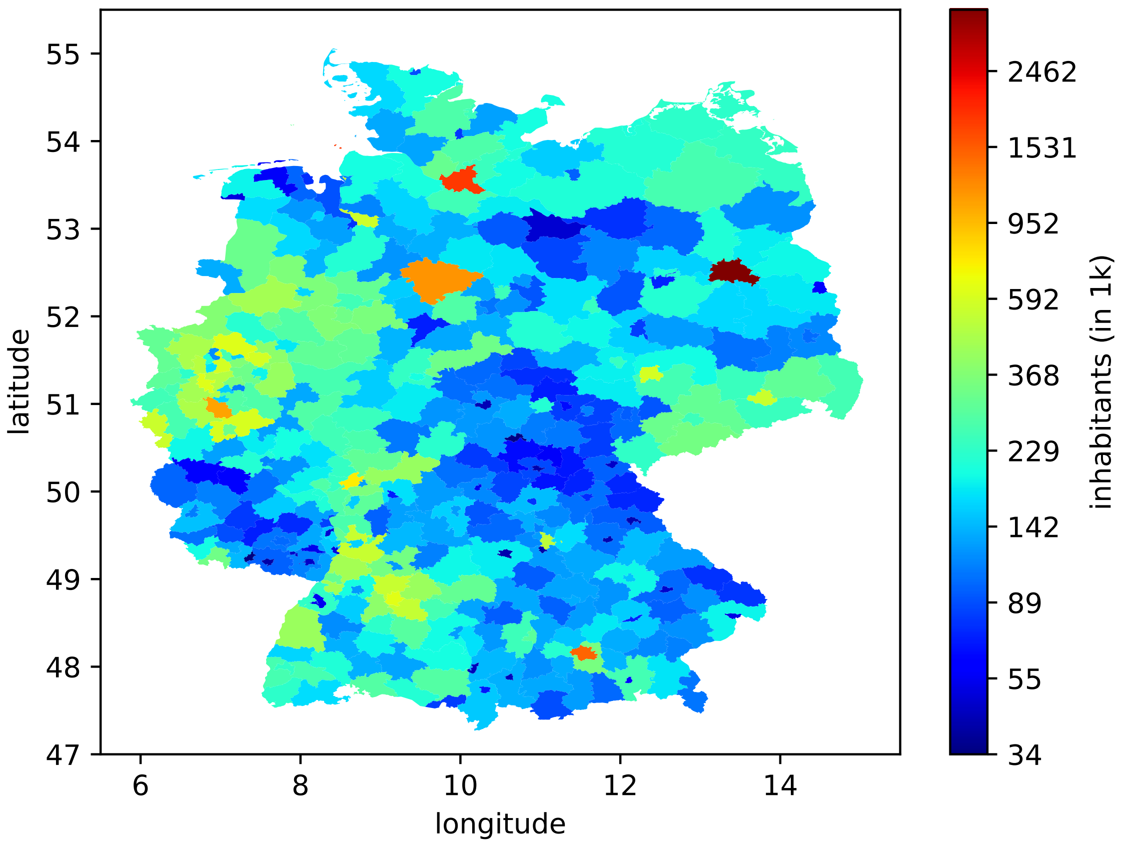
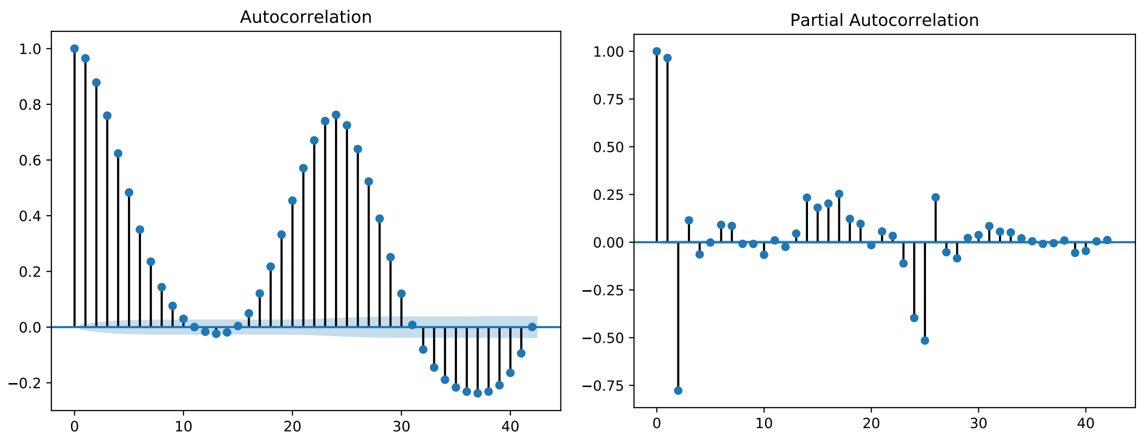


Figure 4.4. Population of Germany for each region respectively using a log scale for better distinction.



(a) (b)

Figure 4.5. Autocorrelation (a) and Partial Autocorrelation (b) plots used to select the order of ARMA and ARMAX models.

5. Discussion

This chapter is supposed to discuss your results. Point out what your results mean. What are the limitations of your approach, managerial implications or future impact?
Explain the broader picture but be critical with your methods.

6. Conclusion

It needs to be clarified, that in contrast to most of the presented works, this thesis uses reanalysed data from ECMWF as weather predictions which means, that the forecasts might behave differently from forecasts in other works as what here is assumed to be a weather forecast is more accurate than usually. This also means that results from this thesis may not exactly match results using the same procedure with real-time data.

Terms and abbreviations

AnEn Analog Ensemble. 4, 8

AR Autoregressive. 5

ARIMA Autoregressive Integrated Moving Average. 5, 8, 9

ARMA Autoregressive-Moving Average. v, 5, 8, 9, 16

ARMAX Autoregressive-Moving Average with Exogenous Inputs. v, 5, 8, 16

ARWD Autoregressive model using an average weekly profile. 4, 8

ARWDY Autoregressive model using an average weekly profile including annual seasonality. 4, 8

C3S Copernicus Climate Change Service. 3

CARDS Coupled Autoregressive and Dynamical System. 5, 8

CRO Coral Reefs Optimization. 5, 8

DWD Deutscher Wetterdienst. 5

ECMWF European Centre of Medium-Range Weather Forecasts. vii, 3–6, 9, 11

ELM Extreme Learning Machine. 5, 8

EMOS Ensemble Model Output Statistics. 4, 8

EPEX SPOT European Power Exchange. 5

GGA Grouping Genetic Algorithm. 5, 8

GHI Global Horizontal Solar Irradiance. 5

HWT-ESM Holt-Winters-Taylor Exponential Smoothing Method. 5, 8

IDW Inverse Distance Weighting. 5, 8

KDE Kernel Density Estimation. 4, 8

LASSO Least Absolute Shrinkage Selection Operation. 5, 8

LR Linear Regression. 4, 8, 9

MARS Multivariate Adaptive Regression Splines. 5, 8

MLR Multiple Linear Regression. 8, 9

MM5 Fifth-generation Mesoscale Model. 5

MOS Model Output Statistics. 5, 8

NN Neural Networks. 4, 5, 8, 9

NOAA/ESRL National Oceanic and Atmospheric Administration - Earth System Research Laboratory. 4

NUTS Nomenclature des Unités territoriales statistiques. 11

OK Ordinary Kriging. 6

PCA Principal Component Analysis. 4, 8

PDF Probability Density Function. 4, 8

PE Persistence Ensemble. 4, 8

PV photovoltaic. 3

RAMS Regional Atmospheric Modelling System. 4

RF Random Forests. 5, 8

SSLR Simple Seasonal Linear Regression. 4, 8

SVM Support Vector Machines. 4, 8, 9

SVR Support Vector Regression. 5, 8

VD Variance Deficit. 4, 8

WNN Wavelet Neural Networks. 5, 8

WRF Weather Research and Forecasting Model. 5

Bibliography

Aguiar, L. M., B. Pereira, P. Lauret, F. Díaz, and M. David (2016). *Combining solar irradiance measurements, satellite-derived data and a numerical weather prediction model to improve intra-day solar forecasting*. In: *Renewable Energy*, Vol. 97, pp. 599–610.

Alessandrini, S., L. Delle Monache, S. Sperati, and J. N. Nissen (2015). *A novel application of an analog ensemble for short-term wind power forecasting*. In: *Renewable Energy*, Vol. 76, pp. 768–781.

BASE. Bielefeld University. URL: `https://www.base-search.net/` (visited on 07/01/2019).

Bofinger, S. and G. Heilscher (2014). *Solar electricity forecast : Approaches and first results.* In: No. January 2006.

Davò, F., S. Alessandrini, S. Sperati, L. Delle Monache, D. Airoldi, and M. T. Vespucci (2016). *Post-processing techniques and principal component analysis for regional wind power and solar irradiance forecasting*. In: *Solar Energy*, Vol. 134, pp. 327–338.

De Felice, M., A. Alessandri, and F. Catalano (2015). *Seasonal climate forecasts for medium-term electricity demand forecasting*. In: *Applied Energy*, Vol. 137, pp. 435–444.

Diagne, M., M. David, P. Lauret, J. Boland, and N. Schmutz (2013). *Review of solar irradiance forecasting methods and a proposition for small-scale insular grids*. In: *Renewable and Sustainable Energy Reviews*, Vol. 27, pp. 65–76.

ECMWF. European Centre for Medium-Range Weather Forecasts. URL: `https://www.ecmwf.int` (visited on 07/01/2019).

Eurostat. Eurostat. URL: `https://ec.europa.eu/eurostat/` (visited on 07/01/2019).

Fairley, I., H. C. Smith, B. Robertson, M. Abusara, and I. Masters (2017). *Spatio-temporal variation in wave power and implications for electricity supply*. In: *Renewable Energy*, Vol. 114, pp. 154–165.

Google Scholar. Google. URL: `https://scholar.google.de/` (visited on 07/01/2019).

Haben, S., G. Giasemidis, F. Ziel, and S. Arora (2018). *Short term load forecasting and the effect of temperature at the low voltage level*. In: *International Journal of Forecasting*, No. xxxx.

Hyndman, R. and G. Athanasopoulos (2018). *Forecasting: principles and practice, 2nd edition.* OTexts: Melbourne, Australia. URL: `https://otexts.com/fpp2` (visited on 07/01/2019).

Kaggle. Google. URL: `https://www.kaggle.com/` (visited on 07/01/2019).

Kamińska-Chuchmala, A. (2014). *Spatial internet traffic load forecasting with using estimation method*. In: *Procedia Computer Science*, Vol. 35, No. C, pp. 290–298.

Li, Y., V. G. Agelidis, and Y. Shrivastava (2009). *Wind-solar resource complementarity and its combined correlation with electricity load demand*. In: *2009 4th IEEE Conference on Industrial Electronics and Applications, ICIEA 2009*, pp. 3623–3628.

Ludwig, N., S. Feuerriegel, and D. Neumann (2015). *Putting Big Data analytics to work: Feature selection for forecasting electricity prices using the LASSO and random forests*. In: *Journal of Decision Systems*, Vol. 24, No. 1, pp. 19–36.

Morley, J., K. Widdicks, and M. Hazas (2018). *Digitalisation, energy and data demand: The impact of Internet traffic on overall and peak electricity consumption*. In: *Energy Research and Social Science*, Vol. 38, No. August 2017, pp. 128–137.

Open Power System Data. Neon Neue Energieökonomik et al. URL: `https://open-power-system-data.org/` (visited on 07/01/2019).

Salcedo-Sanz, S., R. C. Deo, L. Cornejo-Bueno, C. Camacho-Gómez, and S. Ghimire (2018). *An efficient neuro-evolutionary hybrid modelling mechanism for the estimation of daily global solar radiation in the Sunshine State of Australia*. In: *Applied Energy*, Vol. 209, No. July 2017, pp. 79–94.

Sperati, S., S. Alessandrini, and L. Delle Monache (2016). *An application of the ECMWF Ensemble Prediction System for short-term solar power forecasting*. In: *Solar Energy*, Vol. 133, pp. 437–450.

Terna. Terna. URL: `http://www.terna.it` (visited on 07/01/2019).

arXiv. Cornell University. URL: `https://arxiv.org/` (visited on 07/01/2019).

A. Appendix

A.1. First Section

...

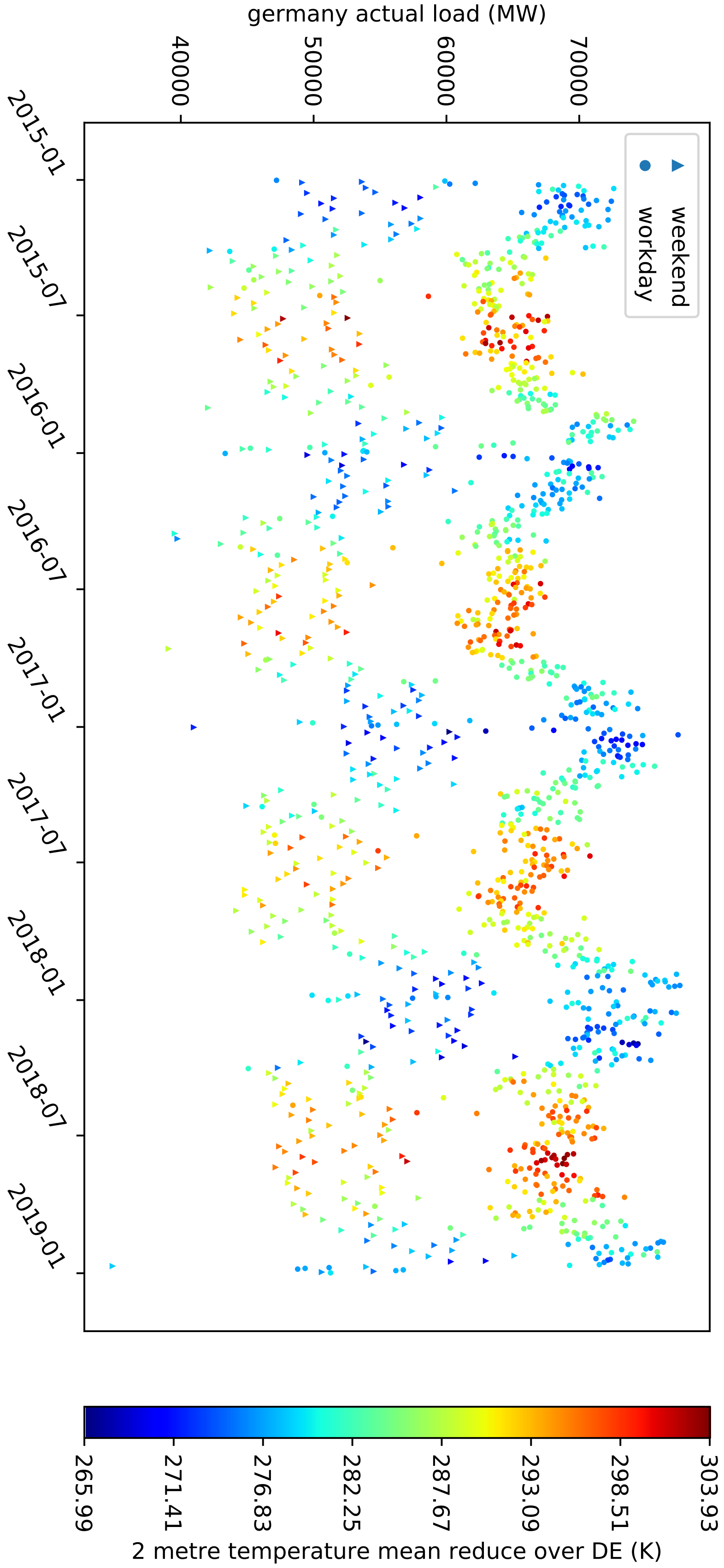


Figure A.1. Load curve with mean of 2 meter height measured temperature in germany as color from 2015/1/1 to 2018/12/31 with one single point per day at 12am utc time respectively.