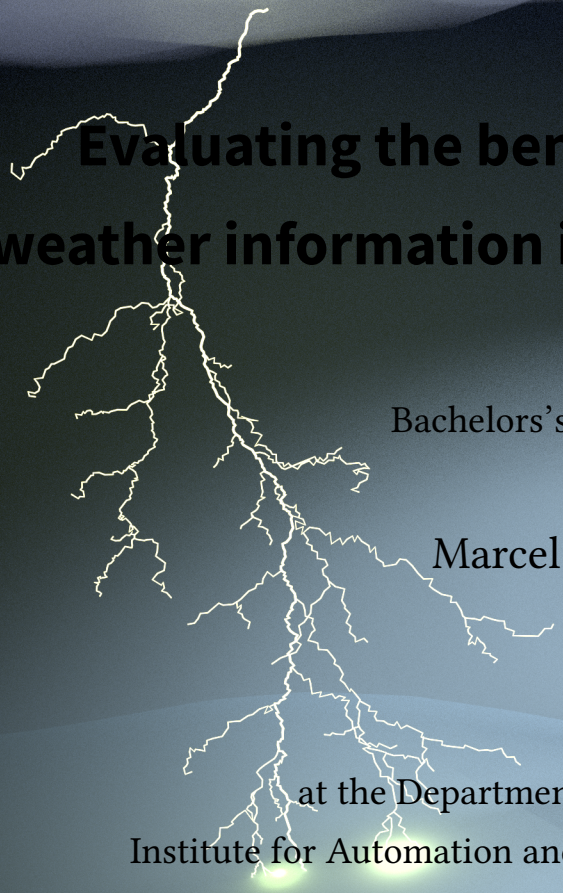
Karlsruhe Institute of Technology

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

This work is supposed to refer about including grid-based data into load forecasts using different methods.

2. Related Work

In this chapter, the subject of this thesis will be compared to similar work, a few points will be outlined and considered to be either valuable in terms of relevance for this thesis. There will also be a few points about the process of research.

2.1. Research

Gathering information is a key element in research. Therefor, *arXiv*,*Google Scholar* and *BASE* have been used in order to find suitable reading.

As the title of a work often gives a good overview, a major criteria at searching for similar papers was if the title implies working with geographic or grid-based data and/or has an application in the field of energy networks and aims at forecasting correlating values. Another criteria is the abstract and/or introduction, where the priority was to check wether the used data is grid based and if not directly mentioned in the title, if or how the forecasting is done.

2.2. Related Work

When it comes to weather based prediction of power, a lot of papers have been puplished. Some of them also used data from ECMWF, but in the process of research, there didn't come up any that had a focus on the aspect of grid-based data.

A lot of the papers that combine forecasting energy demand or generation with using weather data are focusing either on forecasting 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).

For time series forecasting, often used methods are e.g. Autoregressive (AR)-Moving Average (ARMA) models as mentioned in Hyndman and Athanasopoulos (2018) which are (TODO rough explanation of ARMA and related models). But also Neural Networks (NN) are often used, 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. Last there are also some papers that use regression models other than ARMA such as simple Linear Regression (LR),Multiple Linear Regression (MLR) or Support Vector Machines (SVM).
E.g. Aguiar et al. (2016) uses Artificial NN (ANN) to do intra-day forecasting of solar radiance (SR) on Gran Canaria and, as in this thesis, data from ECMWF is used.
Alessandrini et al. (2015) propose a novelty by applying an Analog Ensemble (AnEn) method to retain a probabilistic wind power forecast. Here the data from ECMWF is indirectly used by feeding it into the Regional Atmospheric Modeling System (RAMS) to get forecast data for the prediction.
Bofinger and Heilscher (2014) also used data from ECMWF, but also data from local weather stations to forecast solar power output. The data is then refined using Model Output Statistics (MOS), spatially interpolated and then simulated for germany.
Another paper that used data from ECMWF is Davò et al. (2016). Here also data from National Oceanic and Atmospheric Administration - Earth System Research Laboratory (NOAA/ESRL) was used which was used for an online competition hosted by *Kaggle*. Reference power data was obtained from *Terna*. This is the only paper so far using Principal Component Analysis (PCA) to reduce dimensionality, but as in Alessandrini et al. (2015), AnEn is used, whereas here, NN are used before. The target values here are both SR and wind power produced over Sicily.
Similar to this thesis, De Felice et al. (2015) aims to forecast electricity demand using data from ECMWF, though for italy and with a medium-term temporal range in contrast to the short-term range targeted in this thesis. Therefor LR and SVM are used. As power prediction is a rather complex problem, it is not very surprising that the non-linear SVM perform better than LR.
In terms of this thesis, a very interesting paper is Diagne et al. (2013) where different forecasting methods are reviewed, even though for solar radiation forecasting. Also different data sources are compared, specifically ECMWF,Fifth-generation Mesoscale Model (MM5) and Weather Research and Forecasting Model (WRF). The paper focuses on AR methods and NN considering a very short time range from 5 min up to 6h.
In contrast, Ludwig et al. (2015) does not consider NN, but therefor compares Least Absolute Shrinkage Selection Operation (LASSO) and Random Forests (RF) next to ARMA

and ARMA with Exogenous Inputs (ARMAX) models. The target value here is the german electricity price for the next day and the used data is obtained by distributed measures from Deutscher Wetterdienst (DWD) for weather data and from European Power Exchange (EPEX SPOT) for the price history. A desirable side effect from RF is the output of the variable importance which is useful in order to filter considered used variables.
Further methods are presented in Salcedo-Sanz et al. (2018) with combinations of Coral Reefs Optimization (CRO),Extreme Learning Machine (ELM),Grouping Genetic Algorithm (GGA),Multivariate Adaptive Regression Splines (MARS),Support Vector Regression (SVR) for short-term solar radiation forecast in Australia. The used data comes mostly from ECMWF, but also from Scientific Information for Land Owners (SILO) and therefor also uses gridded and non-gridded data.
Another application of ECMWF data is proposed in Sperati et al. (2016) for short-term (0-72h) solar power forecasting using a probability density function (PDF) combined with NN,variance deficit (VD),Ensemble Model Output Statistics (EMOS) and Persistence Pnsemble (PE).

Table 2.1 provides an overview about some of the mentioned related works including further information in terms of spatial distribution of the used data that does not correspond to the target value, used methods, origin of the data, temporal scope and the target value. It is to mention that regarding the temporal scope, short term means up to a few days, middle term refers to up to a few months and long term is about seasonal forecasting which possibly includes multiple years.

One key difference of the presented works to this thesis is that here, reanalyzed data from ECMWF is used for prediction which means, that the forecast tends to be far better than forecasts that are usually used in other works. This also means that results from this thesis might not exactly match results using the same procedure with real-time data.

paper	distribution	methods	data origin	scope	target
Aguiar et al. (2016)	grid-based	NN	ECMWF,HC3	intra-day	SR
Alessandrini et al. (2015)	distributed	AnEn	ECMWF/RAMS	short term	wind power
Bofinger and Heilscher (2014)	mixed	MOS,IDW	ECMWF	short term	solar power
Davò et al. (2016)	grid-based	PCA,AnEn,RAMS	ECMWF,NOAA/ESRL,*Terna*	short/mid term	wind power,SR
De Felice et al. (2015)	grid-based	LR,SVM	ECMWF	medium term	electricity demand
Diagne et al. (2013)	grid-based	ARMA,ARIMA,CARDS,ANN,WNN	ECMWF,MM5,WRF	short term?!	SR
Ludwig et al. (2015)	distributed	ARMA,ARMAX,LASSO,RF	EPEX SPOT,DWD	short term	energy prices
Salcedo-Sanz et al. (2018)	mixed	ELM,CRO,MARS,MLR,SVR,GGA	ECMWF,SILO	short term	SR
Sperati et al. (2016)	grid-based	PDF,NN,VD,EMOS,PE	ECMWF	short term	solar power

Table 2.1. List of related works and used methods respectively as well as some further details.

3. Methodology

Using weather data from ECMWF Copernicus Climate Change Service (C3S).
Using load data from `https://data.open-power-system-data.org/`.
First downloaded whole Datasets from 2006-2019, but as the load for germany is properly available since 2015, now reduced dataset to 2015-2019.
Also checked for non-existing values, only 2 last timestamps values for the load are missing.

3.1. Method 1

Maybe use Random Forests for variable selection as in Nicoles paper? (Ludwig et al. 2015)

You can also use equation numbering if you need to refer to an equation later e. g. Equation (3.1).

$$a^2 + b^2 = c^2 \tag{3.1}$$

Additionally, simple equations can be put inline with the text, for example, $x \in X$. Remember to set all variables in math font i. e. all x, i and so on.

3.2. Method 2

…

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 reanalized, 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.

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 each point in the ECMWF dataset is checked wether it is within germany or not. The result can be seen in Figure 4.1.

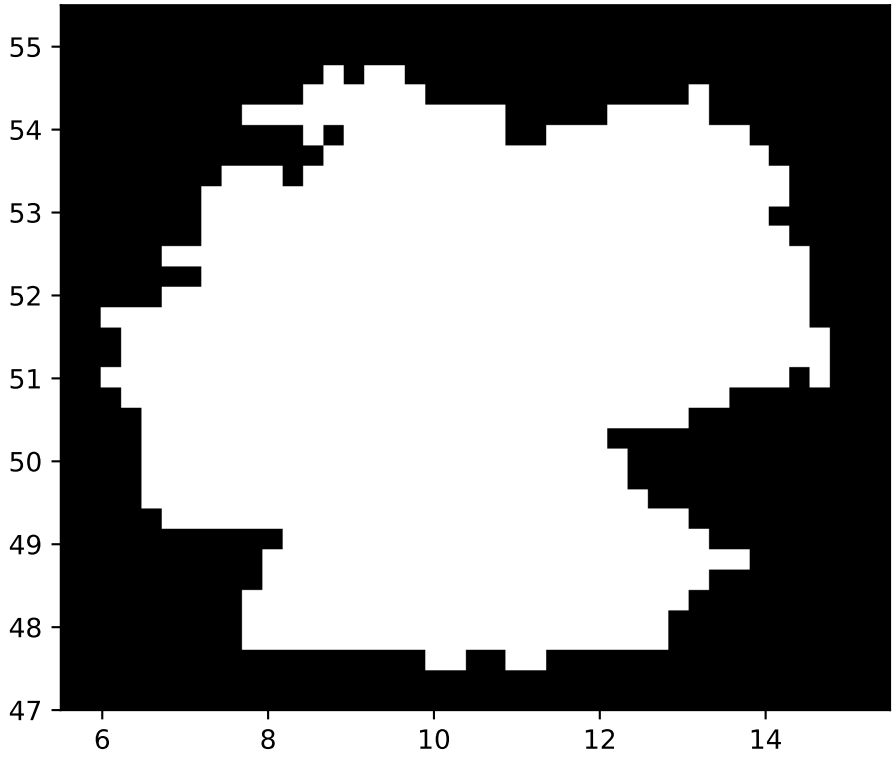


Figure 4.1. 2D boolean numpy.ndarray used to filter grid squares that are within germany. It was created by using a shapefile of germany (TODO insert source `https://ec.europa.eu/eurostat/cache/GISCO/distribution/v2/nuts/nuts-2016-files.html`) and checking for each point of the grid if it is within the shapefile. (TODO shorter explanation, put explanation in text)

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*.

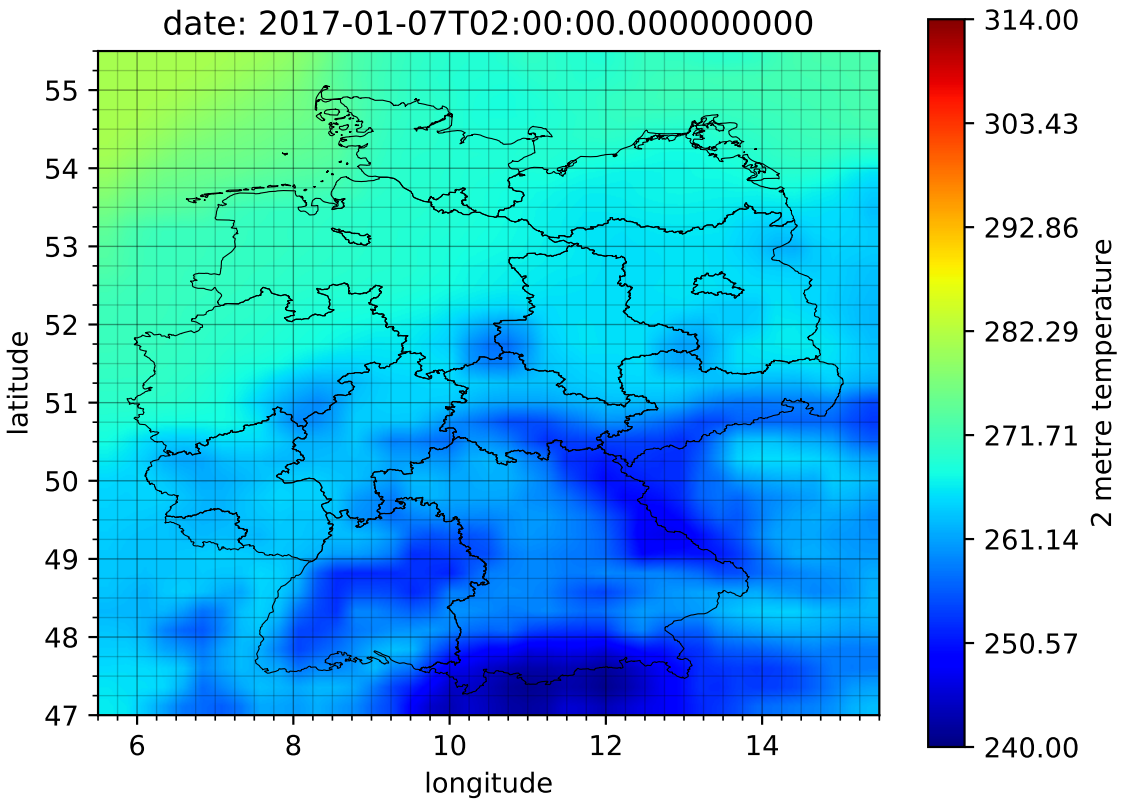


Figure 4.2. Map showing day with lowest temperature in germany.

4.1.2. Load data

The load data has been retained from *Open Power System Data*.
Figure 4.4 shows the distribution of loads over time with one point per day at 12am utc time. The color shows the mean temperature measured at 2 meters height.

4.1.3. Population

4.2. Programming part

4.2.1. Programming Language

For the programming part, Python3.6+ 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.

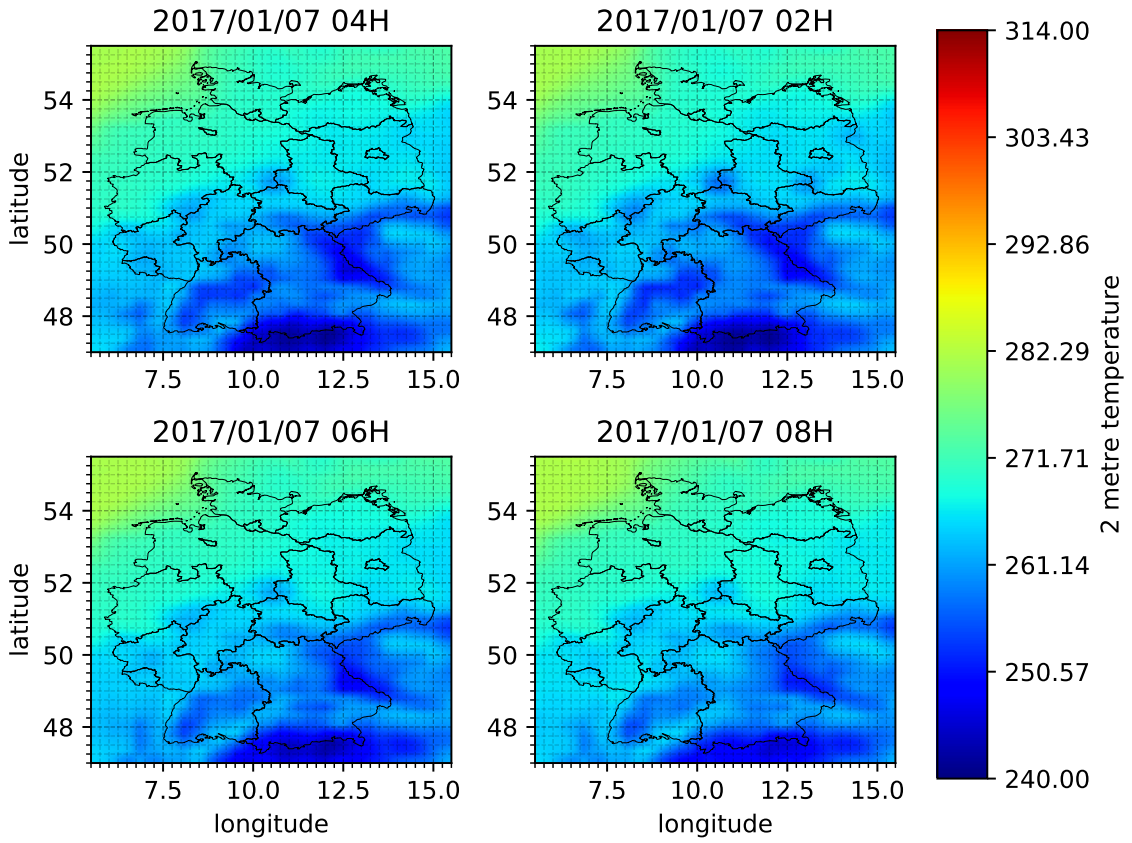


Figure 4.3. Map showing 4 times with highest temperature variance in germany, where top left is highest, top right second highest, bottom left third highest and bottom right fourth highest variance (TODO put this in text).

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. Results

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

4.3.1. Experiment 1

...

4.3.2. Experiment 2

...

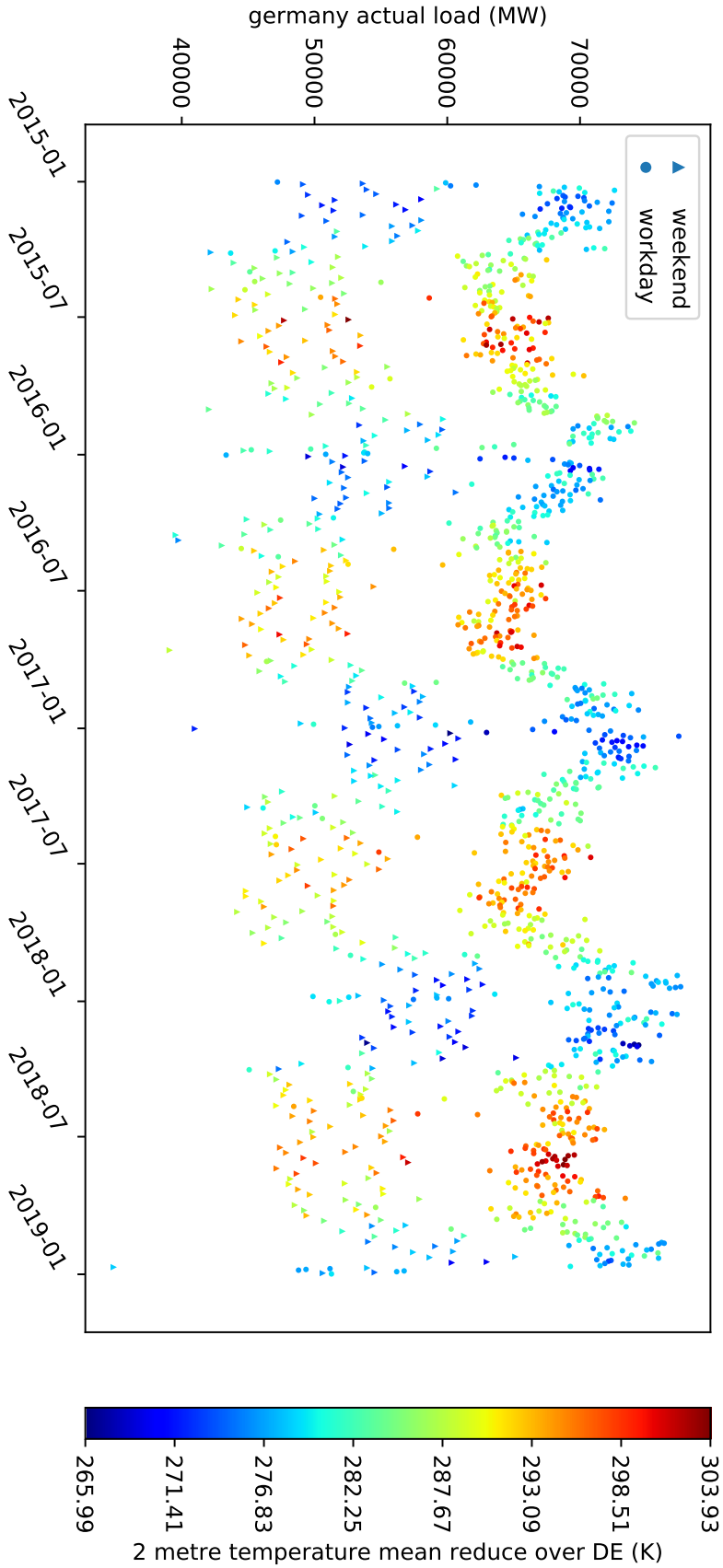


Figure 4.4. Load curve with mean of 2 meter height measured temperature in germany as color from 2015/1/1 to 2018/12/31 with one single point per day at 12am utc time respectively.

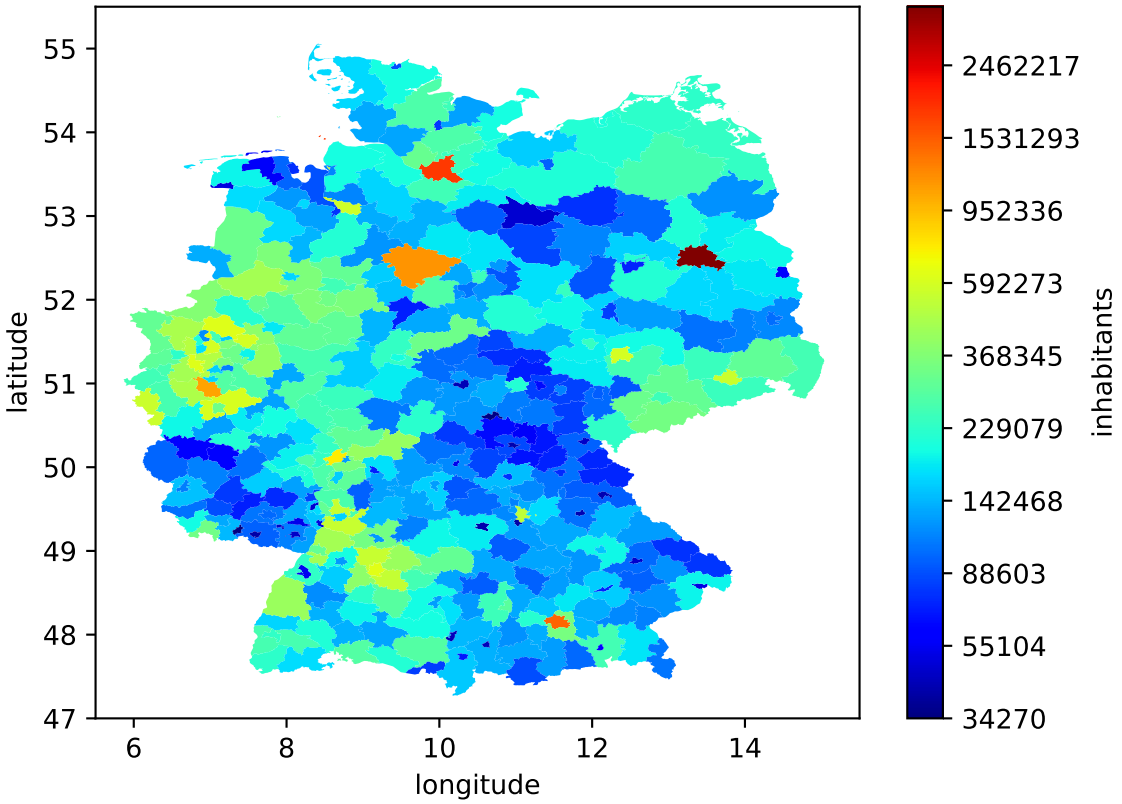


Figure 4.5. Population of germany for each region respectively using a log scale for better distinction.

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

Repeat the problem and its relevance, as well as the contribution (plus quantitative results).
Look back at what you have written in the introduction.
Provide an outlook for further research steps.

Terms and abbreviations

AnEn Analog Ensemble. 4, 6

ANN Artificial NN. 4, 6

AR Autoregressive. 4, 21

ARIMA AR Integrated Moving Average. 6

ARMA AR-Moving Average. 4–6, 21

ARMAX ARMA with Exogenous Inputs. 5, 6

CARDS Coupled AR and Dynamical System. 6

CRO Coral Reefs Optimization. 5, 6

DWD Deutscher Wetterdienst. 5, 6

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

ELM Extreme Learning Machine. 5, 6

EMOS Ensemble Model Output Statistics. 5, 6

EPEX SPOT European Power Exchange. 5, 6

GGA Grouping Genetic Algorithm. 5, 6

HC3 Helioclim-3. 6

IDW Inverse Distance Weighted. 6

LASSO Least Absolute Shrinkage Selection Operation. 4, 6

LR Linear Regression. 4, 6

MARS Multivariate Adaptive Regression Splines. 5, 6

MLR Multiple Linear Regression. 4, 6

MM5 Fifth-generation Mesoscale Model. 4, 6

MOS Model Output Statistics. 4, 6

NN Neural Networks. 4–6, 21, 22

NOAA/ESRL National Oceanic and Atmospheric Administration - Earth System Research Laboratory. 4, 6

NUTS Nomenclature des Unités territoriales statistiques. 9

PCA Principal Component Analysis. 4, 6

PDF probability density function. 5, 6

PE Persistence Pnsemble. 5, 6

PV photovoltaic. 3

RAMS Regional Atmospheric Modeling System. 4, 6

RF Random Forests. 4–6

SILO Scientific Information for Land Owners. 5, 6

SR solar radiance. 4, 6

SVM Support Vector Machines. 4, 6

SVR Support Vector Regression. 5, 6

VD variance deficit. 5, 6

WNN Wavelet NN. 6

WRF Weather Research and Forecasting Model. 4, 6

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).

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

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).

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.

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

…